BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 7 2012, pages 955–961
doi:10.1093/bioinformatics/bts057

Systems biology

Advance Access publication January 28, 2012

A co-module approach for elucidating drug–disease associations and revealing their molecular basis

Shiwen Zhao and Shao Li*

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing, China

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Understanding how drugs and diseases are associated in the molecular level is of critical importance to unveil disease mechanisms and treatments. Until recently, few studies attempt end to discover important gene modules shared by both drugs and diseases.

Results: Here, we propose a novel presentation of drug–gene–disease relationship, a 'co-module', which is characterized by closely related drugs, diseases and genes. We first define a network-based gene closeness profile to relate drug to disease. Then, we develop a Bayesian partition method to identify drug–gene–disease co-modules underlying the gene closeness data. Genes share similar notable patterns with respect not only to the drugs but also the diseases within a co-module. Simulations show that our method, comCIPHER, achieves a better performance compared with a popular co-module detection method, PPA. We apply comCIPHER to a set consisting of 723 drugs, 275 diseases and 1442 genes and demonstrate that our co-module approach is able to identify new drug–disease associations and highlight their molecular basis. Disease co-morbidity emerges as well. Three co-modules are further illustrated in which new drug applications, including the anti-cancer metastasis activity of an anti-asthma drug Pranlukast, and a cardiovascular stress-testing agent Arbutamine for obesity, as well as potential side-effects, e.g. hypotension for Triamterene, are computationally identified.

Availability: The compiled version of comCIPHER can be found at http://bioinfo.au.tsinghua.edu.cn/comCIPHER/. The 86 co-modules can be downloaded from http://bioinfo.au.tsinghua.edu.cn/comCIPHER/Co_Module_Results.zip.

Contact: shaoli@mail.tsinghua.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 30, 2011; revised on December 25, 2011; accepted on January 25, 2012

1 INTRODUCTION

Drugs achieve their therapeutic functions by targeting gene products relevant to an abnormal state. On the one hand, drugs do not always follow a 'one gene, one drug, one disease' paradigm (Hopkins, 2008). Such promiscuities result in unanticipated actions, some of which may lead to serious side-effects (Campillos *et al.*, 2008); others may induce drug new applications and therefore guide drug repositioning (Chong and Sullivan, 2007). On the other hand, complex human diseases are often multi-factor-driven, involving dysfunctions of dozens of genes (Goh *et al.*, 2007). Identifying such disease-relevant gene modules helps us decipher how an abnormal phenotype is induced and therefore in turn offers opportunities for the development of new therapies (Barabási *et al.*, 2011).

With this understanding, it is of great importance to investigate how drugs exert their activities directly or indirectly via those gene modules, how pathophenotypes are influenced by the abnormality of gene modules, and most notably, how drugs and disease phenotypes are associated on the basis of gene modules (Schadt *et al.*, 2009). However, few analyses address these questions together in a systematic view. Segal *et al.* (2004) initiatively characterized different tumor types by predefined gene modules using gene expression data. However, their method is largely dependent on the prior knowledge of gene modules and does not involve drugs. Wong *et al.* (2008) connected gene modules with human cancers by defining a module map, which was shown to guide new disease therapies. Suthram *et al.* (2010) demonstrated that common functional gene modules underlie similar diseases, and highlighted the therapeutic importance of those modules. Chiang and Butte (2009) proposed a network-based approach to discover new drug indications by connecting disease pairs sharing therapies. Novel drug uses were predicted. Recently, Gottlieb *et al.* (2011) developed a computational method, PREDICT, to identify drug–disease associations and predict new drug indications. A 0.9 area under the curve (AUC) demonstrates the power of their method. Nevertheless, few existing studies attempt to identify gene modules important both in drugs and diseases.

In current study, we aim to investigate drug–disease associations and their shared gene modules on a network basis. It has been widely shown that the modularity of genes can be characterized by a tightly interconnected subnetwork of their products in the protein–protein interaction network or the interactome network (Barabási *et al.*, 2011). In our previous studies, we proposed a measurement termed *gene closeness* to describe such interconnectedness and then used it to predict drug targets (Zhao and Li, 2010) and disease genes (Wu *et al.*, 2008). The gene closeness is calculated according to known drug–target (disease–gene) relations as well as the shortest distance in the interactome network. A gene whose products is more highly interconnected in interactome network with drug targets (disease gene products) receives a higher closeness score with respect to that drug (disease) (Section 4). Here, we use such a closeness index to define a gene closeness profile to relate drugs to diseases, and then identify important gene modules that both tightly interconnect with drug targets and disease genes from this profile.

*To whom correspondence should be addressed.

