Bioinformatics, 31(15), 2015, 2523–2529
doi: 10.1093/bioinformatics/btv181
Advance Access Publication Date: 29 March 2015
Original Paper

Systems biology

Global optimization-based inference of chemogenomic features from drug–target interactions

Songpeng Zu[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 and [2]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 29, 2014; revised on February 22, 2015; accepted on March 24, 2015

Abstract

Motivation: Gaining insight into chemogenomic drug–target interactions, such as those involving the substructures of synthetic drugs and protein domains, is important in fragment-based drug discovery and drug repositioning. Previous studies evaluated the interactions locally, thereby ignoring the competitive effects of different substructures or domains, but this could lead to high false-positive estimation, calling for a computational method that presents more predictive power.
Results: A statistical model, termed Global optimization-based InFerence of chemogenomic features from drug–Target interactions, or GIFT, is proposed herein to evaluate substructure-domain interactions globally such that all substructure-domain contributions to drug–target interaction are analyzed simultaneously. Combinations of different chemical substructures were included since they may function as one unit. When compared to previous methods, GIFT showed better interpretive performance, and performance for the recovery of drug–target interactions was good. Among 53 known drug–domain interactions, 81% were accurately predicted by GIFT. Eighteen of the top 100 predicted combined substructure-domain interactions had corresponding drug–target structures in the Protein Data Bank database, and 15 out of the 18 had been proved. GIFT was then implemented to predict substructure-domain interactions based on drug repositioning. For example, the anticancer activities of tazarotene, adapalene, acitretin and raloxifene were identified. In summary, GIFT is a global chemogenomic inference approach and offers fresh insight into drug–target interactions.
Availability and implementation: The source codes can be found at http://bioinfo.au.tsinghua.edu.cn/software/GIFT.
Contact: shaoli@mail.tsinghua.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein domains are independent folding structures and form different proteins as the functional units (Chothia, 1984). Compared with sequence similarity, domain structures are much more conserved among proteins (Koehn and Carter, 2005). Compound chemical substructures, or functional scaffolds, are the basic structural and functional building blocks of small molecules (Bemis *et al.*, 1996). According to Bredel and Jacoby (2004), chemogenomics, or chemical genomics, involves the systematic screening of targeted chemical libraries of small molecules against individual drug target families with the goal of identifying new drugs and drug targets. Thus, gathering chemogenomic data about protein domains and the chemical substructures of drugs underlying drug–target interactions could foster the development of fragment-based drug discovery, drug repositioning and the understanding of drug-induced side effects (Murray and Rees, 2009; Yamanishi *et al.*, 2011; Duran-Frigola *et al.*, 2013), thereby supplementing the network pharmacology methods of target prediction (Zhao and Li, 2010).

Among 1584 FDA-approved drugs and 89 nutraceuticals from the DrugBank database (Wishart *et al.*, 2008), only 30% of them are contained in the Protein Data Bank (PDB) database (Berman *et al.*, 2000; Gallina *et al.*, 2013). The large gap between the number of known binding positions and the total number of drug–target pairs calls for computational models to bridge them. Such computational models can be classified as structure-based methods and non-structure-based methods. Structure-based methods, such as docking, have been developed over the course of many years, but they are constrained by the dependence on three dimensional structures of proteins (Sousa *et al.*, 2013; Yang *et al.*, 2013). In recent years, several non-structure-based methods, which are not limited by the structure information have been developed, along the accumulation of large-scale database in both chemistry and biology.

Yamanishi *et al.* (2011) developed a sparse canonical correspondence analysis method to extract the associated sets of chemical substructures and protein domains from drug–target interactions, but the amount of detectable protein domains was relatively small. Takigawa *et al.* (2011) applied graph mining and sequence mining to automate the search for significant substructure patterns, leading to the interpretation of polypharmacology in drug–target network. The substructures in drugs or proteins did not need to be defined from the literatures. However, protein substructures, which are generally composed of two or three residues, were difficult to correlate with biological interpretations. Tabei *et al.* (2012) treated this question as a classification problem in machine learning. They used the logistic regression and support vector machine (SVM) algorithm to seek the drug chemical substructures and protein domains pairs that could be used to determine whether drugs and proteins interact or not. This method showed good results in predicting drug–target interactions.

However, none the methods noted above was able to evaluate uncertainty or variance of results, and they also ignored the possible combinations of drug chemical substructures that bind to protein domains as a whole. More importantly, the competitive effects of different substructure–domain interactions have never been considered since prediction has, thus far, been performed locally.

