Quantitative Biology 2015, 3(3): 124–134
DOI 10.1007/s40484-015-0051-0

RESEARCH ARTICLE

Network-based method to infer the contributions of proteins to the etiology of drug side effects

Rui Li[1], Ting Chen[1,2,*] and Shao Li[1,*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China
[2] Program in Computational Biology and Bioinformatics, University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: shaoli@mail.tsinghua.edu.cn, tingchen@tsinghua.edu.cn

Received April 15, 2015; Revised July 21, 2015; Accepted September 24, 2015

Studying the molecular mechanisms that underlie the relationship between drugs and the side effects they produce is critical for drug discovery and drug development. Currently, however, computational methods are still unavailable to assess drug-protein interactions with the aim of globally inferring the contributions of various classes of proteins toward the etiology of side effects. In this work, we integrated data reflecting drug-side effect relationships, drug-target relationships, and protein-protein interactions to develop a novel network-based probabilistic model, *SidePro*, to evaluate the contributions of proteins toward the etiology of side effects. For a given side effect, the method applies an expectation–maximization algorithm and a diffusion kernel-based approach to estimate each protein's contribution. We applied this method to a wide range of side effects and validated the results using cross-validation and records from the Side Effect Resource database. We also studied a specific side effect, nephrotoxicity, which is known to be associated with the irrational use of the Chinese herbal compound triptolide, a diterpenoid epoxide in the Thunder of God Vine, *Tripterygium wilfordii* (*Lei-Gong-Teng*). Using triptolide as an example, we scored the target proteins of triptolide using our model and investigated the high-scoring proteins and their related biological processes. The results demonstrated that our model could differentiate between the potential side effect targets and therapeutic targets of triptolide. Overall, the proposed model could accurately pinpoint the molecular mechanisms of drug side effects, thus making contribution to safe and effective drug development.

Keywords: network pharmacology; drug targets; side effects; triptolide

INTRODUCTION

Drug side effects are unintended and usually undesirable consequences resulting from use of medications. Drugs exert side effects by interacting with molecular targets through protein-protein interactions (PPIs), disturbing related biological processes and ultimately causing the observed changes [1,2]. Despite extensive research investment in drug discovery, many experimental drugs fail in the clinical trial stages owing to drug safety concerns [3,4]. Even for drugs that gain FDA approval, side effects may occur during the post-market stages, causing severe health hazards [5,6]. Thus, studying the molecular mechanisms underlying drug side effects has become a very important issue in drug development [2,7]. However, to curb the incidence of side effects and facilitate the development of safe drugs, more efficient methods are needed to determine the relationships between side effects and proteins.

By the development of systems biology and network pharmacology, the integration of different types of datasets has become possible. Network perspectives and network-based approaches are powerful and are widely used in biomarker identification [8], drug discovery [9,10], and traditional Chinese medicine (TCM) research [11–13]. Using network-based drug side effect analysis, the relationship between side effects and chemical features [14], pathways [15] or Gene Ontology (GO)

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

biological processes [16] can be discovered, helping researchers understand the mechanisms underlying the manifestation of certain side effects exhibited by any given drug. Some methods have been developed to predict side effects by using drug properties [17–19], drug-target relationships [18,20], and network information [18,20] to construct machine learning-based models for trial drugs. Recently, studies have systematically evaluated the relationships between drug targets and side effects [21–23]. From such reports, it can be seen that drug-target relationships provide important information [18,20–22,24], PPI networks make it easier to predict drug side effects [20], and gene expression datasets help to gain insight into the mechanism of drug side effects [15,16]. Although side effects are generated at the molecular level [1,2], methods that use drug-target and network information to infer the contributions of proteins to the etiology of side effects have, to the best of our knowledge, never been developed.

Based on the molecular mechanism of side effects [1,2], we have developed a network-based method, termed SIDE effects of PROteins, or *SidePro*, to measure the contributions of proteins in a PPI network to a side effect. We propose that each protein makes some contribution to each corresponding side effect, and we express the contribution of each protein as a score. We use an expectation-maximization (EM)-based method to give scores for targets directly connected to drugs based on drug-side effect observations and drug-target relationships, and a kernel-based method is used to evaluate the scores of non-targets based on PPIs.

We applied our method to a wide range of side effects, and each protein was assigned a score for its contribution to the corresponding side effect. The side effect proteins for 86% of side effects in the Side Effect Resource (SIDER) database [25] had significantly high scores with our method, which, in turn, validated the scores obtained for the proteins. The scores were further validated by using cross-validation on drug-side effect observations and by comparing the results with a support vector machine (SVM) prediction model [20]. The results show that our method achieves better performance.

In recent years, the so-called "Chinese herb nephropathy" has received considerable attention, but the molecular mechanisms remain unclear [26]. For instance, triptolide, an active compound in *Tripterygium wilfordii* (*Lei-Gong-Teng* in Chinese), is reported to cause nephrotoxicity if irrational use [27,28]. When we applied *SidePro* to investigate the proteins associated with renal side effect, as well as triptolide, we were able to identify the potential responsible proteins interacting with triptolide and, in turn, the resulting nephrotoxicity. *SidePro* could also differentiate between side effect targets and therapeutic targets of triptolide, showing its promise as a method to eliminate dangerous targets during drug development.

Developing a computational method to find specific proteins responsible for side effects will help us to understand the etiology of side effects resulting from unintended drug-protein interactions, affording a tool able to guide the choice of safe drug targets during drug discovery and development.