We call a gene module and those drugs and diseases associated with it a 'co-module', the word that was initially proposed by Kutalik *et al.*,in studying drug–gene associations along different cell lines (Kutalik *et al.*, 2008). In short, co-module suggests to divide biological elements into sets that share similarly significant pattern in order to study their interconnections. We introduce this concept here to analyze drug–disease relationships and, at the same time, elucidate their molecular connections by functional gene modules in an interactome network. We further develop a Bayesian partition method to identify such drug–gene–disease co-modules. We name our method comCIPHER (***co-m***odule) following the drugCIPHER (Zhao and Li, 2010) and CIPHER (Wu *et al.*, 2008) methods due to their commonality in using network closeness to study drug–gene or disease–gene relations. Within a co-module, genes share similar notable patterns in the closeness profile with respect not only to the associated drugs but also the diseases. By defining the indicator variables for each co-module, comCIPHER constructs a Markov chain that traverses the variable space and seeks to find the indicator variables that fit the statistical model to the largest extent. Monte Carlo Markov Chain (MCMC) strategies are introduced to determine transition moves. Specifically, comCIPHER first partitions the genes in the closeness profiles into different gene modules. Then, in each gene module, comCIPHER partitions the drugs and diseases into two categories: those that are associated with the gene module, and those that are not. Drugs and diseases associated with the same gene module as well as those genes themselves form a co-module (Fig. 1). In this article, we first compare comCIPHER with a state-of-the-art modular algorithm PPA (Ping-Pong Algorithm) (Kutalik *et al.*, 2008) in simulation and demonstrate a better performance of comCIPHER. Then, we use comCIPHER to identify drug–gene–disease co-modules based on the closeness profile data. The 86 co-modules are identified, in which not only drugs with common targets and diseases with shared genes are significantly enriched, but disease co-morbidity emerges as well. After multiple test corrections, 24 co-modules are selected in which new drug–disease associations and their molecular connections are indicated. Our co-module approach renders a promising perspective to investigate drug–disease associations and provides computational evidence to reveal their mechanisms basis.

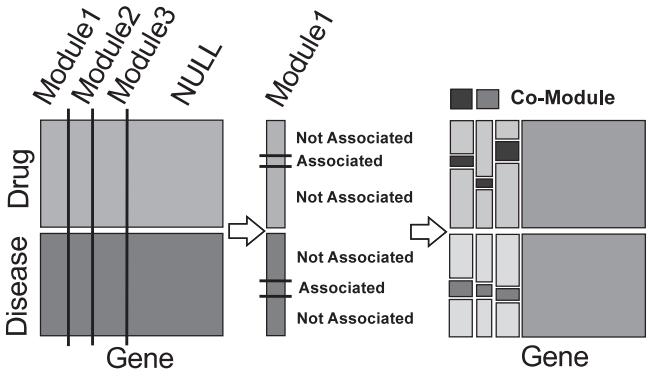

Fig. 1. Scheme of comCIPHER. ComCIPHER first partitions the genes into different gene modules, including a NULL module. In each gene module, comCIPHER further partitions drugs and diseases into two categories: those that are associated with the gene module and those are not. Drugs and diseases associated with the same gene module and those genes themselves form a co-module.

2 RESULTS

We selected FDA-approved 'promiscuous' drugs (Yildirim *et al.*, 2007) with multiple targets and diseases with multiple susceptibility genes in current study for co-module analysis. As we were also interested in human cancers, antineoplastic drugs and cancers were included. Finally, 723 drugs and 275 diseases from the DrugBank (Wishart *et al.*, 2008) and Online Mendelian Inheritance in Man (OMIM) (McKusick, 2007) databases, respectively, were selected as our candidate set. The interactome network data was an integration of five protein interaction databases (Section 4). We mapped drug targets and disease genes onto the integrated interactome network, generating 3702 drug–target and 877 disease–gene relations.

2.1 Performance of ComCIPHER

In short, comCIPHER defines the partition indicator variables for each co-module and constructs a Markov chain to sample from their posterior distribution given the observation data. After the chain burn-in, the partition configuration receiving a higher posterior density will be more likely to be sampled (See Section 4 for details). To evaluate the performance of comCIPHER, we compared it with a popular modular algorithm, PPA (Kutalik *et al.*, 2008) and a method directly using correlation coefficient. PPA is an iterative algorithm used to discover co-modules residing in two data matrices sharing one dimension and has been shown as the most robust and accurate co-module detection algorithm (Kutalik *et al.*, 2008). With a large number of random initial seeds, PPA iteratively computes three weight vectors for each dimension until a convergence condition satisfied. The converged weight vectors describe the associations between the three dimensions and are therefore interpreted as co-module indicators.

To make it comparable, we used the method described in (Kutalik *et al.*, 2008) to generate simulation data. In general, we defined three kinds of 'factors': (i) drug–disease co-factors, which induce the involvement of certain genes in drug activity, and at the same time, stimulate the susceptibilities of those genes to some diseases; (ii) pure drug factors, which only induce gene involvement in drug activities; and (iii) pure disease factors, which only affect the susceptibilities between genes and diseases. Drug–disease co-factors determine the co-modules, and pure factors determine the respective drug or disease modules. In the performed simulations, we generated 18 factors underlying 100 drugs, 80 diseases and 250 genes. There were six co-factors, six pure drug factors and six pure disease factors. For each of those six factors, we set three as positive and three as negative. We defined the complexity level as the maximum number of factors per drug or disease. With a complexity level larger than one, drugs or diseases could be associated with multiple co-modules. We further define the noise level as the variance of a normal noise with zero mean added to the simulation data (See Supplementary Material for details about simulation data). We chose noise levels of 0.2, 04, 0.6 and 0.8 with a complexity level of two, and complexity levels of one, two, three and four with a noise of 0.5 in simulation, and generated 10 matrix pairs of the drug–gene and disease–gene data for each of the noise and complexity combination. An example of the simulation data with a complexity level two and noise level 0.4 is shown in Figure 2A and B. Then, comCIPHER was applied to the