Therefore, in this article, a Global optimization-based InFerence of chemogenomic features from drug–target interactions, termed GIFT, was proposed (Fig. 1). Since different drug chemical substructures and protein domains may competitively contribute to drug–-target interactions, GIFT takes into consideration all the substructure–domain interactions and treats them as latent variables leading to drug–target interactions. A substructure–domain interaction is defined as a direct physical interaction between a chemical substructure and a protein domain, for example, a hydrogen bond between an atom in a chemical substructure and an atom in an amino acids (Gallina *et al.*, 2013). It should be noted that because some drug–target interactions have not been recorded in DrugBank, we incorporated two parameters, namely *fp* (false-positive rate) and *fn* (false-negative rate), to describe data inaccuracy and incompleteness. Then the substructure–domain contributions to different drug–target pairs were globally analyzed by the expectation–maximum (EM) algorithm framework, a statistical method to find the maximum likelihood estimates when the latent variables exist (Dempster *et al.*, 1977). Variances in results were calculated by the observed Fisher Information (Efron *et al.*, 1978). GIFT was further extended to include the combinations of different drug chemical substructures based on the likelihood that some may function together. More than 700 substructure-domain interactions were found by GIFT. Some were validated by the PDB database. Our approach is a promising method for investigating drug–target binding positions and contributes to our understanding in this field.

2 Results

2.1 Robust analysis of *fn* and *fp* on recovering drug–target interactions

As noted above, two parameters, *fp* and *fp* should be given in GIFT. Based on the assumptions in GIFT (see Section 4), we roughly estimated that the value of *fn* was no less than 0.4 and the value of *fp* was no more than 0.001 (see Section 4).

To analyze the robustness of GIFT with different parameters, we used 5-fold cross validation to detect the recoveries of drug–target interactions on different combinations of *fn* and *fp*. Here the negative samples were randomly selected from the drug–protein pairs without interacting records. The results showed that the performance for recovering drug–target interactions remained stable for different combinations of *fn* and *fp* (See Table 1). In GIFT, *fn* was set as 0.85 and *fp* was set as 0.0001.

Following the same cross validation procedure above, we compared GIFT with previous methods, namely, L1-log, L1-SVM (Tabei *et al.*, 2012), SCCA (Yamanishi *et al.*, 2011) and the association method (see Section 4), based on their performance for recovery of drug–target interactions. The association method was a naive approach to inference of substructure–domain interactions. The area under the ROC curve (AUC) of the association method is 0.72 (data not shown). Based on the AUC (see Table 2), GIFT performed better than the association method, L1-Log method, as well as SCCA, and it was comparable to L1-SVM for predicting drug–protein interactions.

Although a large number of substructure–domain interactions were extracted by the previous methods, for example, more than 350 000 interactions by SCCA (Yamanishi *et al.*, 2011), and around 5000 interactions by L1-SVM (Tabei *et al.*, 2012), it was difficult to assess the accuracy of these interactions. In the following sections, we demonstrate that GIFT not only perform better for drug–target predictions, but also provided better interpretation of the extracted substructure–domain interactions.

2.2 Performance of prediction on drug–domain interactions

GIFT was then evaluated for its performance in predicting drug–domain interactions. A total of 108 interactions between 53 proteins and 106 drugs were extracted from the crystallographic structures of known drug–protein interactions in the PDB database (Kruger *et al.*, 2012). Domains coexisting in the same proteins were merged, so that only multidomain proteins were considered after data preprocessing. Finally 53 pairs of drug–protein interactions including 53 pairs of drug–domain interactions were left. Then the *k* value was used to describe the number of binding (amino acid) residues of a given drug were located in a given protein domain. Here *k* value was defined as the proportion of the number of the binding residuals lying within the protein domain over the total number of binding residuals, given a drug–target pair. When k value was larger than 0.5, the drug was considered to interact with the domain (Kruger *et al.*, 2014).