RESULTS

Data extraction

Drug-side effect relationships were extracted from SIDER [25], and drug-target relationships were extracted from the DrugBank database [29]. For the natural small-molecule compound triptolide, which was not recorded in DrugBank, the drug-target relationships were obtained from the STITCH database [30]. PPI datasets were integrated from five databases, including the Human Protein Reference Database [31], the Biomolecular Interaction Network Database [32], the IntAct Database [33], the Molecular Interaction Database [34] and the Online Predicted Human Interaction Database [35], to obtain the maximum connected subnetwork, which contained 137,008 non-redundant PPIs for 13,337 human proteins. As a result, we obtained 645 drugs with their corresponding recorded side effect relationships. To reduce the influence of sample imbalance, *SidePro* was conducted on all side effects with the number of positive drugs larger than 100.

Performance of *SidePro*: retrieval of proteins causing side effects

For each given side effect, *SidePro* gives a score for each protein by measuring its contribution to the side effect, thus producing a ranked list of proteins. We used protein-side effect relationships for each side effect from SIDER to validate the results. For each side effect, we used GSEA to determine if the proteins given by SIDER had high scores by our algorithm. The results showed that the proteins given by SIDER had high scores for 86% of the side effects with a significance of $P < 0.05$, as shown in Figure 1A. The side effect-causing proteins in SIDER were ranked significantly high in our algorithm.

Performance of *SidePro*: prediction of drug-side effect

Testing the performance of drug-side effect prediction is another way to validate the accuracy of *SidePro*. Given the drug-protein relationships, as well as the scores of proteins for the side effect, we predict the scores of drug-

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

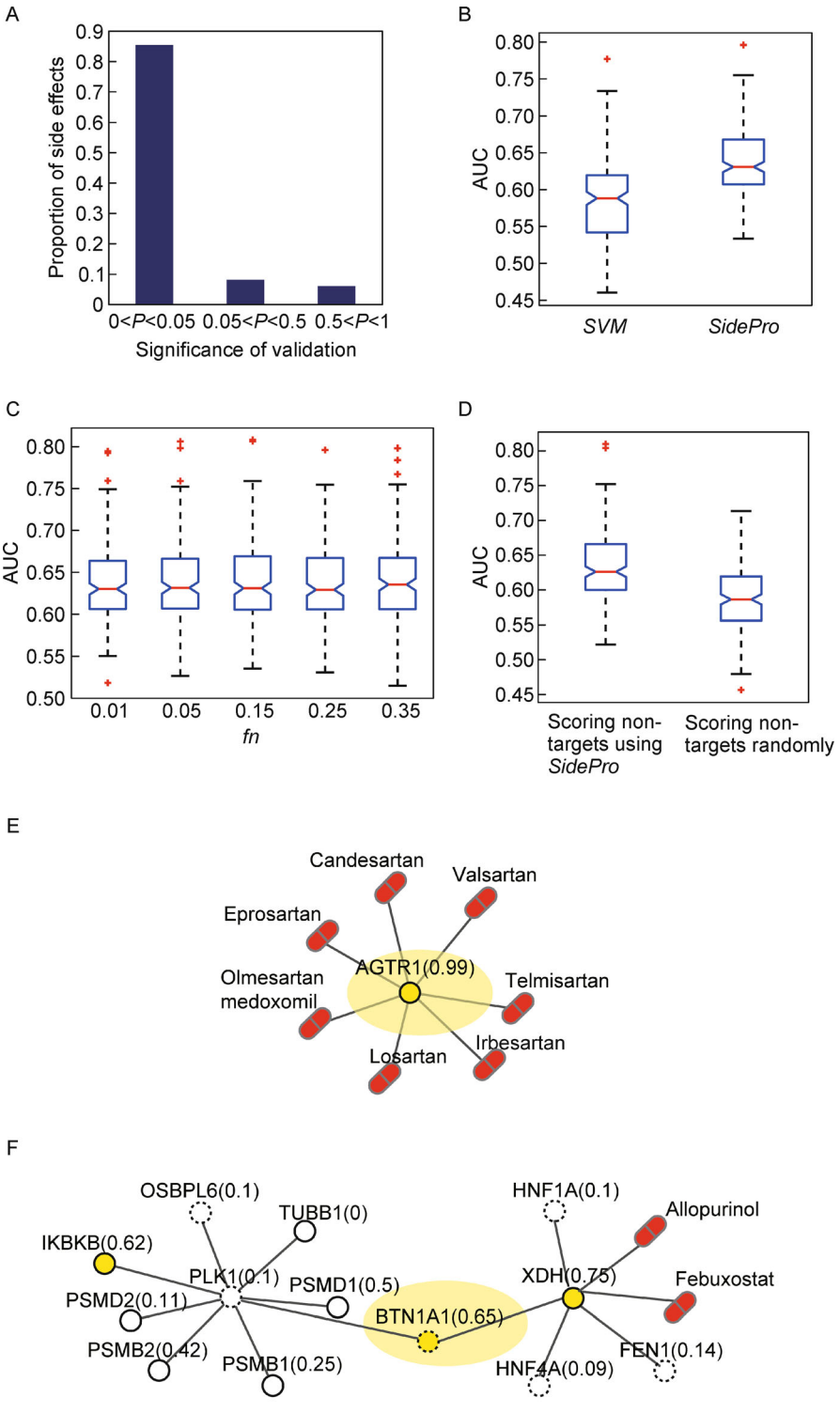

Figure 1. The performance of *SidePro*. (A) 86% of the protein-side effect relationships from a side effect database, SIDER, were calculated as significant ($P < 0.05$) by *SidePro*. (B) Comparison of performance of drug side effect prediction between *SidePro* and SVM. (C) Influence of the parameter *fn*. (D) Difference of performance by measuring versus not measuring non-targets. (E) An example of inferred results around a target protein AGTR1. (F) An example of inferred results around a non-target protein BTN1A1. For E and F, numbers in parentheses are the inferred scores for proteins. The cutoff score for proteins causing the side effect is set to be 0.5.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

side effect relationships using Equation 2. Thus we compared our method with the SVM prediction model [20] using a10-fold cross-validation by leaving the drug-side effect relationships out. The AUC (Area Under the Curve) values were compared across different side effects, and paired t-test was used to compare the AUC values for all of the side effects. Overall, our method achieved higher AUC values than the approach by the SVM prediction model (Figure 1B, $P < 0.001$).

