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.

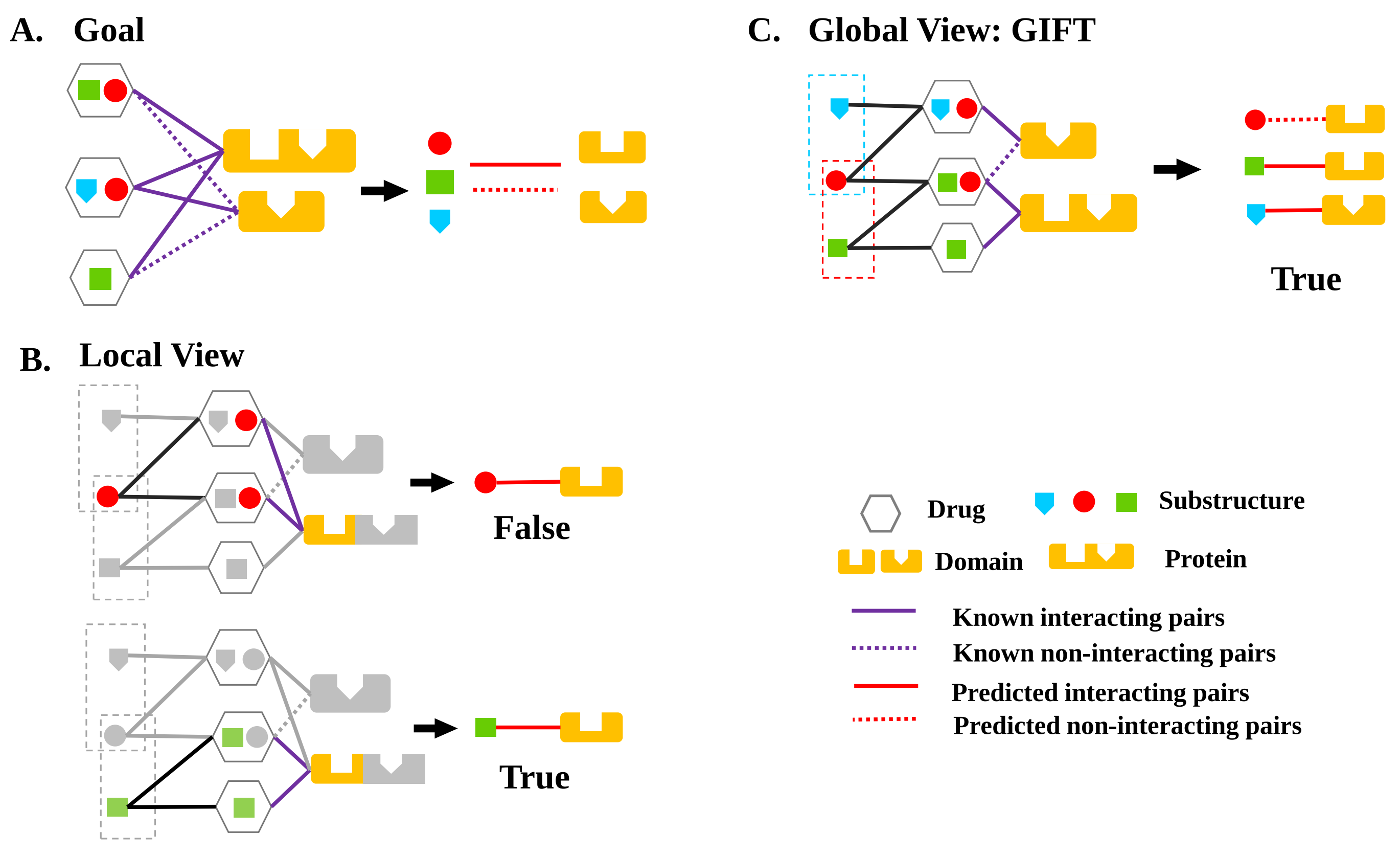

Figure 1: Schematic diagram of GIFT. A. The goal of GIFT is to infer the underlying drug substructure-protein domain interactions, given a set of drug-protein interactions. B. The local view of previous approaches that test one pair of substructure-domain at a time while ignoring other substructures 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.

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 $Y_1 \ldots Y_T$ 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.

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 thedrugs and proteins containing them, as shown by

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

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

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

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

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

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

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

The log likelihood function is followed

$$l(\theta) = log(Pr(O|\theta)) \tag{5}$$

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:

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

M Step:

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

Results

Following 5-fold cross validation procedure, we compared GIFT with previous methods, namely, L1-log, L1-SVM, SCCA and the association method, 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 1), 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.

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

GIFT was then evaluated for its performance in predicting drug-domain interactions. The prediction of drug-domain interactions were followed Equation 2 in Methods, 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. 81% of the 53 known drug-domain interactions could be predicted by GIFT. The representative results were shown in Figure 2.