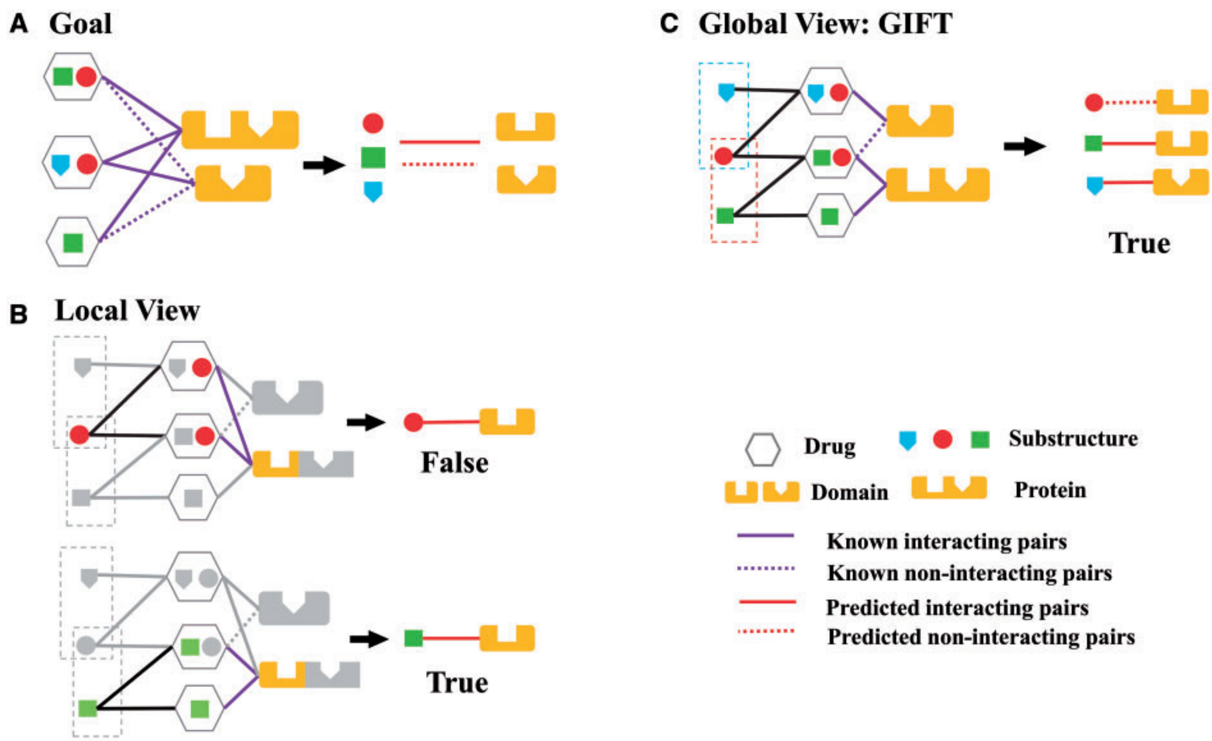

Fig. 1. Schematic diagram of GIFT. **A**. The goal of GIFT is to infer the underlying drug substructure-protein domain interactions, given a set of drug–protein interactions. **B**. The local view of previous approaches that test one pair of substructure-domain at a time while ignoring other substructures or protein domains. **C**. The global view of GIFT that considers contributions of all substructure-domain pairs to the given set of drug–target interactions

Table 1. Performance on recovery of drug–target interactions on different combinations of the two fixed parameters

fn	0.10	0.10	0.40	0.40	0.80	0.80
fp	0.0001	0.001	0.0001	0.001	0.0001	0.001
	0.845 (0.006)	0.837 (0.005)	0.846 (0.006)	0.838 (0.005)	0.847 (0.006)	0.840 (0.005)

The values are the mean areas under the ROC curves with the standard variances.

Table 2. Performance on recovery of drug–target interactions

Ratio	GIFT	L1-Log	L1-SVM	SCCA
1	0.835 (0.006)	0.829 (0.001)	0.830 (0.001)	0.798 (0.002)
5	0.847 (0.006)	0.838 (0.001)	0.855 (0.001)	0.798 (0.002)

The values are the mean areas under the ROC curves with the standard variances. Ratio is the proportion of negative samples over the total number of training samples.

The prediction of drug–domain interactions were followed Equation 2 in Section 4, and each domain was treated as one protein. A drug and a protein domain were predicted to interact if the score given by GIFT was larger than zero.

Eighty-one percent of the drug–domain interactions could be predicted by GIFT. The representative results were shown in Table 3, and the complete results can be found in Supplementary Material.

2.3 Validation of substructure–domain interactions

A total of 726 substructure–domain interactions, whose scores, given by GIFT, were larger than their standard deviations, were extracted, including 255 interactions between substructure combinations and domains. Since some chemical substructures may function as one unit, more than 1800 combinations were recorded using the SMARTS format and added in our model (See Supplementary Material).

Interactions between combinations of chemical substructures and domains were validated by the known crystal complex containing the relevant drugs and proteins from the PDB database. Eighteen of the top 100 predicted interactions between combinations of chemical substructures and domains have the corresponding drug-–target structures in the PDB database, and 15 of the 18 ones were successfully predicted (see Table 4).

Both methotrexate and pemetrexed are antitumor drugs (Wishart *et al*., 2008), and can target against the dihydrofolate reductase and thymidylate synthase. One combination of the substructures interact with dihydrofolate reductase domain were predicted by GIFT, in good agreement with the real data (Fig. 2A and B). DHF, the methotrexate analog, can also interact with thymidylate synthase and showed the same results (Fig. 2C). It was shown that methotrexate targeted two different domains with the same substructures. Trimetrexate is one agent against amoebiasis and other protozoal disease. Although it lacked one predicted substructure, when compared with methotrexate and pemetrexed, trimetrexate could still target against dihydrofolate reductase domain by the other predicted substructure (Fig. 2D). Since it shares the same functional substructure with methotrexate and pemetrexed, trimetrexate may have similar biological activity. Indeed, it has been tested as an antitumor drugs (Wishart *et al*., 2008).

2.4 Interpretation of the drug–protein interactions by GIFT

We further explored the drug–protein interactions by GIFT, using four different proteins from Homo sapiens: RARA, RXRA, RXRG and ESR1 as examples. All have the ligand-binding domain of

Table 3. Representative results of the predictions on drug–domain interactions

Protein	Drug	Domain	*k* value	Prediction
DNA (cytosine-5)-methyltransferase 1	*S*-Adenosylhomocysteine	C-5 cytosine-specific DNA methylase	1	TRUE
Alcohol dehydrogenase 1B	*N*-benzylformamide	Alcohol dehydrogenase GroES-like domain	0.58	TRUE
Androgen receptor	Flufenamic Acid	Ligand-binding domain of nuclear hormone receptor	0.9	TRUE
Ornithine carbamoyltransferase	*N*-(Phosphonoacetyl)-L-ornithine	Asp/Orn binding domain	0.51	TRUE
Progesterone receptor	Norethindrone	Ligand-binding domain of nuclear hormone receptor	1	TRUE
Rho-associated protein kinase 1	Hydroxyfasudil	Protein kinase domain	0.94	TRUE
Tissue-type plasminogen activator	Benzamidine	Trypsin	1	TRUE

k value is the proportion of the number of the binding residues (amino acids) lying within the protein domain over the total number of the binding residues. When k value is larger than 0.5, the drug is considered to interact with the domain. TRUE means the prediction that the drug interacts with the domain by GIFT fits the real data well.