Performance of *SidePro*: using the PPI network

We conducted the following simulation to investigate the importance of evaluating the contributions of non-targets to side effects using PPI network information. For a given side effect, 10% percent of the target proteins were chosen randomly, and the corresponding drug-target relationships were deleted from the training set. Thus, these target proteins became non-targets during the training, and their scores causing the side effect would have remained unknown if we had not evaluated them. However, using *SidePro*, we were able to integrate the PPI network to predict the scores for these deleted proteins. Thus, we compared the following two strategies for scoring non-targets: scoring non-targets using the PPI network by a diffusion kernel-based method as in *SidePro* versus scoring non-targets randomly. For each side effect, a 10-fold cross-validation was conducted, and the boxplot of the AUC of all the side effects under the two circumstances was plotted in Figure 1D. It could be seen from the figure that the performance of *SidePro* did not significantly suffer by the loss of these target proteins, in turn demonstrating the significance of using the PPI network in *SidePro*. On the other hand, the performance of scoring non-targets randomly drops significantly lower.

Performance of *SidePro*: example of proteins causing renal failure

To help understand why nephropathy caused by a Chinese herb has aroused extensive attention in recent years, we demonstrated *SidePro* with the side effect of renal failure as an example.

To make the output more visual and easier to understand, we demonstrated the inferred results with a target protein and a non-target protein. In the evaluation of each protein, our method considered the information of the global network. For the convenience of the display, only parts of the surrounding neighbors of the node under consideration were plotted. Figure 1E shows the inferred results for the target AGTR1. Seven drugs target AGTR1, all of which cause renal failure. Based on drug-target relationships and drug-side effect observations for AGTR1, which is known to be involved in renal pathogenesis [36], it was assigned with a high score. As shown in Figure 1F, BTN1A1, a non-target protein, was given a relatively high score of 0.65, mainly because its neighbors target XDH, while, in contrast, PLK1 was assigned a relatively lower score because of the low score of neighboring proteins. As shown in Figure 1E and 1F, the obtained scores accurately reflect the global topological information of the network.

For renal failure as a side effect, the top 1% of the high-scoring proteins was used to analyze GO biological processes and pathway enrichment. As shown in Table 1, the identified proteins were enriched in renal failure-related GO terms and pathways. Those pathways were also recorded as related to renal failure in the Comparative Toxicogenomics Database (CTD). It could be shown that these proteins cause renal failure by acting on the key biological processes involved in pathological kidney changes, which, in turn, supports the effectiveness of our method.

Inferring proteins causing renal failure as a side effect of a herbal compound triptolide

Using the protein scores obtained for renal failure as a side effect and the targets of triptolide, triptolide was successfully predicted to cause renal failure with a score of 0.99. Furthermore, among the targets of triptolide, BCL2 and PTGS2 were prioritized as the responsible targets. BCL2 plays important roles in the regulation of apoptosis and hypoxia-related apoptosis [37]. Studies have shown that abnormal changes of BCL2 might induce apoptosis and cause damage to renal cells [38,39]. PTGS2 (COX2) also participates in oxidative- and apoptosis-related biological processes. It has been widely reported that abnormal PTGS2 expression may be related to renal disease [40,41]. This evidence supports the idea that prioritized BCL2 and PTGS2 may be candidate proteins enabling triptolideto elicit renal toxicity. The biological processes related to renal failure, in which BCL2 and PTGS2 participate, were shown in Figure 2B. By targeting BCL2 and PTGS2, triptolide may interfere with those renal biological processes, which, in turn, causes the observed side effect.

Triptolide is a novel anti-inflammatory and immunosuppressive agent. Combined with evidence from the literature and GO biological information, PTGS2, TNF, NFKB1, and IL2 were identified among the targets of triptolide and were determined to participate in the therapeutic process-related inflammatory response.

DISCUSSION

This article reports an algorithm, *SidePro*, designed to identify proteins that contribute to the etiology of side

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

Table 1. Enriched renal failure-related GO terms/pathways of renal failure proteins by *SidePro*.

Category	Enriched GO terms/pathways	*P*-value
GO biological processes	Renal system process	0.00213
	Renal system process involved in regulation of systemic arterial blood pressure	0.0629
	Renal control of peripheral vascular resistance involved in regulation of systemic arterial blood pressure	0.0274
	Kidney development	0.0592
	Regulation of blood volume by renin-angiotensin	0.00228
	Regulation of systemic arterial blood pressure by renin-angiotensin	0.00441
	Regulation of blood vessel size by renin-angiotensin	0.027466
	Regulation of systemic arterial blood pressure by circulatory renin-angiotensin	0.07159
	Response to oxidative stress	0.018
	Regulation of apoptosis	0.0029
Pathways	Renin-angiotensin system	0.0347
	Renal cell carcinoma	0.0334
	mTOR signaling pathway	0.0125
	MAPK signaling pathway	0.00644
	B cell receptor signaling pathway	0.0416
	VEGF signaling pathway	0.00185
	ErbB signaling pathway	0.0176
	GnRH signaling pathway	0.00704

effects during drug therapy. For a given side effect, the method applies a network-based approach to estimate each protein's contribution. Proteins with high scores were identified from the PPI network, demonstrating the performance of *SidePro*. Then, protein scores were validated by using a database and computing drug-side effect relationships. Furthermore, we studied a specific side effect called renal failure, which is known to be associated with the Chinese herbal compound triptolide. We scored every target protein of triptolide using our model and investigated the high-scoring proteins and their related biological processes. Our model could help distinguish between side effect targets and therapeutic targets.

