Discovery

The correlation between proteins sequence similarities and ligand-based similarities

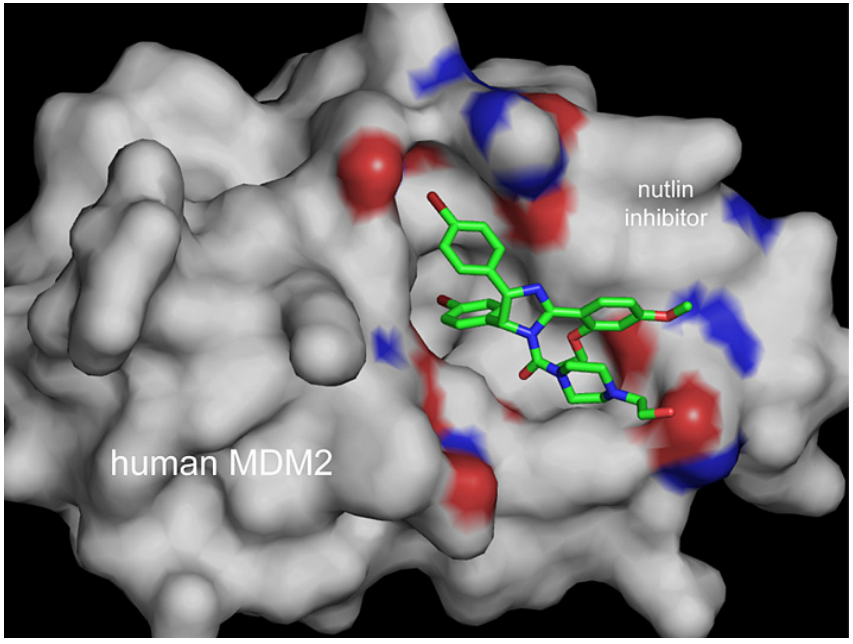
Songpeng Zu

FIT 1-108, Tsinghua University

September 9, 2013

The Problem

- Drugs can target on several proteins (Mestres *et al*,2008).
- It's related to both drugs side effects and drugs efficiency.



Outline

- SUDO: Inferring Drug Substructures and Protein Domains Interactions from Drug Target Interactions
- LIGNET: Incorporation of Ligand-based and Network-based Approaches for Prediction of Drug Target Interactions
- Summary and Future Plan

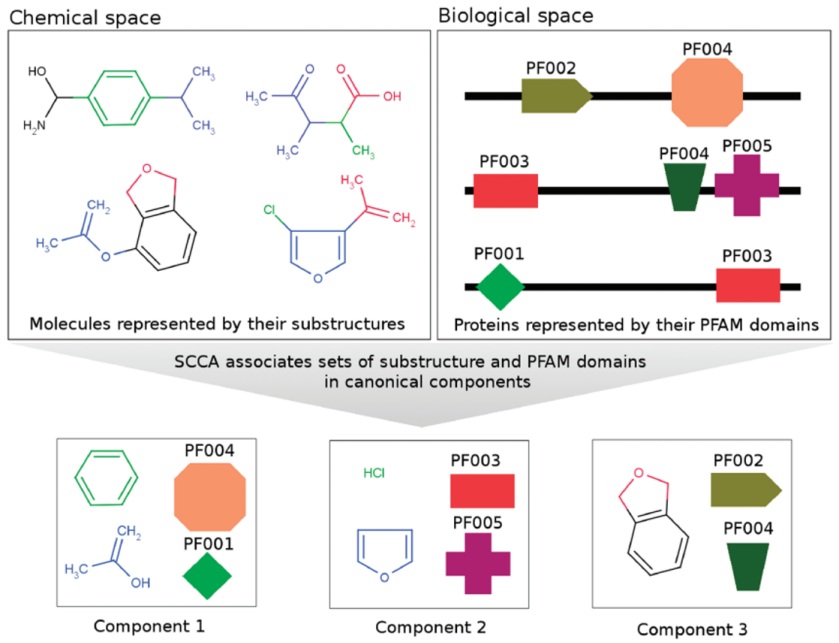
Part I

SUDO: Inferring Drug Substructures and Protein Domains Interactions from Drug Target Interactions

Content

Background

- Drugs are composed of different chemical substructures.
- Proteins are composed of functional domains.
- To infer the associations of drug substructures and protein domains governing drug target interactions.