Table 4. Results of the prediction interactions between drug chemical substructure combinations and protein domains

Domain	Chemical substructure	Chemical substructure	Prediction (variance)	Proof from PDB
Eukaryotic-type carbonic anhydrase	N–S	Sc1cc(S)ccc1	1(0.21)	1zgf
Dihydrofolate reductase	C(∼N)(:C)(:N)	O–C–C–C–C–C–O	0.32 (0.03)	1u70
Thymidylate synthase	O=C–C=C–N	O–C–C–C–C–N	0.18 (0.01)	1lcb
Dehydratase family	O(∼C)(∼P)	O=C–C=C–N	0.18 (0.005)	1n7h
Dihydrofolate reductase	C(∼Cl)(:C)	N–C–N–C:C	0.17 (0.01)	1j3j
Phosphoenolpyruvate carboxykinase	C(∼N)(:C)(:N)	O=C–C–C–N	0.17 (0.01)	1khb
Cation transporting ATPase	N–S	Sc1cc(S)ccc1	0.16 (0.02)	1zgf
5’ nucleotidase family	C(∼C)(:N)(:N)	O–C–C–C–C–O	0.15 (0.01)	2xcw
Phosphorylase superfamily	N:C:N:C	O–C–C–O–[#1]	0.13 (0.005)	1je1
Carbon-nitrogen hydrolase	C(∼C)(∼C)(∼H)(∼N)	O=C–C–C–C–C=O	0.11 (0.006)	3syt
Angiotensin-converting enzyme	O=C–C–N–C	O–C–C–C–C–C–N–C	0.11 (0.002)	1j37
Glutaminase	C(∼C)(∼C)(∼H)(∼N)	O=C–C–C–C–C=O	0.11 (0.003)	3vp0
Thymidylate synthase	C(–H)(–N)(=C)	O=C–N–C–[# 1]	0.08 (0.005)	4fqs
GMP reductase domain	OC1CC(O)CCC1	CC1C(C)CCCC1	0.08 (0.001)	4fo4
Carboxylesterase	O–C:C–O–C	Cc1c(C)cccc1	0.08 (0.005)	1eve

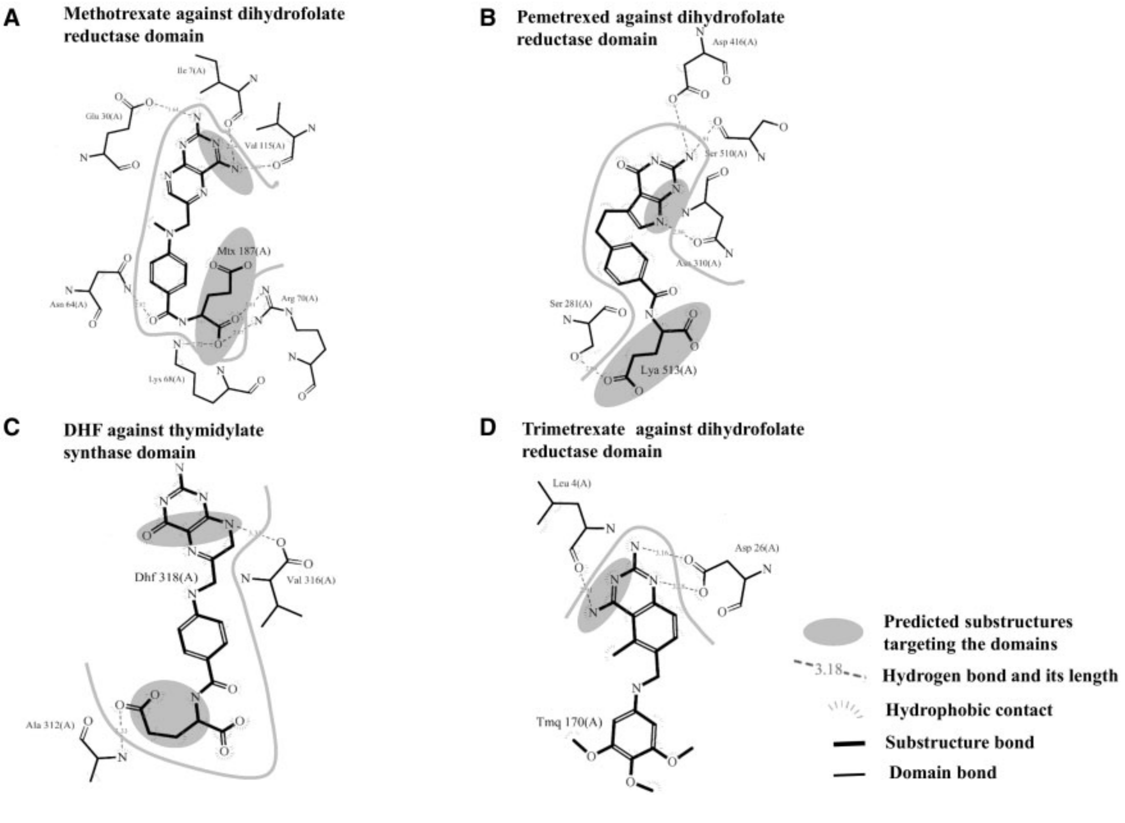
A
Methotrexate against dihydrofolate reductase domain
Ile 7(A)
Glu 30(A)
Val 115(A)
Mtx 187(A)
Arg 70(A)
Asn 64(A)
Lys 68(A)
B
Pemetrexed against dihydrofolate reductase domain
Asp 416(A)
Ser 510(A)
Asn 310(A)
Ser 281(A)
Lya 513(A)
C
DHF against thymidylate synthase domain
Dhf 318(A)
Val 316(A)
Ala 312(A)
D
Trimetrexate against dihydrofolate reductase domain
Leu 4(A)
Asp 26(A)
Tmq 170(A)
Predicted substructures targeting the domains
3.18
Hydrogen bond and its length
Hydrophobic contact
Substructure bond
Domain bond