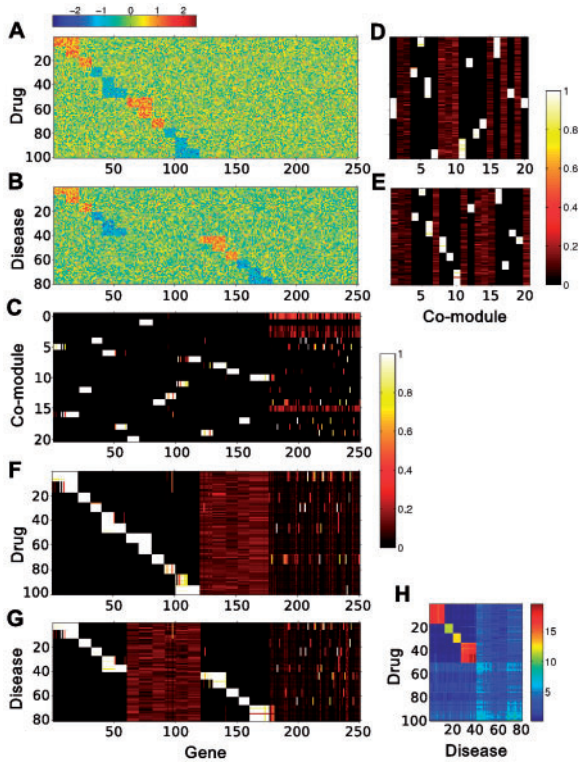

Fig. 2. An results of comCIPHER in simulation. (**A** and **B**) Drug–gene and disease–gene simulation data with complexity level two and noise level 0.4. (**C**) The estimated posterior probabilities of gene indicator variables. A row presents the probabilities of genes belonging to the co-module. (**D** and **E**) The estimated posterior probabilities of drug and disease indicator variables. A column presents the probability that different drugs or diseases belong to the co-module. (**F** and **G**) The drug–gene and disease–gene posterior association score matrices. (**H**) The drug–disease posterior association score matrix.

simulation data with a predefined co-module number M=20. We also adopted the method in PPA to compute the receiver operating characteristic (ROC) curve. We defined the drug–gene (disease–gene) posterior association score for drug i (disease j) and gene g, AS_{ig} (AS_{jg}), as the sum of the products of posterior probabilities of drug and gene (disease and gene) indicators across all co-modules except the NULL, and defined the drug–disease posterior association score for i,j, AS_{ij}, as the sum of the products of AS_{ig} and AS_{jg} across all genes (Section 4). The ROC curve was computed based on AS_{ij} as well as on the 'true' drug–disease relations. Figure 2F, G and H, respectively, depicts the AS_{ig}, AS_{jg} and AS_{ij} under the posterior probabilities of indicator variables shown in Figure 2C, D and E. For the PPA algorithm, the parameters were set to the same values mentioned in the supplement file of Kutalik *et al.*'s (2008) work. For the correlation method, we simply calculated the Pearson's correlation coefficient between drug i and disease j across all genes as the associate score. The results are shown in Table 2.1. Both of comCIPHER and PPA outperform the correlation method and exhibit good performance under low complexity and noise levels. At high complexity and noise levels, comCIPHER outperforms the PPA method.

The selection of co-module number M has impacts in the results of comCIPHER. In simulation, we found that if we chose M large,

Table 1. The areas under ROC curve

Noise level	Complexity level (average/std)			
0.5	1	2	3	4
comCIPHER	0.99/0.005	0.99/0.007	0.98/0.010	0.96/0.015
PPA	1.00/0.000	0.98/0.007	0.97/0.011	0.91/0.010
Corr	0.88/0.011	0.89/0.010	0.84/0.013	0.81/0.018
Complexity level	Noise level (average/std)			
2	0.2	0.4	0.6	0.8
comCIPHER	0.99/0.000	0.99/0.002	0.95/0.007	0.90/0.012
PPA	1.00/0.000	0.98/0.004	0.92/0.007	0.84/0.045
corr	1.00/0.000	0.98/0.005	0.79/0.021	0.65/0.016

comCIPHER was able to identify co-modules precisely. As a price, this will increase its computational burden. However if M is small, it will lead misclassifications. We also found if there were empty co-modules left after thresholding the posterior probabilities, the ROC integral for discovering drug–disease associations was fairly good. Therefore, we practically used a criterion that whether there existed empty co-modules to determine the sufficiency of M when applying to real data. See Supplementary Material for more details.

2.2 Drug–gene-disease co-module detection in gene closeness profile data

Before running comCIPHER on the drug and disease gene closeness profile data, we normalized the data to reduce the biases for drugs (diseases) with larger number of targets (disease genes). We further eliminated non-specific genes that received scores with small variation across both drugs and diseases, leaving 1442 genes for co-module analysis (Section 4). We ran comCIPHER on this subset in the interactome with M=100 and chose 0.9 as the posterior probability threshold for the indicator variables. Under these criteria, 86 co-modules were identified along with eight drug modules and two disease modules. There were four empty co-modules left, suggesting the number of M was sufficient.

We also ran PPA on the profile data to compare the performance of comCIPHER and PPA in a real application. Totally 19 921 co-modules with redundancy were obtained. We compared the following six indexes: averaged and minimum network shortest distances between drug targets (DTs) and co-module genes (CGs), disease genes (PGs) and CGs as well as DTs and PGs. The results are shown in Table 2.2. It can be seen that comCIPHER has shorter distances in all these indexes except the averaged distance between DTs and PGs, indicating drugs, diseases and genes are more interconnected in co-modules obtained by comCIPHER. We also found in the 86 co-modules identified by comCIPHER, drugs tended to have similar structural similarities (P <0.001, permutation test) and share targets (11.7-fold enrichment, P<2.2E-16, Fisher's exact test, one-sided); diseases tended to have similar phenotypic similarities (van Driel *et al.*, 2006) (P <0.001, permutation test) and share susceptibility genes (37-fold enrichment, P <2.2E-16, Fisher's exact test, one-sided). See Supplementary Material for details.