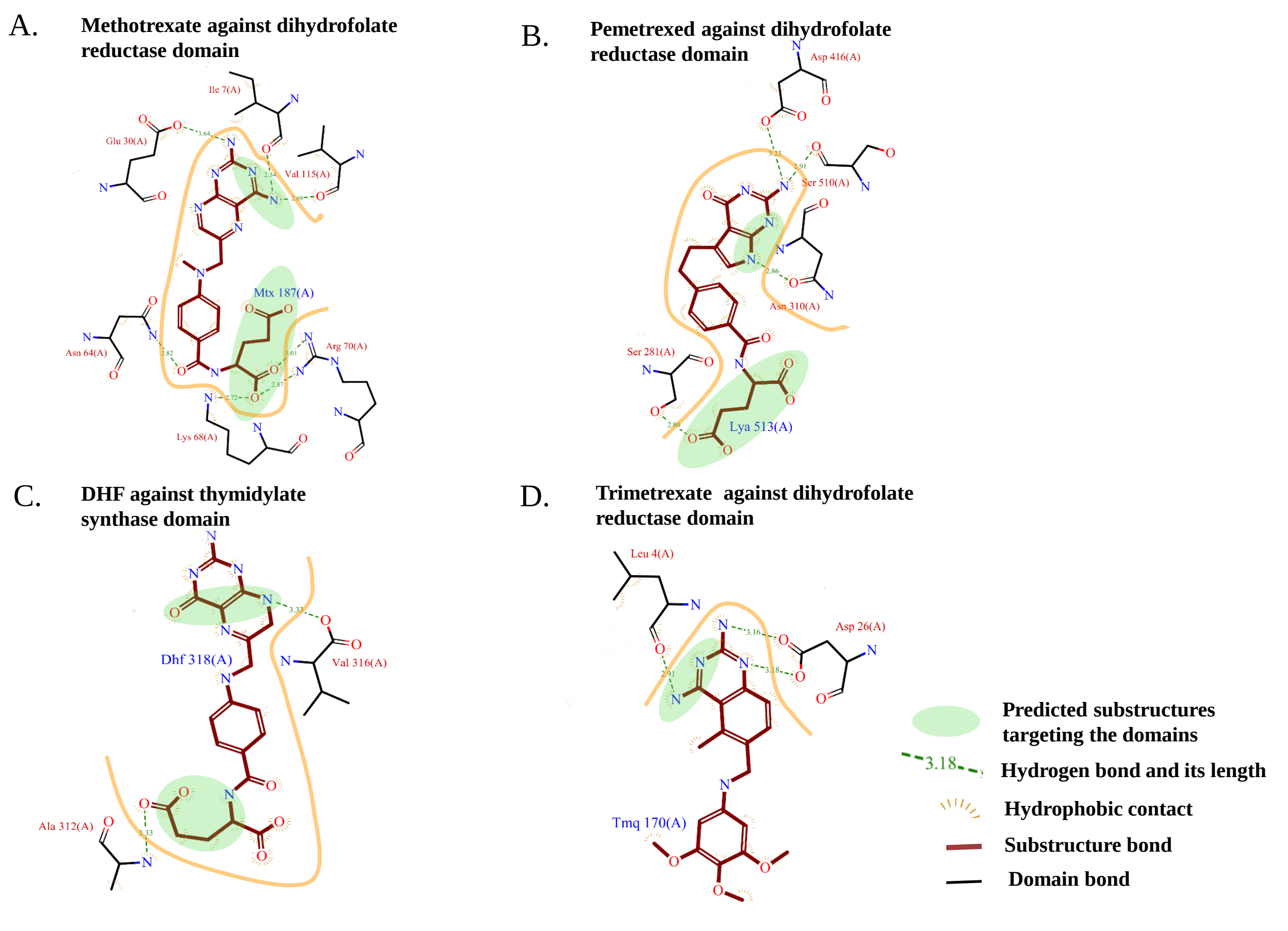
A.
Methotrexate against dihydrofolate reductase domain
Ile 7(A)
Glu 30(A)
Val 115(A)
Mtx 187(A)
Asn 64(A)
Arg 70(A)
Lys 68(A)
B.
Pemetrexed against dihydrofolate reductase domain
Asp 416(A)
Ser 510(A)
Asn 310(A)
Ser 281(A)
Lya 513(A)
C.
DHF against thymidylate synthase domain
Dhf 318(A)
Val 316(A)
Ala 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

Figure 2: Examples of the substructure-domain interactions validated from the Protein Data Bank database: (A) PDB entry 1u70, (B) PDB entry 3k2h, (C) PDB entry 1lcb, (D) PDB entry 1bzf. Dark grey: the drugs. Black: the amino acids. All the figures are generated by LigPlot.

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.