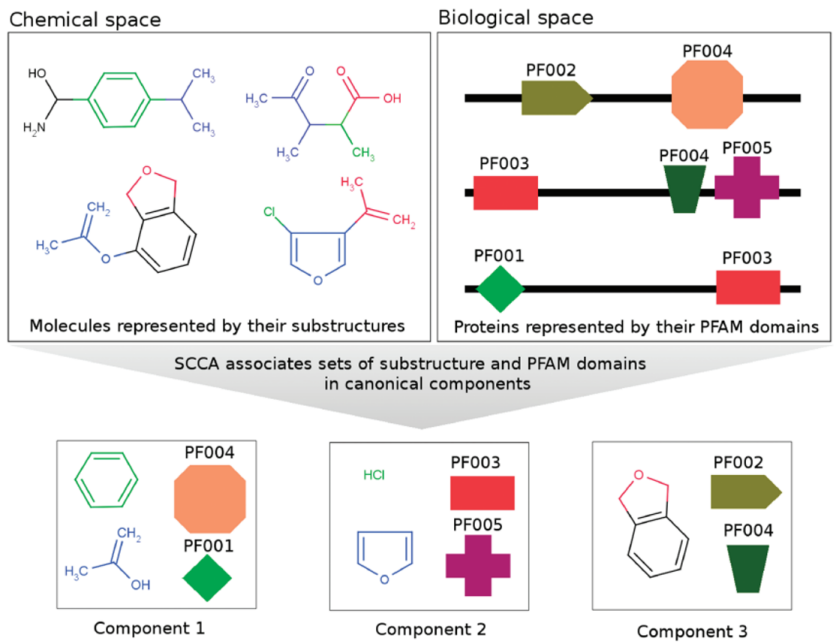


- Yoshihiro *et al.*, *J.Chem.Inf.Model.* 2011, SCCA method.
- Yasuo *et a.l, Bioinformatics* 2012, *L*1 regularized SVM

Background

- Drugs are composed of different chemical substructures.
- Proteins are composed of functional domains.
- To infer the associations of drug substructures and protein domains governing drug target interactions.

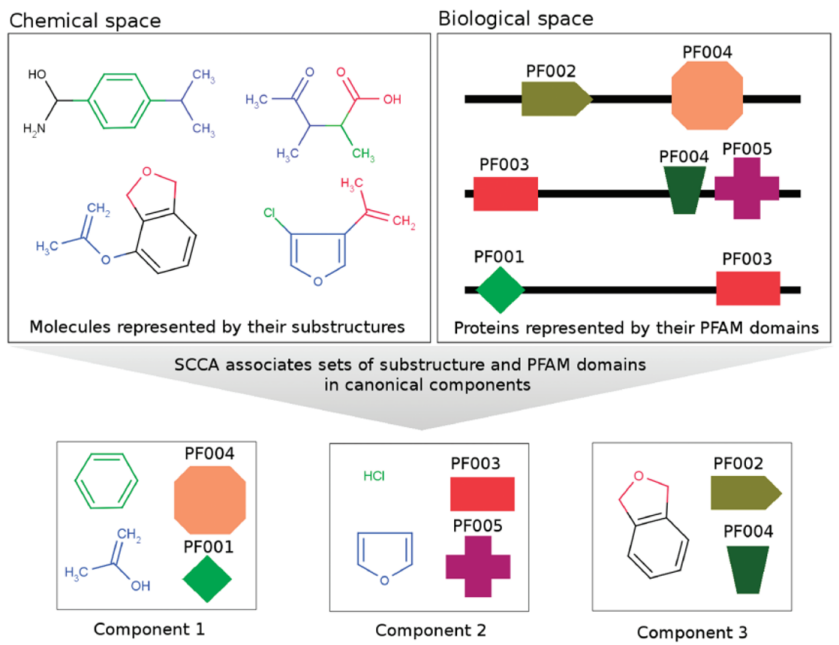
- Yoshihiro *et al.*, *J.Chem.Inf.Model.* 2011, SCCA method.
- Yasuo *et al.l, Bioinformatics* 2012, *L*1 regularized SVM



Background

- Drugs are composed of different chemical substructures.
- Proteins are composed of functional domains.
- To infer the associations of drug substructures and protein domains governing drug target interactions.