It is shown that diseases may co-occur according to shared metabolic pathways (Lee *et al.*, 2008) or cellular networks (Park

Table 2. Comparison of comCIPHER and PPA in read data

	DT-CG		PG-CG		DT-PG	
	Ave	Min	Ave	Min	Ave	Min
comCIPHER	2.20	1.35	2.19	1.52	2.66	1.30
PPA	2.49	1.91	2.37	1.91	2.54	1.36
Random	2.77 ± 0.005	2.36 ± 0.005	2.67 ± 0.012	2.37± 0.012	3.06 ± 0.011	2.36 ± 0.015

et al., 2009), resulting in disease co-morbidity. We found that diseases in the co-modules also tended to have co-morbidity. We computed the co-morbidity *P*-value for each disease pair according to co-occurrence diagnosis data in (Park *et al.*, 2009) study. We further selected disease pairs with significant co-morbidity under a 0.05 false discovery rate (FDR) control (Benjamini–Hochberg correction). In all, 902 disease pairs were selected in our candidate set, out of which 27 were found in the 86 co-modules. The results generated a 6.8-fold enrichment compared to random selection (P = 7.95E-15, Fisher's exact test, one-sided).

It is noted that high network interconnectedness might indicate but not guarantee drug–disease associations [See ref (Yildirim *et al.*, 2007) and Supplementary Material]. We, therefore, used the known drug–disease associations in the Comparative Toxicogenomics Database (CTD) database (Davis *et al.*, 2009) to filter out those co-modules unlikely to have new drug–disease associations. This is based on the assumption that if significant drug–disease pairs in a selected co-module are known to be associated, other drugs or diseases, which are topologically closely related, might also tend to be associated. For each co-module, we computed a *P*-value in which the known drug–disease associations were observed by chance (Fisher's exact test, one-sided). Under a 0.05 FDR control, 24 co-modules were selected (Benjamini–Hochberg correction). We selected three representative co-modules to provide demonstrations of our results.

A co-module relevant to cell proliferation and human cancers: there are 14 genes, 3 drugs and 3 diseases in this co-module. We find the drug targets and disease genes are either directly connected to each other or linked tightly by the co-module genes in the interactome network (Fig. 3A, Supplementary Table S2). We use gene functional annotation analysis provided by DAVID (Huang *et al.*, 2008) to investigate the functional enrichment of the 14 co-module genes. Two GO terms, regulation of apoptosis (P = 5.1E-9, FDR = 8.2E-7) and regulation of programmed cell death (P = 5.6E-9, FDR = 7.1E-7) are ranked as the most significantly enriched biological process (BP) terms (Supplementary Table S2), indicating that those genes serve important roles in the regulation of cell proliferation. Two disease pairs show significant co-morbidity: Gastric Cancer and Lung Cancer (P = 0.000003, FDR = 0.00007), Gastric Cancer and Breast-Ovarian Cancer (P = 0.000237, FDR = 0.000611). In this co-module, the angiogenesis inhibitor, Thalidomide, is found to have associations with the three diseases in the CTD database (P = 0.0012, FDR = 0.0015). Interestingly, a drug used for the treatment of asthma (Pranlukast) and an anti-bacterial agent (Minocycline) are also included in this

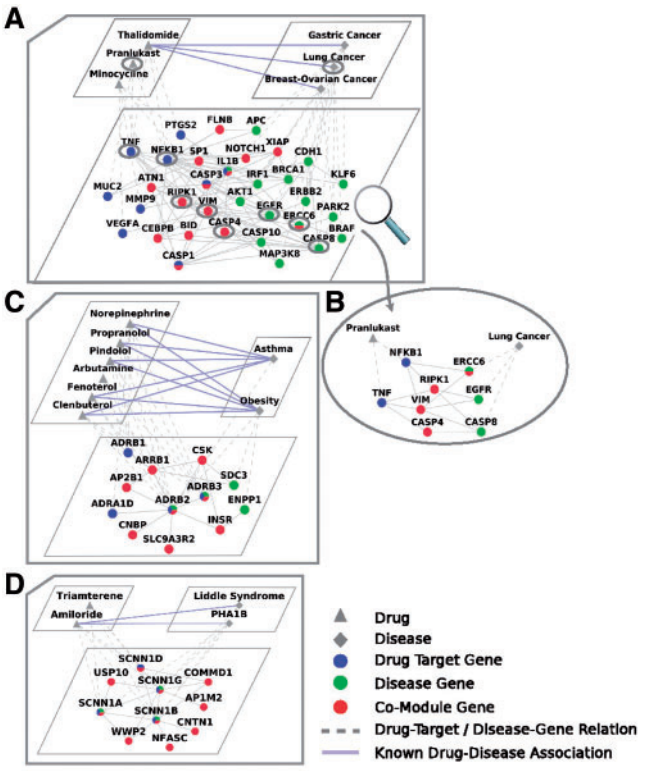

Fig. 3. Three co-module examples. (**A**) The most enriched GO term in this co-module is regulation of apoptosis. The anti-angiogenic and anti-metastatic activities respective for Minocycline and Pranlukast were reported previously and may suggest their new applications in human cancer. (**B**) A subnetwork of the co-module presented in (A) demonstrates the molecular network connections between Pranlukast and lung cancer. (**C**) Two β-adrenergic receptors (ADRB2 and ADRB3) are known drug targets, disease susceptible genes and co-module genes. Further analysis reveals Arbutamine's potential uses for asthma and obesity. (**D**) Two phenotypically opposite diseases are included in this co-module, demonstrating that different mutations in the same genes could cause distinct phenotypes. We further demonstrate the potential application of Triamterene to Liddle Syndrome, and predict Triamterene's PHA1B-like side-effects.