Fig. 2. Examples of the substructure-domain interactions validated from the PDB database: (**A**) PDB entry 1u70, (**B**) PDB entry 3k2h, (**C**) PDB entry 1lcb, (**D**) PDB entry 1bzf. Bold black: the drugs. Thin black: the amino acids. All the figures are generated by LigPlot (Wallace *et al.*, 1995)

nuclear hormone receptor. ESR1 also contains the oestrogen receptor domain. Using GIFT, several chemical substructures were predicted to interact with the above domains. Drugs having these chemical substructures that can target at least one of the four proteins were shown in Figure 3. In the original drug-target interactions, tretinon, adapalene and acitretin targeted against RARA, RXRA and RXRG. Fulvestrant, raloxifene and letrozole interacted with ESR1. Tazarotene only targeted RARA. RARA shares the same protein domain as tretinon, adapalene and acitretin target RXRA and RXRG. Based on the results of GIFT, the four drugs have the same chemical substructures predicted to target the ligand-binding domain of nuclear hormone receptor (labeled in black in Fig. 3) and they may, therefore, share the same therapeutic functions. Since tretinoin has anticancer bioactivity, it suggests that acitretin, adapalene, tazarotene and raloxifene may have similar anticancer activity similar to that of tretinoin. Meanwhile, fluvestrant has the same predicted chemical substructures with tretinoin, adapalene, and tazarotene. It suggests that fluvestrant may interact with ESR1 by targeting the ligand-binding domain instead of oestrogen receptor domain.

Recent researches partly confirm the GIFT prediction above. It is found that adapalene could block cell proliferation and induce apoptosis on hepatoma cells (Ocker *et al.*, 2004). It was reported that tazarotene profoundly inhibits murine Basal cell carcinoma by inducing PI3K-AKT signaling (So *et al.*, 2014). Acitretin and several analogs exhibited antiproliferative activities against human breast MCF-7 epithelial cells (Magoulas *et al.*, 2011). Raloxifene, one of the selective estrogen-receptor modulators, was effective as a preventive choice in reducing the risk of breast cancer for postmenopausal woman (Vogel *et al.*, 2010).

3 Discussion

In this article, a statistical model, termed GIFT, was constructed to infer the interactions between drug chemical substructures and protein domains based on drug–protein interactions data. For the first time, a global optimization perspective has been introduced into this computation, meaning that all the substructure-domain contributions on drug–target interactions are simultaneously analyzed.

GIFT has several advantages compared with previous methods. First, it offers a more detailed description on the drug–protein interactions. Instead of adhering to the constraint which holds that ‘similar drugs target similar proteins’, we attempt to reveal the possible mechanisms underlying drug–protein interactions. Second, more than 350 000 interactions were predicted by SCCA (Yamanishi *et al.*, 2011), and around 5000 interactions were predicted by L1-SVM (Tabei *et al.*, 2012). However, these figures represent local estimates, in addition to the lack statistical significance evaluations, thus potentially overestimating substructure–domain interactions. GIFT, on the other hand, can reduce such overestimation by incorporating a global perspective that accommodates variance estimation. It should be noted that GIFT is limited in that it only focuses on two-dimensional chemical space, as presented by the 881 PubChem substructures. Therefore, in the future, three-dimensional structure information should be incorporated into the calculations (See Section 4). As reflected by data already accumulated in the PDB database, more than 20,000 compound-protein interactions can be found, including more than 9000 distinct compounds (Gallina *et al.*, 2013; de Beer *et al.*, 2014). Extraction of these binding information might result in more precise estimations of the substructure-domain interactions in the future.

In summary, we propose a novel chemogenomic approach to infer substructure-domain interactions. It provides a chemogenomic view of the mechanisms of drug–protein interactions, and can be used as a new method that contributes to target prediction, drug repositioning and drug combination studies.