SidePro takes aim at the mechanism(s) by which drug-protein interaction results in certain side effects [2] by assigning contribution factors to each protein. *SidePro* integrates information drawn through observation of drug-side effect relationships, drug-target relationships, and PPIs at the system's level, and it gives clear and comprehensive results at the molecular level. *SidePro* presents results that are biologically pictorial and interpretable. To score each protein-side effect relationship in a probabilistic manner, *SidePro* employs Maximum Likelihood Estimation and network topological measurements based on observations of drug-side effect relationships. This model could be further used to predict the probability of drugs causing the side effect. Different probabilities of proteins represent different contributions to the side effect. Thus for a given drug, the proteins and biological processes involved in side effects could be inferred (Figure 2).

Drug-side effect relationships differ between observation and reality, and no databases record drugs that do not cause side effects. *SidePro*, however, takes all of this into account by introducing the parameters *fp* and *fn*. With *fn*, our method could use the drugs in databases not annotated as causing the side effect as negative drugs.To the best of our knowledge, *SidePro* evaluates, for the first time, the whole set of proteins relative to the contributions made to side effects in the PPI network. *SidePro* also evaluates the contributions of non-target proteins for the following reasons. First, side effects normally begin when a drug binds to its targets, but while some targets are therapeutic, others may be off-target proteins, and binding with these proteins can result in the development of unwanted side effects. Still other non-target proteins that may also play some part in this process. Essentially, it is protein-protein interactions that elicit signaling from a drug which ultimately causes the side effect. Second, during drug design, some new proteins suitable as therapeutic targets may be developed; therefore, it is beneficial to evaluate the contributions of non-targets to side effects. By simulation study in Figures 1 and 3, results show the rationale for measuring non-targets.

Target records exist for many drugs, such as those in

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

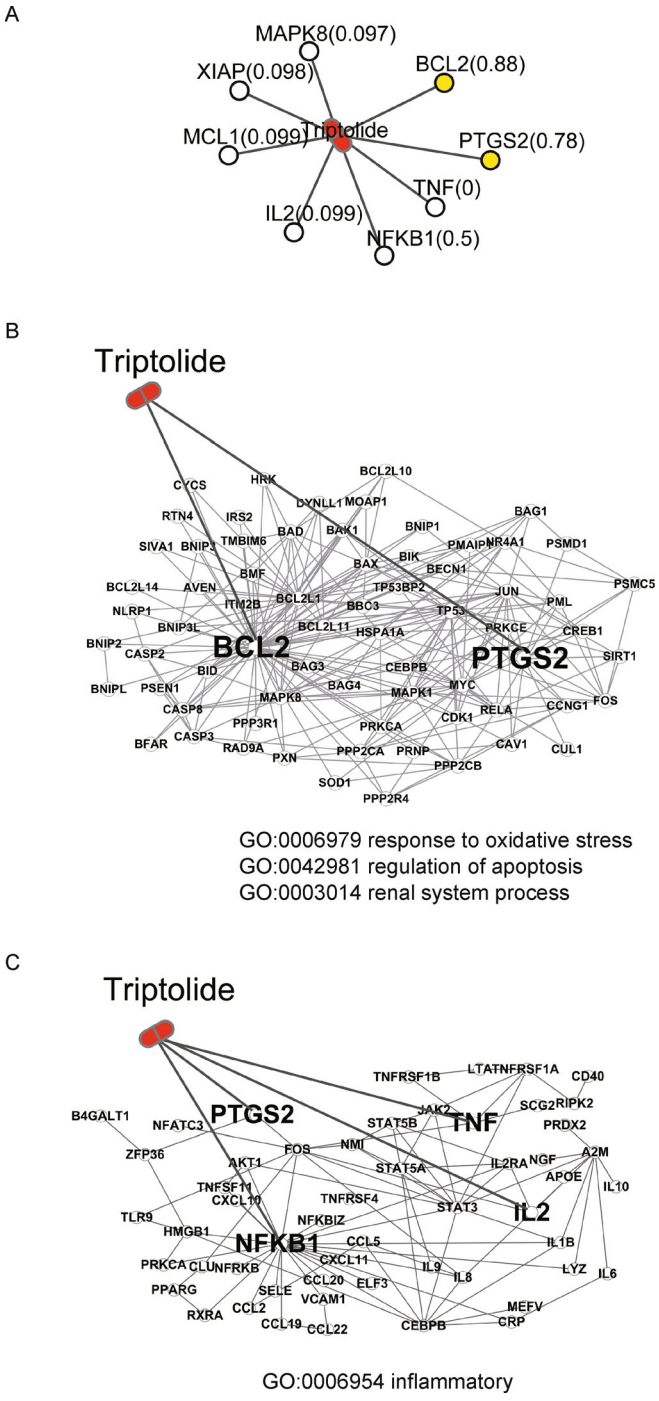

Figure 2. The inferred proteins associated with the side effect of renal failure and the anti-inflammatory efficacy exerted by the Chinese herbal compound triptolide. (A) Inferred results for nephrotoxicity caused by triptolide. (B) Nephrotoxicity-related processes. (C) Anti-inflammatory-related processes.

DrugBank, but it is difficult to understand the roles these targets play, particularly in the case of drugs that exert both therapeutic effects and side effects via different targets, making the underlying mechanisms of side effects difficult to interpret. Drugs or compounds with multiple targets fall into this category. Therefore, as an example, we studied a specific side effect, nephrotoxicity, which is known to be associated with symptoms of nephropathy arising from the irrational use of the herbal compound triptolide, a diterpenoid epoxide extracted from *Tripter-*

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

ygium wilfordii. For triptolide, we scored the target proteins of triptolide using our model and investigated the high-scoring proteins and their related biological processes. The results demonstrated that our model was able to differentiate between the side effect targets and therapeutic targets of triptolide (Figure 2), providing some clues to guide the rational and precision use of this compound. Thus, by using our algorithmic framework, we could potentially distinguish between targets that are toxic and those that are therapeutic, making it possible to choose safe proteins as targets, while avoiding those that are dangerous.

