

Inference of Chemogenomic Features from Drug-Protein Interactions by EM Algorithm

An Exploration of Global Optimization

Songpeng Zu, Ting Chen, Shao Li *

Department of Automation, Tsinghua University

zusongpeng@gmail.com



Abstract

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.

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. The source codes and results can be found at <http://bioinfo.au.tsinghua.edu.cn/software/GIFT>.

Introduction

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, thereby supplementing the network pharmacology methods of target prediction.

However, previous methods were not 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.

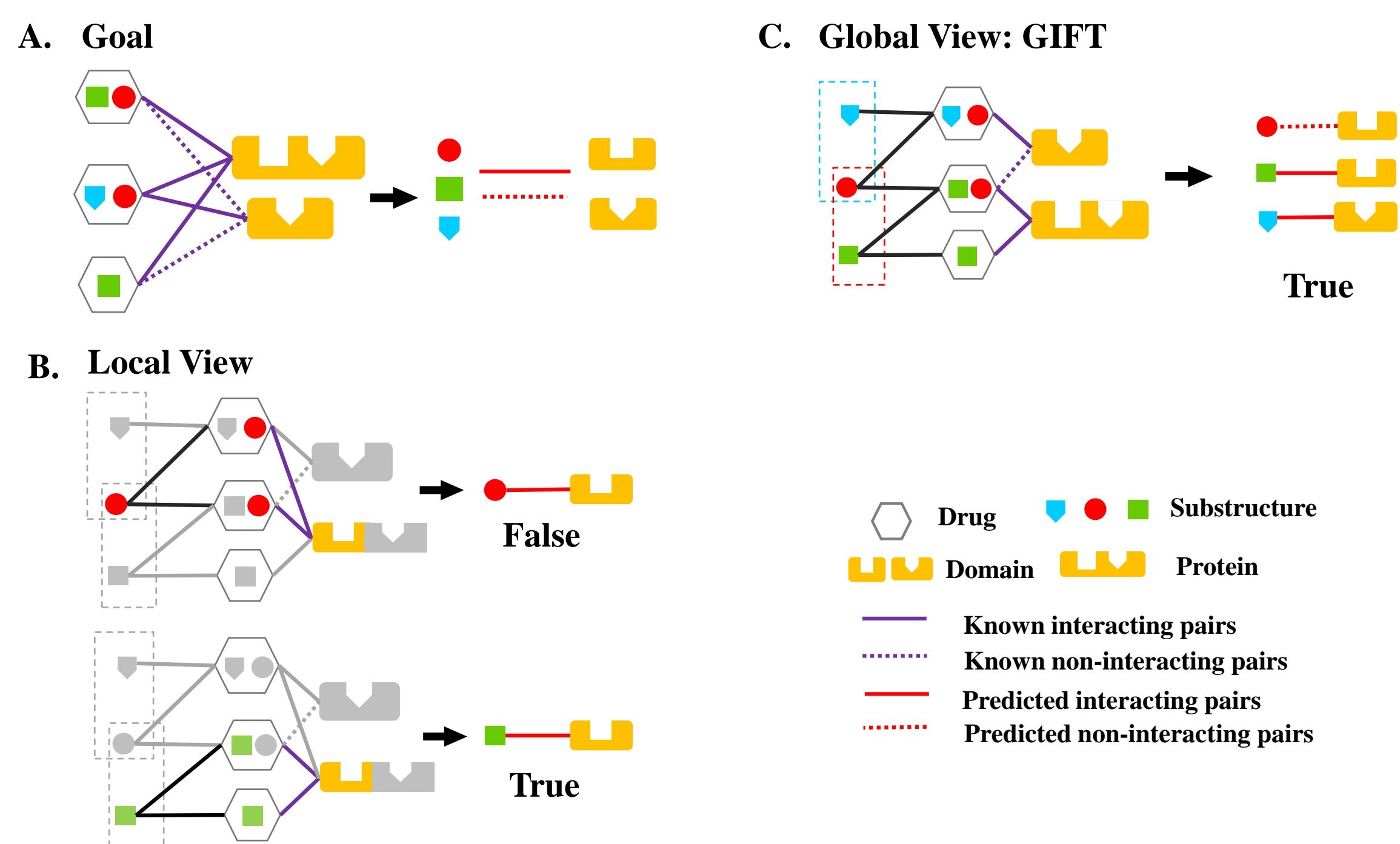


Figure 1: Figure caption

Materials and Methods

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. 4809 interactions exist between the drugs and the proteins.