4 Methods

4.1 Data sources of GIFT

The information of drug–protein interactions, drug substructures, and protein domains is obtained from Tabei *et al.*, 2012. A total of 1862 drugs are represented by 881-dimensional chemical substructure binary vectors from PubChem database, and 1554 proteins are represented by 876-dimensional protein domain binary vectors from the Pfam database (Bateman *et al.*, 2004). Here the substructure is defined as a specific kind of two-dimensional chemical fingerprint by CACTVS. 4809 interactions exist between the drugs and the proteins. We deleted the drug chemical substructures or protein domains that never appeared in the drugs or proteins, and we merged those substructures or domains that appeared in the same drugs or proteins.

The drug–domain interactions were extracted from PDB database by the script from Kruger *et al.*, 2012. We only chose proteins that had multiple domains for our data. Finally, 53 pairs of drug–protein interactions with the records of drug–domain interaction were used.

4.2 The EM framework of GIFT

Let $Y_1 \ldots Y_T$ denote the T drugs, and $P_1 \ldots P_S$ denote the S proteins. Let $Z_1 \ldots Z_M$ denote the M drug chemical substructures and let $D_1 \ldots D_N$ denote the N protein domains. Let $ZD^{(ij)}$ denote the set of the pairs of chemical substructures and domains from drug Y_i and protein P_j correspondingly. Let ZD_{mn} denote the interaction result between the chemical substructure D_m and the domain D_n.$ZD_{mn}^{(ij)} = 1$ if they interact and $ZD_{mn}^{(ij)} = 0$ otherwise. Let YP_{ij} denotes the interaction result between the drug Y_i and the protein P_j. $YP_{ij} = 1$ if they interact and $YP_{ij} = 0$ otherwise.

Inspired by the work (Deng *et al.*, 2002), we herein propose a statistical method to evaluate the possible interactions between drug chemical substructures and protein domains. For our calculations, it was assumed that (i) the interactions of the drug chemical substructures and the protein domains are independent, given a pair of a drug and a protein pair; (ii) interactions between a given drug chemical substructure and protein domain would remain unchanged between different pairs of the drugs and proteins containing them, as shown by

$$\theta_{mn} = Pr(D_{mn}^{(ij)} = 1) \qquad (1)$$

in which $\theta_{mn} = Pr(D_{mn} = 1)$; (iii) drug and the protein will interact if, and only if, one pair of chemical substructures and domains from them interact. Based on these assumptions, we can get

$$Pr(YP_{ij} = 1|\theta) = 1 - \prod_{D_{mn}^{(ij)}} (1 - \theta_{mn}) \qquad (2)$$

We include two types of errors in the data of the drug protein interactions: *fp* (false positive rate), in which the drug and the protein do not interact, but are recorded to be interacting, and *fn*, in which the drug and the protein interact, but are not recorded. Let O_{ij} be the result of observed interaction between drug Y_i and protein P_j:$O_{ij} = 1$ if the interaction is observed and $O_{ij} = 0$ otherwise.

Then

$$fp = Pr(O_{ij} = 1|YP_{ij} = 0) \qquad (3)$$

$$fn = Pr(O_{ij} = 0|YP_{ij} = 1) \qquad (4)$$

Both *fn* and *fp* are fixed in GIFT.