By assigning a contribution metric for proteins in relation to the etiology of side effects, the algorithm demonstrates many merits. First, for side effects, we are able to determine which target and non-target proteins are important out of the thousands of proteins. Second, most drugs have more than one targets, and the obtained results indicate all proteins interacting with the drug of interest, finally resulting in the observed side effects. Third, for drugs in ongoing clinical trials, we can predict side effects based on drug-target relationships and the probabilities of targets. Thus, by using existing knowledge in our algorithm during drug development, we are able to predict the risks of potential side effects before choosing targets, which is important for drug safety. Such identification efforts could be used to aid in designing or selecting drugs that target proteins with improved side effect profiles in future drug development efforts.

We also need to emphasize that the drug-side effect relationships are much more complex than what we have assumed in this study. We do not consider how dose of a drug affects its side effects, the quantitative and the tissue specific information of targets and PPI, and neither do we consider drug tolerability and genetic background of each individual, which plays a critical role of the side effects. Nevertheless, this study presents an important step toward our understanding of drug-side effect relationships.

METHODS

Outline of *SidePro*

As shown in Figure 3, *SidePro* employs the following three steps to evaluate the contributions of proteins to the etiology of a side effect.

First, for a given side effect, all drugs are classified into two groups: one observed with the side effect and the other without.

Second, the scores for drug-target proteins are evaluated by globally integrating information from two sources: drug-side effects and drug-target relationships. The scores for other non-target proteins are evaluated based on their connections with target proteins in the protein-protein interaction network.

Third, proteins with high scores are considered side effect-causing candidates, and those proteins are subjected to further analysis.

Inferring the side effect scores of target proteins

The score produced by *SidePro* was defined as follows. Given a specific side effect, define $P_j = 1$ if protein P_j causes the side effect and $P_j = 0$ otherwise. Let λ_j denote the probability that protein P_j causes the side effect:

$$\lambda_j = \Pr(P_j = 1). \tag{1}$$

Let D_i denote a drug. $D_i = 1$ if drug D_i causes the side effect, and $D_i = 0$ otherwise. We call a drug causing the side effect a positive drug and a drug not causing the side effect a negative drug. Let T_{ij} denote the *j-th* target (protein) of drug D_i. $T_{ij} = 1$ if drug D_i interacts with protein P_j and $T_{ij} = 0$ otherwise.

To evaluate the contributions of the target proteins to the side effect, we maximize the likelihood of the observation of drug-side effect relationships. Thus we propose a Maximum Likelihood Estimation method with an EM algorithm to infer target-side effect probabilities based on data about drug-target and drug-side effect relationships.

Assumption: We assume that a given drug causes a side effect if and only if at least one protein targeted by the drug causes the side effect.

Under this assumption, we obtain

$$\Pr(D_i = 1) = 1 - \prod_{j:=1}^{m} (1 - \Pr(P_j = 1))^{T_{ij}} = 1 - \prod_{j:=1}^{m} (1 - \lambda_j)^{T_{ij}}, \tag{2}$$

where m is the number of target proteins.

However, an observed side effect can be caused by some factor other than the drug itself. Therefore, we need to take into account both test errors and observation errors in a clinical study. Since a conclusion drawn from the clinical study through statistical testing may not be true, we introduce two types of errors: false positives (*fp*), in which a positive drug-side effect relationship was observed (or established), but in reality does not occur, and false negatives (*fn*), in which a positive drug-side effect relationship was not reported (established), but in reality does occur.

Let O_i denote the observed drug-side effect relationship for drug D_i, where $O_i = 1$ if a positive relationship is observed and $O_i = 0$ otherwise. Then,

$$fp = \Pr(O_i = 1 | D_i = 0),$$

$$fn = \Pr(O_i = 0 | D_i = 1).$$

Thus, the following equation can be derived:

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

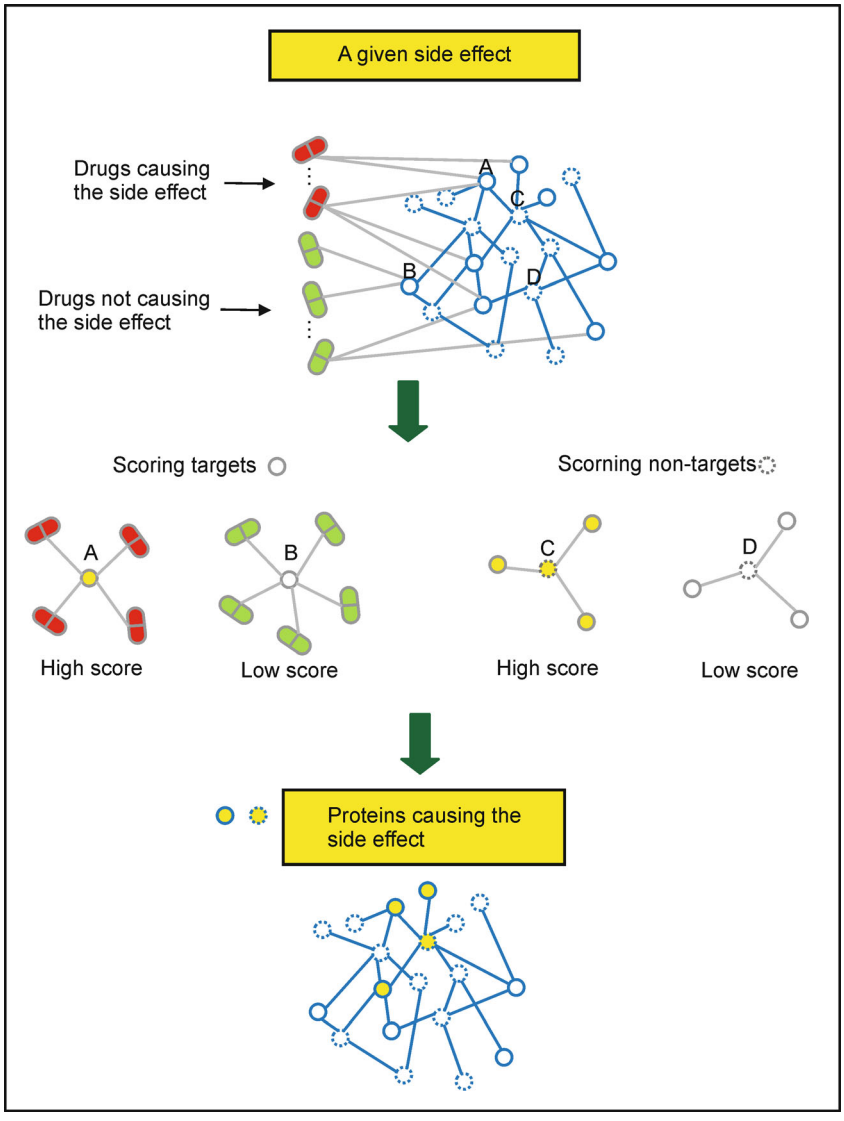

Figure 3. The Workflow of *SidePro*. For a given side effect, drugs causing the side effect and drugs not causing the side effect were collected and denoted with red capsules and green capsules, respectively. Proteins that are targets of those drugs are denoted with solid circle, and non-target proteins are denoted with dotted line. Edges between drugs and proteins denote drug-target interactions, while those between proteins and proteins denote protein-protein interactions. The representation of color, shape and interactions is consistent for all the figures in this work.

$$
\begin{aligned}
&\Pr(O_i=1)\\
&=\Pr(D_i=1)\Pr(O_{i=1}|D_i=1)+(1-\Pr(D_i=1))\\
&\quad\Pr(O_i=1|D_i=0)\\
&=\Pr(D_i=1)(1-fn)+(1-\Pr(D_i=1))fp\\
&=\left[1-\prod_{j:=1}^{m}(1-\lambda_j)^{T_{ij}}\right](1-fn)+\left[\prod_{j:=1}^{m}(1-\lambda_j)^{T_{ij}}\right]fp\\
&=1-fn-(1-fn-fp)\left[\prod_{j:=1}^{m}(1-\lambda_j)^{T_{ij}}\right]. \qquad (3)
\end{aligned}
$$

The likelihood function, i.e., the probability of the observed relationships between a specific side effect and all drugs, is given by

$$L=\prod_{i=1}^{n}(\Pr(O_i=1))^{O_i}(1-\Pr(O_i=1))^{1-O_i}, \qquad (4)$$

where n is the number of drugs.

The likelihood L is a function of $\theta=(\{\lambda_j\},fp,fn)$. In the following steps, we fix fp and fn, and we develop the following EM algorithm to estimate each λ_j. The EM algorithm distinguishes the observed data Y from the complete data Z. In the expectation (E) step, we calculate the expectation of the complete data Z given the observed

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

data Y, $\hat{Z} = E(Z \mid Y, \theta^{(t-1)})$. In the maximization (M) step, we obtain the MLE of θ, $\theta^{(t)}$, based on $\hat{Z}$. Thus we obtain a recursive formula to estimate θ.

In our case, the observed data are the observed drug-side effect relationships $O = \{o_i, i = 1, \cdots, n\}$, and the complete data include all target-side effect relationships for each drug-side effect relationship $Z = \{O,P\}$, where O is given above and $P = \{P_j^{(i)} \mid T_{ij} = 1\}$. $P_j^{(i)} = 1$ if protein P_j is a target for drug D_i and causes the side effect and $P_j^{(i)} = 0$ otherwise.

The E-step is calculated as follows:

$$\begin{aligned} & E(P_j^{(i)} | O_k = o_k, \forall k, \theta^{(t-1)}) \\ &= E(P_j^{(i)} | O_i = o_i, \theta^{(t-1)}) \\ &= \frac{\Pr(P_j^{(i)} = 1, O_i = o_i | \theta^{(t-1)})}{\Pr(O_i = o_i | \theta^{(t-1)})} \\ &= \frac{\Pr(P_j^{(i)} = 1 | \theta^{(t-1)}) \Pr(O_i = o_i | P_j^{(i)} = 1, \theta^{(t-1)})}{\Pr(O_i = o_i | \theta^{(t-1)})} \\ &= \frac{\lambda_j^{(t-1)} (1 - fn)^{o_i} fn^{1-o_i}}{\Pr(O_i = o_i | \theta^{(t-1)})}, \end{aligned} \tag{5}$$

where the denominator can be calculated by Equation 3. Let Λ_j be the set of drugs for which protein P_j is a target. The MLE of λ_j is the fraction of the drugs in Λ_j for which target protein P_j causes the side effect, $P_j^{(k)} = 1$. We thus obtain the recursive formula for the M-step:

$$\begin{aligned} \lambda_j^{(t)} &= \frac{1}{|\Lambda_j|} \sum_{k \in \Lambda_j} E(P_j^{(k)} | O_k = o_k, \forall k, \theta^{(t-1)}) \\ &= \frac{\lambda_j^{(t-1)}}{|\Lambda_j|} \sum_{k \in \Lambda_j} \frac{(1 - fn)^{o_k} fn^{1-o_k}}{\Pr(O_k = o_k | \theta^{(t-1)})}. \end{aligned} \tag{6}$$

The EM algorithm is performed as follows:

i. Choose initial values for $\{\lambda_j, j = 1, \cdots, m\}$ and compute the probabilities $\Pr(D_i = 1)$ by Equation 2 and $\Pr(O_i = 1)$ by Equation 3.

ii. Update $\{\lambda_j, j = 1, \cdots, m\}$ by Equation 6 and compute the likelihood function by Equation 4.

iii. Repeat Step (ii) until the value of the likelihood function is unchanged (within certain error).