co-module. After further analysis, we find that Pranlukast's activity in the inhibition of tumor metastasis by targeting TNF, NF-κB and MUC2 was reported (Kishiokaa *et al.*, 2005) and is currently being investigated as a new treatment for human cancers. Minocycline also was confirmed to have anti-angiogenic activity due to its interaction with VEGF (Yao *et al.*, 2007). ComCIPHER successfully identified these new drug applications. We extracted a subnetwork of this co-module to demonstrate the possible mechanisms of association between Pranlukast and lung cancer (Fig. 3B). In one way, by targeting TNF, Pranlukast triggers cell apoptosis (Wang *et al.*, 1996) by inducing the expression of CASP4 and CASP8 (Chu and Chen, 2008), therefore exhibiting anti-cancer activity. In another way, its interactions with TNF and NF-κB influence the activity of VIM (Huber *et al.*, 2004) and RIPK1 (Wang *et al.*, 2008) , which further inhibits the EGFR gene and therefore suppresses tumor cell growth (Nicholson *et al.*, 2001). Moreover, ERCC6, an important gene in DNA repair, is directly connected to the target NF-κB. This gene is also reported as a susceptibility gene for various human

cancers, including lung cancer (Chiu *et al.*, 2008). It is possible that Pranlukast induce the expression of ERCC6 and therefore increase the DNA repair capacity in cancer cells. Our co-module approach identified new drug–disease associations for the two drugs in this co-module and highlighted their possible molecular basis.

A co-module related to β-adrenergic receptors: eight genes, six drugs and two diseases exist in this co-module (Fig. 3C, Supplementary Table S3). The three most significantly enriched GO BP terms are positive regulation of MAPKKK cascade ($P = 2.6E\text{-}4$, FDR = 8.9E-2), regulation of GPCR (G protein-coupled receptor) protein signaling pathway ($P = 2.9E\text{-}4$, FDR = 5.1E-2) and fat cell differentiation ($P = 3.1E\text{-}4$, FDR = 3.7E-2). Co-morbidity significantly exists between Asthma and Obesity ($P < E\text{-}8$, FDR < E-8). It was shown that the three β-adrenergic receptor subtypes (ADRB1, ADRB2 and ADRB3) have important influences on human muscle and fat metabolism (Mersmann, 1998) and have therefore been used as targets for the treatment of both asthma and obesity(Lazarus *et al.*, 2001). Two of the adrenergic receptors are successfully identified as co-module genes by comCIPHER. Out of the 12 drug–disease relations in this co-module, 10 are found in the CTD database (P = 6.55E-15, FDR = 6.87E-15), leaving Arbutamine as the only drug with no disease associations. Same as a β-adrenergic agonist, we believe it could serve as a new treatment for asthma. Moreover, It was reported that Arbutamine is used as a cardiovascular stress-testing agent and could achieve similar effects as exercise by targeting β-adrenergic receptors (McDOWELL, 2000). Therefore, we suggest Arbutamine could be used to treat obesity.

A co-module related to epithelial sodium channel: there are 10 genes, 2 drugs and 2 diseases in this co-module (Fig. 3D, Supplementary Table S4). GO enrichment analysis shows that Sodium channel activity ($P = 2.7E\text{-}7$, FDR = 7.6E-6) and Sodium ion binding ($P = 1.4E\text{-}5$, FDR = 2.0E-4) are ranked as the most significantly enriched molecular function terms. Interestingly, Liddle Syndrome disease, which is characterized by severe hypertension (Shimkets *et al.*, 1994), and the Pseudohypoaldosteronism Type one (PHA1B) disease, which shows vomiting, dehydration and hypotension (Chang *et al.*, 1996), are both identified as co-module diseases. The two diseases not only show significant co-morbidity ($P < E\text{-}8$, FDR < E-8) but also share two susceptibility genes. This example demonstrates the possibility that different mutations in the same genes may cause opposite phenotypes (Berger *et al.*, 2010). In this co-module, the three genes (SCNN1A, SCNN1B and SCNN1G) encoding the subunits of epithelial sodium channel (ENaC) are all identified as co-module genes. These genes are also the known drug targets and disease susceptibility genes. Two drug–disease associations are supported by the CTD database ($P = 0.0037$, FDR = 0.0047). Amiloride was demonstrated antihypertensive action by inhibiting sodium re-absorption through blocking sodium channels, therefore is associated with Liddle Syndrome (Shimkets *et al.*, 1994). Its adverse effects on moderate hypertension leading to hypotension (Saini *et al.*, 1998) may account for its association with PHA1B disease. For another drug, Triamterene, no drug–disease association was recorded in the CTD database. Nevertheless, just like those of an epithelial sodium channel blocker, the applications of Triamterene for the treatment of hypertension and Liddle Syndrome were previously reported (Heath and Freis, 1963; Wang *et al.*, 1981). Moreover, Triamterene is also a mild diuretic, and its adverse reactions, including nausea and hypotension, were studied (Knowles *et al.*, 2005). Our co-module approach not only identified a new application for Triamterene but also predicted its PHA1B-like side-effects.