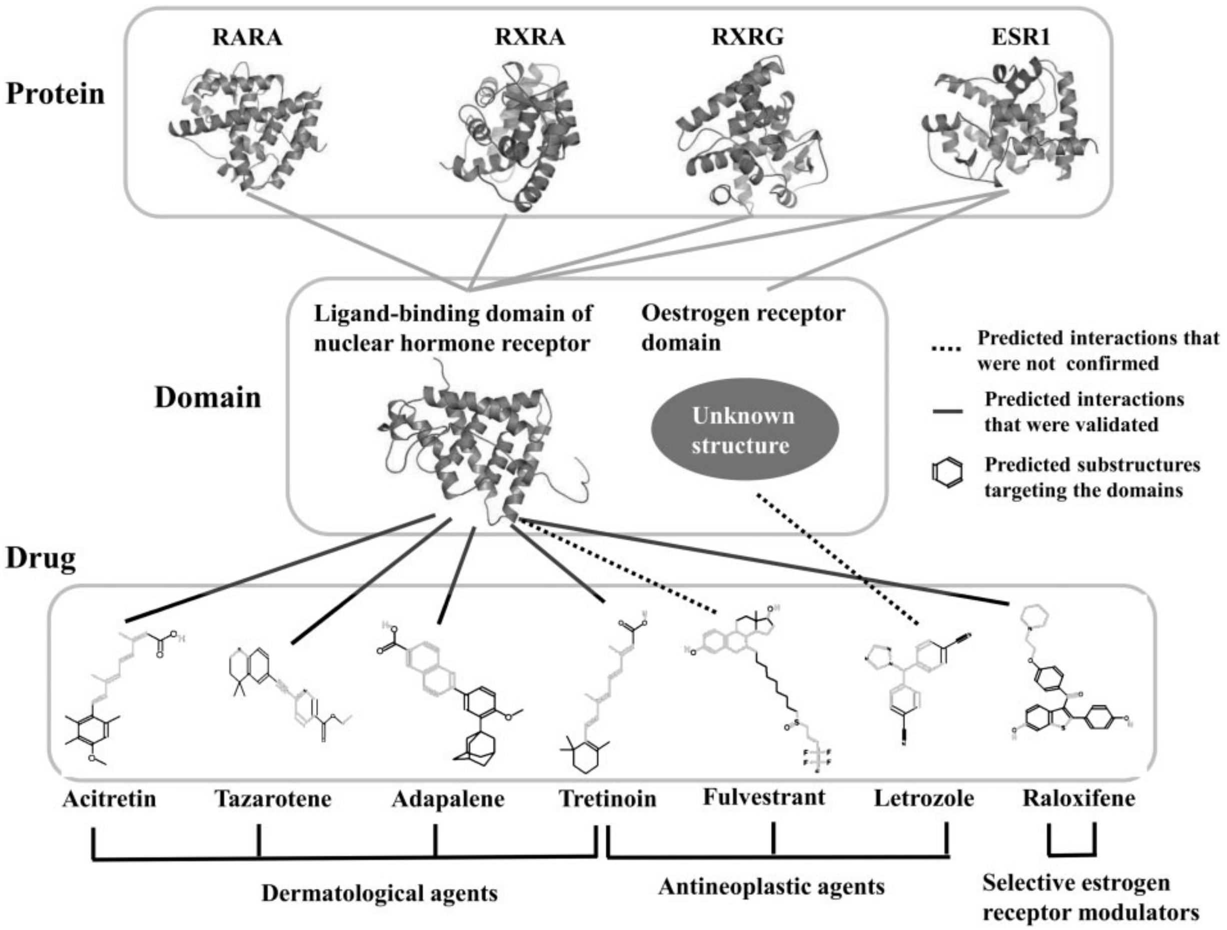
Protein
RARA
RXRA
RXRG
ESR1
Domain
Ligand-binding domain of nuclear hormone receptor
Oestrogen receptor domain
Unknown structure
.... Predicted interactions that were not confirmed
— Predicted interactions that were validated
Predicted substructures targeting the domains
Drug
Acitretin
Tazarotene
Adapalene
Tretinoin
Fulvestrant
Letrozole
Raloxifene
Dermatological agents
Antineoplastic agents
Selective estrogen receptor modulators

Fig. 3. The example of connecting drug–protein interactions with the chemical substructure-domain interactions by GIFT

$$fn = Pr(O_{ij}=0|YP_{ij}=1) = 1 - \frac{Pr(O_{ij}=1, YP_{ij}=1)}{Pr(YP_{ij}=1)}$$
$$\geq 1 - \frac{Pr(O_{ij}=1)}{Pr(YP_{ij}=1)} \geq 1 - \frac{\text{number of observed interaction pairs}}{\text{number of real interaction pairs}}$$

It was estimated that on average the number of target proteins per drug was about 6.3 (Mestres *et al.*, 2008). The value of *fn* then would be no less than 0.41. We estimate *fp* in a similar way and it would be no more than 0.001.

And the probability for the observed interaction between drug Y_i and the protein P_j is

$$Pr(O_{ij}=1|\theta) = (1-fn)Pr(YP_{ij}=1|\theta) + fp \cdot Pr(YP_{ij}=0|\theta) \quad (5)$$

The log likelihood function is followed

$$l(\theta) = log(Pr(O|\theta)) \quad (6)$$

It is a function of $\theta = \{\theta_{mn}\}$. θ is estimated by the maximum likelihood estimation approach. The EM algorithm (Dempster *et al.*, 1977) is then used here.

The observed data are the interactions from DrugBank Database, $O = \{O_{ij} = o_{ij}\}$. The complete data include all the drug chemical substructures and protein domain interactions and the observed data. Let A_m be the set of drugs containing the chemical substructure Z_m and let A_n be the set of proteins containing the domain D_n. Let N_{mn} be the total number of pairs between A_m and A_n. The EM algorithm as follows:

E Step:

$$E(D^{(ij)}_{mn}|O,\theta^{(t-1)}) = \frac{\theta^{(t-1)}_{mn}(1-fn)^{O_{ij}}fn^{1-O_{ij}}}{Pr(O_{ij}|\theta^{(t-1)})} \quad (7)$$

M Step:

$$\theta^{(t)}_{mn} = \frac{1}{N_{mn}} \sum_{i,j:Zm\in Y_i, Dn\in P_j} E(D^{(ij)}_{mn}|O_{ij},\theta^{(t-1)}) \quad (8)$$

The variance of the parameters can be estimated by the observed Fisher information, which is followed:

$$var(\hat{\theta}) = \frac{1}{I(\hat{\theta})}, I(\theta) = -\frac{d^2 log(Pr(O|\theta))}{d\theta^2} \quad (9)$$

Here $I(\hat{\theta})$ is the observed Fisher information. In GIFT, the observed Fisher information is followed:

$$I(\theta_{mn}) = \sum_{i,j:Zm\in Y_i, Dn\in P_j} \delta^{(i,j)^2}_{mn}\left(\frac{O^{(ij)}_{mn}}{\mu^{(ij)^2}_{mn}} + \frac{1-O^{(ij)}_{mn}}{(1-\mu^{(ij)}_{mn})^2}\right) \quad (10)$$

In which,

$$\delta^{(ij)}_{mn} = \frac{\mu^{(ij)}_{mn}}{\partial\theta_{mn}}, \mu^{(ij)}_{mn} = Pr(O^{(ij)}_{mn}=1|\theta) \quad (11)$$

Since some drug chemical substructures may work as one unit, 1870 pairs of drug chemical substructures that frequently co-occur among the drugs are selected as the extra drug chemical substructures. All the details of GIFT can be found in Supplementary Material.