Inferring the scores of non-target proteins using PPIs

The score of a non-target protein P_u is evaluated based on the scores of all target proteins and the topological relationships of those targets in the PPI network, which is calculated as

$$\begin{aligned} \Pr(P_u = 1) = & (\lambda_1 K_{1,u} + \lambda_2 K_{2,u} + \cdots + \lambda_n K_{n,u}) \\ & / (K_{1,u} + K_{2,u} + \cdots + K_{n,u}), \end{aligned} \tag{7}$$

where λ_1 to λ_n denotes the scores of the all target proteins, and $K_{j,u}$ denotes the influence of drug target protein P_j on protein P_u based on the topology of the PPI network, which can be evaluated by using a diffusion kernel. The diffusion kernel gives a value between each pair of nodes for a network by considering the global information about the network topology. We calculate the diffusion kernel according to the steps of Kondor et al. [42]. A value of 0.5 is used for the diffusion constant.

Parameters of *SidePro*

There are two parameters in the model: *fn* and *fp*. The parameter *fp* indicates that drugs do not cause the side effect although positive drug-side effect relationships are observed; in practice, this should be a low value. Therefore, *fp* was given a relatively low value of 0.0001. Because the parameter *fn* indicates that drugs do cause the side effect, even though positive drug-side effect relationships are not observed, we roughly estimate its value from the old (SIDER 1) and new (SIDER2) versions of SIDER, a database recording drug-side effect relationships. We obtained drug-side effect relationships from SIDER 1 based on the Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) vocabulary and from SIDER 2 based on the Medical Dictionary for Regulatory Activities (MedDRA) vocabulary, respectively. For the common set of side effects, the number of drug-side effect relationships was counted for SIDER 1 and SIDER 2, respectively. The *fn* was calculated as the increase in the number of drug-side effect relationships from SIDER1 to SIDER2 divided by the number of drug-side effect relationships in SIDER 2, which is about 0.13.

We also tested the performance of *SidePro* for five values of *fn*: 0.01, 0.05, 0.15, 0.25 and 0.35. The area under the curve (AUC) of the 10-fold cross-validation was calculated, and a one-way analysis of variance (ANOVA) showed that the performance did not differ significantly for different *fn* values (Figure 1C), which means that our method is robust to changes in the *fn* parameter. The value of *fn* was finally set to 0.15.

Validations and comparisons

By the absence of a large standard dataset of protein-side effect relationships, we adopted two strategies to validate the performance of *SidePro*. We extracted protein-side effect relationships from the SIDER database by Kuhn et al. [22] for validation. In addition, we calculated the statistical significance (*P*-value) of the protein-side effect

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

relationships from the SIDER database that had high scores in *SidePro* using the Gene Set Enrichment Analysis (GSEA) test [43], a method that measures the statistical significance of a set's ranking in another ranked list.