3 DISCUSSION

In this study, we propose a novel interpretation of drug–disease relationships and seek to decipher their molecular basis by proposing the drug–gene–disease co-module using computational means. To the best of our knowledge, this is the first study to present and investigate drug–gene-disease relationships using functionally related co-modules. Then, we develop a Bayesian partition method to identify drug–gene–disease co-modules from gene closeness data. The simulation study demonstrates the superior performance of comCIPHER in co-module detection. As discussed, modular presentation of biological data has several advantages. First, it can discover subtle associations between biological elements that are too weak to detect by considering all of their features as a whole (Bergmann *et al.*, 2003), which also was noted in microarray data analysis (Hu and Qin, 2009). Second, the complex interconnections between elements suggest that modular measurements are more robust than treating each element individually (Kutalik *et al.*, 2008).

The approach used in the current study possesses several merits. First, we find that drugs and diseases in the 86 co-modules tend to share targets or disease genes and may therefore serve as another way to identify new drug targets and disease genes. Moreover, disease co-morbidities also tend to enrich in co-modules, suggesting the potential application of co-modules in disease diagnosis and prognosis. Second, compared with PPA method, the indicator variables used in comCIPHER provide a relatively clear structure of co-modules, which facilitates the interpretation of results. Moreover, simulations show that as long as M is sufficiently large comCIPHER can identify not only the real co-modules but also pure drug and disease modules, which could be missed by PPA. These single dimensional modules serve to find common molecular interconnections between different drugs or diseases and thus are also of value. Third, those unselected co-modules in which no significant drug–disease associations are found are still useful. On the one hand, the incompleteness in current knowledge might suggest unknown associations existing in those co-modules. On the other hand, even though prior knowledge indicates no association in a co-module, understanding why those network-related drug–disease relations do not tend to yield associations may provide new insights into the mechanistic understanding of drug–disease relations and generate new hypotheses (Hopkins, 2008).

Some aspects of the implementation of comCIPHER are worth mentioning. Although the MCMC method theoretically guarantees that the Markov chain will converge to the target distribution, because the number of iterations is always finite, it is possible that the chain will be trapped in a local mode. We handled this problem by adopting parallel tempering (Gilks *et al.*, 1996; Zhang *et al.*, 2010) with multiple chains under different temperatures and with different initial points. Still, the chain must be monitored carefully to determine its burn-in state and convergence. Moreover, when the dimension becomes very large, the chain will move very slowly. Therefore, we only focused on a small subset of genes in this study, though this may make us lose some useful information. In addition, if we choose M larger, comCIPHER tends to break data

into smaller blocks and increases the precision. As a price, this will increase its computational burden. In comparison, PPA has a tolerable computational expense. Therefore, we suggest that an integrated version which uses PPA to estimate the M and uses comCIPHER to learn the clear structure might be preferable.

In summary, we developed a novel co-module approach to discover drug–gene–disease relationships based on an interactome network, and demonstrated its application by showing that new drug–disease associations and their molecular connections were found in co-modules. Different from previous reductionism analyses, such a co-module approach offers a systematic and holistic view to study drug–disease relationships and their molecular basis. It may also provide new insights into unveiling the action mechanisms of traditional Chinese medicine, which is featured by treating multiple diseases using a herb combination from a holistic view (Li *et al.*, 2010, 2011). In total, our analysis reveals a promising perspective to study drug and disease relationships in terms of network pharmacology (Hopkins, 2008) and systems biology.

4 MATERIALS AND METHODS

Data preparation: the protein interaction data used in current study was an integration of five databases (See Supplementary Material). We mapped the UniProt protein ID to the human Entrez gene ID and obtained 137 037 PPIs for 13 388 unique genes. These data were extracted in January 2011.

The information of drugs and their targets was extracted from DrugBank (Wishart *et al.*, 2008) in February 2011. We selected FDA-approved drugs with multiple targets existing in the integrated interactome, generating 685 candidate drugs. Merged with 78 anti-neoplastic drugs in DrugBank, we generated 723 drugs with 3702 drug–target relations for co-module analysis. Disease and susceptibility gene relations were extracted from the OMIM database (McKusick, 2007) in January 2011. In all, 235 diseases were found to have multiple susceptibility genes existing in interactome network. After integrated with 76 cancer diseases, 275 diseases were included in our candidate set. Known drug–disease associations were extracted from CTD database (Davis *et al.*, 2009) in March 2011. The CTD database contains two kinds of chemical(drug)–disease relations: curated and inferred. We extracted two kinds of chemical–disease relations, both curated and inferred, without discrimination of inference score. After chemical-drug and MeSH-OMIM identifier mapping, 5018 drug–disease associations were found in our candidate sets (See Supplementary Material).

Before running comCIPHER on closeness profile data, we normalized the profiles to make the gene scores have a same distribution for different drugs or diseases, aiming to reduce the bias caused by differences in target or disease gene numbers. Then we filtered out genes which received closeness scores with small variations across drugs and diseases, generating 1442 genes for further analysis (See Supplementary Material).

Gene closeness: given a drug i, a gene g and an interactome network, the closeness of g to i is computed as follows,

$$\phi_{ig} = \sum_{g_k \in T(i)} e^{-L^2_{gg_k}}, \tag{1}$$

where g_k is a target gene of i in its target set $T(i)$, and L_{gg_k} is the shortest network distance between gene g and g_k. We computed gene closeness to disease j in a similar manner using disease genes in the network (Wu *et al.*, 2008).