The EM framework of GIFT

Let $Y_1 \dots Y_T$ denote the T drugs, and $P_1 \dots P_S$ denote the S proteins. Let $Z_1 \dots Z_M$ denote the M drug chemical substructures and let $D_1 \dots D_N$ 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} denote the interaction result between the drug Y_i and the protein P_j . $YP_{ij} = 1$ if they interact and $YP_{ij} = 0$ otherwise.

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

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

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

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

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 (false

negative rate), 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

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

Both fn and fp are fixed in GIFT.

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

$$Pr(O_{ij} = 1|\theta) = (1 - fn)Pr(YP_{ij} = 1|\theta) + fp \cdot Pr(YP_{ij} = 0|\theta) \quad (4)$$

The log likelihood function is followed

$$l(\theta) = \log(Pr(O|\theta)) \quad (5)$$

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:

$$E(D_{mn}^{(ij)}|O, \theta^{(t-1)}) = \frac{\theta_{mn}^{(t-1)}(1 - fn)^{O_{ij}}fn^{1-O_{ij}}}{Pr(O_{ij}|\theta^{(t-1)})} \quad (6)$$

M Step:

$$\theta_{mn}^{(t)} = \frac{1}{N_{mn}} \sum_{i,j: Z_m \in Y_i, D_n \in P_j} E(D_{mn}^{(ij)}|O_{ij}, \theta^{(t-1)}) \quad (7)$$

Results

Phasellus imperdiet, tortor vitae congue bibendum, felis enim sagittis lorem, et volutpat ante orci sagittis mi. Morbi rutrum laoreet semper. Morbi accumsan enim nec tortor consectetur non commodo nisi sollicitudin. Proin sollicitudin. Pellentesque eget orci eros. Fusce ultricies, tellus et pellentesque fringilla, ante massa luctus libero, quis tristique purus urna nec nibh.

Nulla ut porttitor enim. Suspendisse venenatis dui eget eros gravida tempor. Mauris feugiat elit et augue placerat ultrices. Morbi accumsan enim nec tortor consectetur non commodo. Pellentesque condimentum dui. Etiam sagittis purus non tellus tempor volutpat. Donec et dui non massa tristique adipiscing. Quisque vestibulum eros eu. Phasellus imperdiet, tortor vitae congue bibendum, felis enim sagittis lorem, et volutpat ante orci sagittis mi. Morbi rutrum laoreet semper. Morbi accumsan enim nec tortor consectetur non commodo nisi sollicitudin.

In hac habitasse platea dictumst. Etiam placerat, risus ac.

Adipiscing lectus in magna blandit:

	Ratio	GIFT	L1-Log	L1-SVM	SCCA
1	0.835	0.829	0.830	0.798	
5	0.847	0.838	0.855	0.798	

Table 1: Performance of recovery of drug-target interactions

Vivamus sed nibh ac metus tristique tristique a vitae ante. Sed lobortis mi ut arcu fringilla et adipiscing ligula rutrum. Aenean turpis velit, placerat eget tincidunt nec, ornare in nisl. In placerat.