To further validate the performance of *SidePro*, we predicted drug-side effect relationships and compared our results with an SVM-based method developed by Huang et al. [20], who also used the information of drug-target relationships and PPIs. The SVM approach was conducted as follows. First, drug-target relationships were expanded using the PPIs; that is, both the target proteins and their direct neighbors were considered. Then, for each drug, the number of targets was counted and listed as a feature. Second, a rank sum test was used to select features with P-value < 0.05. Third, an SVM approach with probabilistic outputs [44,45] was conducted using the selected features to predict drug-side effect relationships. The performance comparison between the SVM approach and *SidePro* for each side effect was based on the same set of drug-side effect relationships, drug-target relationships, and PPIs through a 10-fold cross-validation. The AUC of the receiver operator characteristic (ROC) curve was calculated as a performance indicator.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (Nos. 81225025 and 91229201), and Tsinghua National Laboratory of Information Science and Technology (TNLIST) Big Data Grant.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Rui Li, Ting Chen and Shao Li declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Scheiber, J., Chen, B., Milik, M., Sukuru, S. C. K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., et al. (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. J. Chem. Inf. Model, 49, 308–317
2. Berger, S. I. and Iyengar, R. (2011) Role of systems pharmacology in understanding drug adverse events. Wiley Interdiscip. Rev. Syst. Biol. Med., 3, 129–135
3. Roses, A. D. (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. Nat. Rev. Genet., 5, 645–656
4. Stevens, J. L. and Baker, T. K. (2009) The future of drug safety testing: expanding the view and narrowing the focus. Drug Discov. Today, 14, 162–167
5. Shah, R. R. (2006) Can pharmacogenetics help rescue drugs withdrawn from the market? Pharmacogenomics, 7, 889–908
6. Zhang, W., Roederer, M. W., Chen, W.-Q., Fan, L. and Zhou, H.-H. (2012) Pharmacogenetics of drugs withdrawn from the market. Pharmacogenomics, 13, 223–231
7. Liebler, D. C. and Guengerich, F. P. (2005) Elucidating mechanisms of drug-induced toxicity. Nat. Rev. Drug Discov., 4, 410–420
8. Li, R., Ma, T., Gu, J., Liang, X. and Li, S. (2013) Imbalanced network biomarkers for traditional Chinese medicine Syndrome in gastritis patients. Sci. Rep., 3, 1543
9. Zhao, S. and Li, S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification. PLoS One, 5, e11764
10. Zhang, B., Wang, X. and Li, S. (2013) An integrative platform of TCM network pharmacology and its application on an herbal formula, Qing-Luo-Yin. Evid-Based Compl. Alt. Med., 2013, 456747
11. Li, S., Zhang, B., Jiang, D., Wei, Y. and Zhang, N. (2010) Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae. BMC Bioinformatics, 11, S6
12. Li, S., Zhang, B. and Zhang, N. (2011) Network target for screening synergistic drug combinations with application to traditional Chinese medicine. BMC Syst Biol, 5, S10
13. Liang, X., Li, H. and Li, S. (2014) A novel network pharmacology approach to analyse traditional herbal formulae: the Liu-Wei-Di-Huang pill as a case study. Mol. Biosyst., 10, 1014–1022
14. Scheiber, J., Jenkins, J. L., Sukuru, S. C. K., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J., Urban, L., et al. (2009) Mapping adverse drug reactions in chemical space. J. Med. Chem., 52, 3103–3107
15. Lee, S., Lee, K. H., Song, M. and Lee, D. (2011) Building the process-drug-side effect network to discover the relationship between biological processes and side effects. BMC Bioinformatics, 12, S2
16. Wallach, I., Jaitly, N. and Lilien, R. (2010) A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathways. PLoS One, 5, e12063
17. Atias, N. and Sharan, R. (2011) An algorithmic framework for predicting side effects of drugs. J. Comput. Biol., 18, 207–218
18. Huang, L. C., Wu, X. and Chen, J. Y. (2013) Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. Proteomics, 13, 313–324
19. Yamanishi, Y., Pauwels, E. and Kotera, M. (2012) Drug side-effect prediction based on the integration of chemical and biological spaces. J. Chem. Inf. Model, 52, 3284–3292
20. Huang, L. C., Wu, X. and Chen, J. Y. (2011) Predicting adverse side effects of drugs. BMC Genomics, 12, S11
21. Mizutani, S., Pauwels, E., Stoven, V., Goto, S. and Yamanishi, Y. (2012) Relating drug-protein interaction network with drug side effects. Bioinformatics, 28, i522–i528
22. Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L. J., Gross, C., Gavin, A.-C. and Bork, P. (2013) Systematic identification of proteins that elicit drug side effects. Mol. Syst. Biol., 9, 663
23. Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., et al. (2012) Large-scale prediction and testing of drug activity on side-effect targets. Nature, 486, 361–367
24. Wallach, I., Jaitly, N. and Lilien, R. (2010) A structure-based approach for mapping adverse drug reactions to the perturbation of underlying biological pathways. PLoS One, 5, e12063
25. Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. and Bork, P. (2010) A side effect resource to capture phenotypic effects of drugs. Mol. Syst. Biol., 6, 343
26. Vanherweghem, J. L., Depierreux, M., Tielemans, C., Abramowicz, D.,

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

Dratwa, M., Jadoul, M., Richard, C., Vandervelde, D., Verbeelen, D., Vanhaelen-Fastre, R., et al. (1993) Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs. Lancet, 341, 387–391

27. Allard, T., Wenner, T., Greten, H. J. and Efferth, T. (2013) Mechanisms of herb-induced nephrotoxicity. Curr. Med. Chem., 20, 2812–2819
28. Nowack, R., et al. (2011) Herbal treatments of glomerulonephritis and chronic renal failure: Review and recommendations for research. J. Pharm. Phyt., 3, 124–136
29. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res., 39, D1035–D1041
30. Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J. and Bork, P. (2012) STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res., 40, D876–D880
31. Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K., Gronborg, M., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res., 13, 2363–2371
32. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. and Hogue, C. W. (2001) BIND — The Biomolecular Interaction Network Database. Nucleic Acids Res., 29, 242–245
33. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res., 40, D841–D846
34. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res., 40, D857–D861
35. Brown, K. R. and Jurisica, I. (2005) Online predicted human interaction database. Bioinformatics, 21, 2076–2082
36. Szaszák, M., Chen, H.-D., Chen, H.-C., Baukal, A., Hunyady, L. and Catt, K. J. (2008) Identification of the invariant chain (CD74) as an angiotensin AGTR1-interacting protein. J. Endocrinol., 199, 165–176
37. Basile, D. P., Liapis, H. and Hammerman, M. R. (1997) Expression of bcl-2 and bax in regenerating rat renal tubules following ischemic injury. Am. J. Physiol., 272, F640–F647
38. Zhou, H., Miyaji, T., Kato, A., Fujigaki, Y., Sano, K. and Hishida, A. (1999) Attenuation of cisplatin-induced acute renal failure is associated with less apoptotic cell death. J. Lab. Clin. Med., 134, 649–658
39. Qiu, L.-Q., Sinniah, R. and I-Hong Hsu, S. (2004) Downregulation of Bcl-2 by podocytes is associated with progressive glomerular injury and clinical indices of poor renal prognosis in human IgA nephropathy. J. Am. Soc. Nephrol., 15, 79–90
40. Harris, R. C. (2006) COX-2 and the kidney. J. Cardiovasc. Pharmacol., 47, S37–S42
41. Fujihara, C. K., Antunes, G. R., Mattar, A. L., Andreoli, N., Malheiros, D. M., Noronha, I. L., Zatz, R. and Zatz, R. (2003) Cyclooxygenase-2 (COX-2) inhibition limits abnormal COX-2 expression and progressive injury in the remnant kidney. Kidney Int., 64, 2172–2181
42. Kondor, R. I. and Lafferty, J. D. (2002) Diffusion kernels on graphs and other discrete input spaces. Proceedings, ICML, 2, 315–322
43. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA, 102, 15545–15550
44. Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol., 2, 27
45. Lin, C.-J. and Weng, R. C. (2004) Simple probabilistic predictions for support vector regression. Technical report, National Taiwan University, Taipei

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2015