ComCIPHER: we proposed a Bayesian partition method to identify drug–gene–disease co-modules. The statistical model of comCIPHER was inspired by Zhang *et al.* (2010) work. Given drug–gene and disease–gene profile data and a predefined co-module number M, comCIPHER first partitions genes into $M+1$ modules, including a NULL module. Then, for each non-NULL module, comCIPHER classifies drugs or diseases into two categories: associated with the gene module and not. The genes, drugs and diseases that are associated with that gene module form a co-module.

Consider a case where D drugs, P diseases and G genes exist. We denote the drug–gene profile data as matrix $\mathbf{Y}=((y_{ig}))$ for drug i and gene g. Similarly, the disease–gene profile data are denoted as matrix $\mathbf{Z}=((z_{jg}))$ for disease j. Then, we define gene indicator variable $\mathbf{Ig}=\{Ig_g \in \{0,\cdots M\}; g=1,\cdots G\}$ to indicate the co-module that gene g belongs to; the drug indicator variable $\mathbf{Id}=\{Id_{mi} \in \{0,1\}; m=1,\cdots M; i=1,\cdots D\}$ to determine whether drug i belongs to co-module m or not and $\mathbf{Ip}=\{Ip_{mj} \in \{0,1\}; m=1,\cdots M; j=1,\cdots P\}$ to determine whether disease j belongs to co-module m. In co-module m, we model drug–gene profile values as a normal distribution formulized by following equation:

$$y_{ig} = \delta_{md} + d_{mi} + \alpha_{gd} + \epsilon_{gd}, \tag{2}$$

where δ_{md} is the co-module specific effect on drug–gene data; d_{mi} is the drug-specific factor in co-module m; α_{gd} is the gene-specific factor and ϵ_{gd} is a random noise. The subscript d indicates the specific parameters in the drug–gene data. For drug i' that does not belong to the co-module m,

$$y_{i'g} \sim N(\alpha_{gd}, \tau^2_{md}). \tag{3}$$

In the NULL module, we model the drug–gene profile distribution as

$$y_{ig} \sim N(\alpha_{gd}, \tau^2_{0d}). \tag{4}$$

Similar distributions are set for the disease–gene profile data.

We later define $\mathbf{Bd}=\{\beta d_{mi} = \delta_{md} + d_{mi}; m=1,\cdots M; i:\{Id_{mi}=1\}\}$ and $\mathbf{Bp}=\{\beta p_{mj} = \delta_{mp} + p_{mj}; m=1,\cdots M; j:\{Ip_{mj}=1\}\}$ as co-module center matrices. Joint posterior distribution for **Ig**, **Id**, **Ip** and **Bd**, **Bp** can be expressed as follows given $\mathbf{Y}, \mathbf{Z}$:

$$P(\mathbf{Ig},\mathbf{Id},\mathbf{Bd},\mathbf{Ip},\mathbf{Bp}|\mathbf{Y},\mathbf{Z}) \propto \int P(\mathbf{Y},\mathbf{Z},\mathbf{Bd},\mathbf{Bp}|\Theta,\mathbf{Ig},\mathbf{Id},\mathbf{Ip}) P(\Theta|\mathbf{Ig},\mathbf{Id},\mathbf{Ip}) \pi(\mathbf{Ig},\mathbf{Id},\mathbf{Ip}) d\Theta \tag{5}$$

We set conjugate priors for those parameters to integrate out those nuisance variables and derive the posterior distribution in a closed form. To penalize co-modules with large numbers of elements, we set a prior for indicator variables as follows:

$$\pi(\mathbf{Ig},\mathbf{Id},\mathbf{Ip}) \propto \exp\{-C_G \sum_{m=1}^{M} n^G_m - C_D \sum_{m=1}^{M} n^D_m - C_P \sum_{m=1}^{M} n^P_m\}, \tag{6}$$

where n^G_m, n^D_m, n^P_m are gene, drug and disease number, respectively, in co-module m and C_G, C_D, C_P are penalizing parameters.

MCMC strategies are adopted to construct a Markov chain that converges to the posterior distribution. The Gibbs sampler and Metropolis-Hasting algorithm are used to determine the transition moves of the chain. Parallel tempering (Gilks *et al.*, 1996; Liu, 2001; Zhang *et al.*, 2010) is also adopted to help the chain escape local modes. See Supplementary Material for details.

Association scores for computing ROC integral: we define the posterior association score AS_{ig} (AS_{jg}) between drug i (disease j) and gene g as the sum of the products of posterior indicator probabilities for i (j) and g across all co-modules except the NULL. Then, we define the drug–disease posterior association score AS_{ij} as the sum of the products of AS_{ig} and AS_{jg} across all genes (See Supplementary Material).

Poisson approximation for co-occurrence data: the co-occurrence data from the study of (Park *et al.*, 2009) records the frequencies for observing two diseases respectively and jointly. We adopt the same method to compute a P-value for the significance of co-morbidity of two diseases. Suppose I_{j1} and I_{j2} individuals are observed to have diseases j_1 and j_2, respectively, and C_{j12} individuals have both diseases. The expected number of individuals having both diseases is $C'_{j12} = I_{j1} * I_{j2}/N$ under an independence assumption, where N is the total number of individuals. A Poisson distribution is used to approximate this binomial model. Under this approximation, the P-value for co-morbidity of j_1 and j_2 is the probability of observing more individuals than C_{j12} having both diseases (See Supplementary Material).