4.3 The association method

The association method is a naïve way to estimate the interactions between chemical substructures and protein domain, which is the fraction of interacting drug–protein pairs among all of the drug–protein pairs containing the pair of chemical substructure Z_m and protein domain D_n.

$$\theta_{mn} = \frac{I_{mn}}{N_{mn}} \tag{12}$$

in which I_{mn} is the number of interacting pairs of drug–protein pairs containing the pair of chemical substructure Z_m and protein domain D_n and N_{mn} is the number of total drug–protein pairs containing the pair of chemical substructure Z_m and protein domain D_n.

Acknowledgement

We thank Dr. Felix A Kruger at University of Cambridge for helpful discussion.

Funding

This work is supported by National Natural Science Foundation of China (Nos. 91229201 and 81225025).

Conflict of Interest: none declared.

References

Bateman,A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids. Res*., **32**, D138–D141.

Bemis,G.W. *et al*. (1996) The properties of know drugs. 1. Molecular frameworks. *J. Med. Chem.,* **39**, 2887–2893.

Bredel,M. and Jacoby,E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet*., **5**, 262–275.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Chothia,C. (1984) Principles that determine the structure of proteins. *Annu. Rev. Biohem*., **53**, 537–572.

Deng,M. *et al*. (2002) Inferring domain–domain interactions from protein–-protein interactions. *Genome Res*., **12**, 1540–1548.

de Beer,T.A. *et al*. (2014) PDBsum additions. *Nucleic Acids. Res*., **42**, D292–D296.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series. B*., **39**, 1–38.

Duran-Frigola,M. and Aloy,P. (2013) Analysis of chemical and biological features yields mechanistic insights into drug side effects. *Chem. Biol.* **20**, 594–603.

Efron,B. *et al*. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457–483.

Gallina,A.M. *et al*. (2013) PLI: a web-based tool for the comparison of protein–ligand interactions observed on PDB structures. *Bioinformatics*, **29**, 395–397.

Gaulton,A. *et al*. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*., **40**, D1100–D1107.

Geman,S. *et al*. (1984) Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.,* **23**, 97–115.

Koehn,F.E. and Carter G.T. (2005) The evolving role of natural products in drug discovery, *Nat. Rev. Drug Discov.*, **4**, 206–220.

Kruger,F.A. *et al*. (2012) Mapping small molecule binding data to structural domains. *BMC Bioinformatics*, **13**, S11.

Kruger,F.A. *et al*. (2014) PPDMs—a resource for mapping small molecule bioactivities from ChEMBL to Pfam—A protein domains. *Bioinformatics*, **31**, 776–778.

Magoulas,G.E. *et al*. (2011) Syntheses, antiproliferative activity and theoretical characterization of acitretin-type retinoids with changes in the lipophilic part. *Eur. J. Med. Chem*., **46**, 721–737.

Mestres,J. *et al*. (2008) Data completeness—the Achilles heel of drug–target networks, *Nat. Biotechnol*., **26**, 983–984.

Murray,C.W. and Rees, D.C. (2009) The rise of fragment-based drug discovery, *Nat. Chem*., **1**, 187–192.

Ocker,M. *et al*. (2004) Potentiated anticancer effects on hepatoma cells by the retinoid adapalene. *Cancer Lett*., **208**, 51–58.

So,P.L. *et al*. (2014) PI3K-AKT signaling is a downstream effector of retinoid prevention of murine basal cell carcinogenesis. *Cancer Prev. Res*., **7**, 407–417.

Sousa,S.F. *et al*. (2013) Protein-ligand docking in the new millennium—a restrospective of 10 years in the field. *Curr. Med. Chem*., **20**, 2296–2314.

Tabei,Y. *et al*. (2012) Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*, **28**, i487–i494.

Takigawa,I. *et al*. (2011) Mining significant substructure pairs for interpreting polypharmacology in drug–target network. *PloS One*, **6**, e16999.

Vogel,V.G. *et al*. (2010) Update of the national surgical adjuvant breast and bowel project study of tamoxifen and raloxifene (STAR) P-2 trial: preventing breast cancer. *Cancer Prev. Res*., **3**, 696–706.

Wang,Y. *et al*. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids. Res*., **37**, W623–W633.

Wallace,AC. *et al*. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*., **8**, 127–134.

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

Yabuuchi,H. *et al*. (2011) Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol*., **7**, 472.

Yamanishi,Y. *et al*. (2011) Extracting sets of chemical substructures and protein domains governing drug–target interactions. *J. Chem. Inf. Model*., **51**, 1183–1184.

Yang, J. *et al*. (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.

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