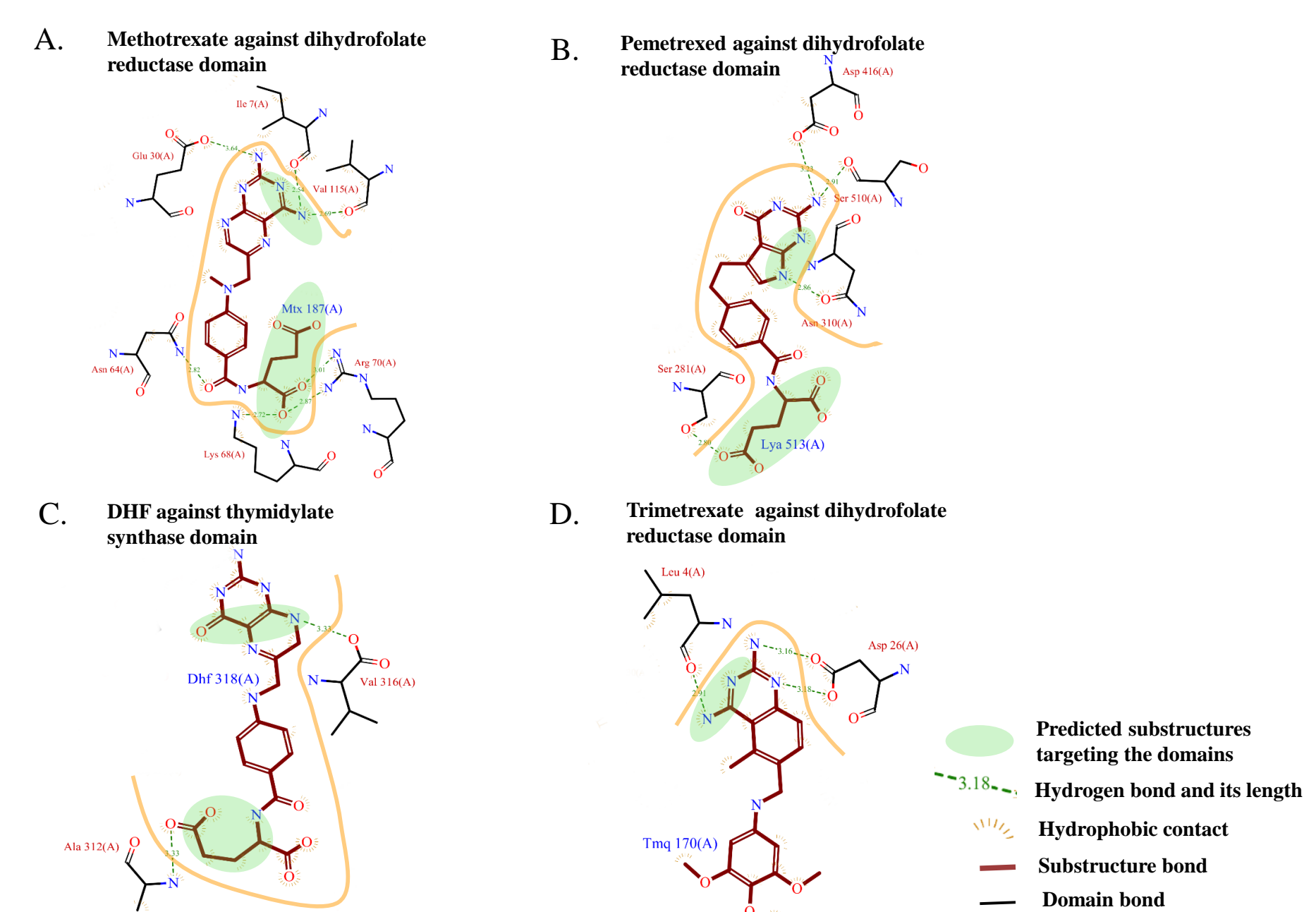


Figure 2: Figure caption

References

- [1] Zu, S., Chen, T., Li, S. (2015) Global Optimization-based Inference of Chemogenomic Features from Drug-Target Interactions. *Bioinformatics, Online*
- [2] Tabei, Y. et al. (2012) Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers. *Bioinformatics*, **28**, i487-i494.