ACKNOWLEDGEMENTS

We thank Prof. Jun Liu at Harvard University and Dr. Tang Wanwan in our laboratory for helpful discussions.

Funding: National Natural Science Foundation of China (Nos. 60934004 and 61021063); innovation scientific fund of Tsinghua University.

Conflict of Interest: none declared.

REFERENCES

Barabási,A. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Berger,S. *et al.* (2010) Systems pharmacology of arrhythmias. *Sci. Signal.*, **3**, ra30.

Bergmann,S. *et al.* (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E*, **67**, 31902–31920.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Chang,S. *et al.* (1996). Mutations in subunits of the epithelial sodium channel cause salt wasting with hyperkalaemic acidosis, Pseudohypoaldosteronism Type 1. *Nat. Genet.*, **12**, 248–253.

Chiang,A. and Butte,A. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.*, **86**, 507–510.

Chiu,C. *et al.* (2008) A novel single nucleotide polymorphism in ERCC6 gene is associated with oral cancer susceptibility in Taiwanese patients. *Oral Oncol.*, **44**, 582–586.

Chong,C. and Sullivan,D. (2007) New uses for old drugs. *Nature*, **448**, 645–646.

Chu,L. and Chen,B. (2008) Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC Syst. Biol.*, **2**, 56.

Davis,A. *et al.* (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids. Res.*, **37**, D786–D792.

Gilks,W. *et al.* (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Goh,K. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. U S A*, **104**, 8685–8690.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Heath,W. and Freis,E. (1963) Triamterene with hydrochlorothiazide in the treatment of hypertension. *JAMA*, **186**, 119–122.

Hopkins,A. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.

Huang,D. *et al.* (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Huber,M. *et al.* (2004) NF-κB is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression. *J. Clin. Invest.*, **114**, 569–581.

Hu,M. and Qin,Z. (2009) Query large scale microarray compendium datasets using a model-based bayesian approach with variable selection. *PloS One*, **4**, e4495.

Kishiokaa,H. *et al.* (2005) Pranlukast inhibits NFκB activation and MUC2 gene expression in cultured human epithelial cells. *Pharmacology*, **73**, 89–96.

Knowles,S. *et al.* (2005) Hydrochlorothiazide-induced noncardiogenic pulmonary edema: an underrecognized yet serious adverse drug reaction. *Pharmacotherapy*, **25**, 1258–1265.

Kutalik,Z. *et al.* (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.*, **26**, 531–539.

Lazarus,S. *et al.* (2001). Long-acting β2-agonist monotherapy vs continued therapy with inhaled corticosteroids in patients with persistent asthma. *J. Am. Med. Assoc.*, **285**, 2583–2593.

Lee,D. *et al.* (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl Acad. Sci. U S A*, **105**, 9880–9885.

Liu,J. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Li,S. *et al.* (2010) Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae. *BMC Bioinformatics*, **11**(Suppl. 11), S6.

Li, S., *et al.* (2011) Network target for screening synergistic drug combinations with application to traditional Chinese medicine. *BMC Syst. Biol.*, **5**(Suppl 1), S10.

McDowell,G. (2000) Comparative physiological study of arbutamine with exercise in humans. *Clin. Sci.*, **98**, 489–494.

McKusick,V. (2007) Mendelian inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

Mersmann,H. (1998) Overview of the effects of beta-adrenergic receptor agonists on animal growth including mechanisms of action. *J. Anim. Sci.*, **76**, 160–172.

Nicholson,R. *et al.* (2001) EGFR and cancer prognosis. *Eur. J. Cancer*, **37**, 9–15.

Park,J. *et al.* (2009) The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.*, **5**, 262.

Saini,R. *et al.* (1998) Tolerability and efficacy of fosinopril and hydrochlorothiazide compared with amiloride and hydrochlorothiazide in patients with mild to moderate hypertension. *Clin. Drug Invest.*, **15**, 91–99.

Schadt,E. *et al.* (2009) A network view of disease and compound screening. *Nat. Rev. Drug Discov.*, **8**, 286–295.

Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Shimkets,R. *et al.* (1994) Liddle's syndrome: heritable human hypertension caused by mutations in the β subunit of the epithelial sodium channel. *Cell*, **79**, 407–414.

Suthram,S. *et al.* (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.*, **6**, e1000662.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Wang,C. *et al.* (1981) The effect of triamterene and sodium intake on renin, aldosterone, and erythrocyte sodium transport in liddle's syndrome. *J. Clin. Endocrinol. Metab.*, **52**, 1027–1032.

Wang,C. *et al.* (1996) TNF-and cancer therapy-induced apoptosis: potentiation by inhibition of NF-κB. *Science*, **274**, 784–787.

Wang,L. *et al.* (2008) TNF-α induces two distinct caspase-8 activation pathways. *Cell*, **133**, 693–703.

Wishart,D. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids. Res.*, **36**, D901–D906.

Wong,D. *et al.* (2008) Revealing targeted therapy for human cancer by gene module maps. *Cancer. Res.*, **68**, 369–378.

Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

Yao,J. *et al.* (2007) Comparison of doxycycline and minocycline in the inhibition of VEGF-induced smooth muscle cell migration. *Neurochem. Int.*, **50**, 524–530.

Yildirim,M. *et al.* (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.

Zhang,W. *et al.* (2010) A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput. Biol.*, **6**, e1000642.

Zhao,S. and Li,S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS One*, **5**, e11764.