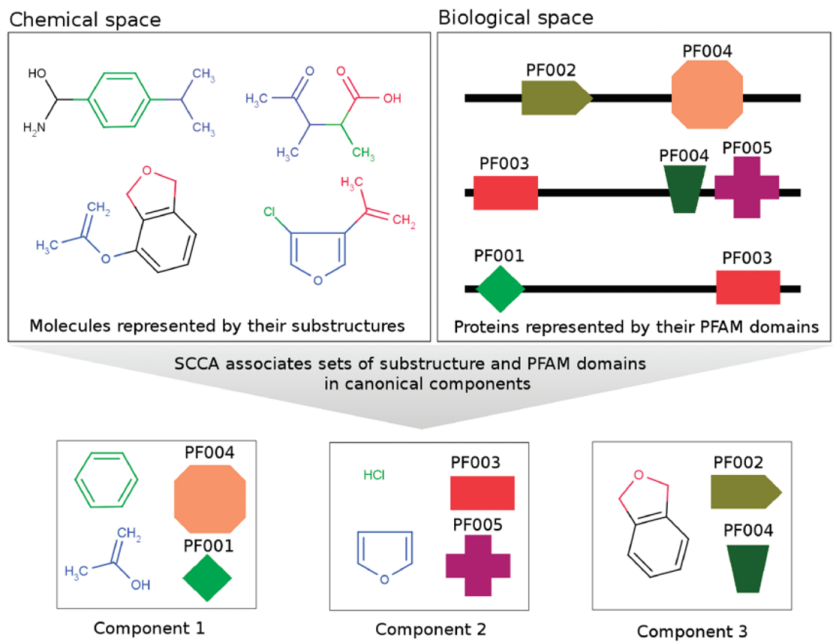
- Yoshihiro *et al.*, *J.Chem.Inf.Model.* 2011, SCCA method.
- Yasuo *et a.l, Bioinformatics* 2012, *L*1 regularized SVM



Background

- Drugs are composed of different chemical substructures.
- Proteins are composed of functional domains.
- To infer the associations of drug substructures and protein domains governing drug target interactions.

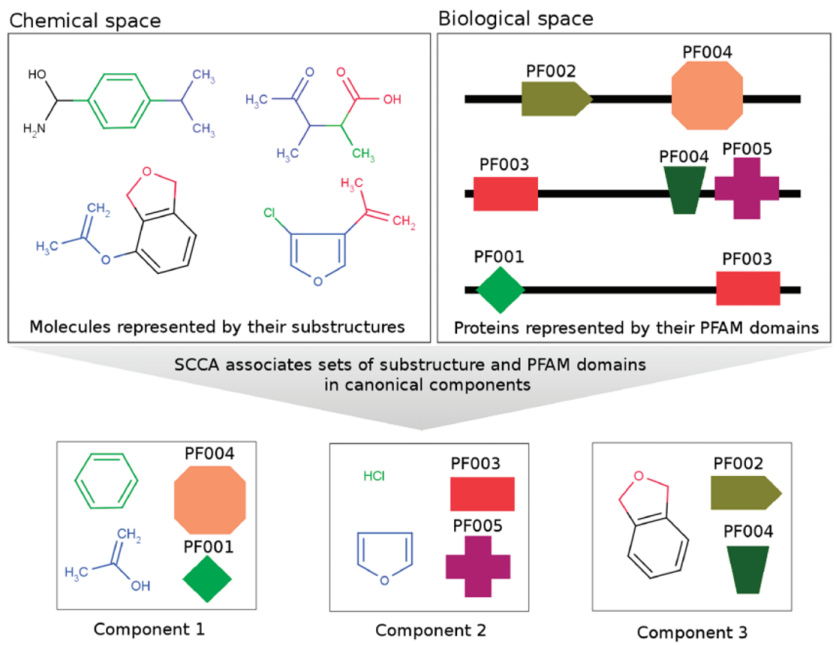
- Yoshihiro *et al.*, *J.Chem.Inf.Model.* 2011, SCCA method.
- Yasuo *et a.l, Bioinformatics* 2012, *L*1 regularized SVM



Background

- Drugs are composed of different chemical substructures.
- Proteins are composed of functional domains.
- To infer the associations of drug substructures and protein domains governing drug target interactions.

- Yoshihiro *et al.*, *J.Chem.Inf.Model.* 2011, SCCA method.
- Yasuo *et al.l, Bioinformatics* 2012, *L*1 regularized SVM



Content

Content

Definition

O_{ij} The observations of drug and protein interactions.

YP_{ij} The binary variable of drug i and protein j interactions.

$ZD_{mn}^{(ij)}$ The interaction result of substructure m from drug i and domain n from protein j

θ_{mn} $\theta_{mn} = Prob(ZD_{mn} = 1)$

fn $\text{fn} = Prob(O_{ij} = 0 | YP_{ij} = 1)$

fp $\text{fp} = Prob(O_{ij} = 1 | YP_{ij} = 0)$

Content

Probabilistic Model : SUDO

- **Reliance**: The drug protein interact if and only if at least one pair of drug substructures and protein domains interact.
- **Independence**: The interactions between drug substructures and protein domains are independent.
- Under the assumptions above,we can get:

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

YP_{ij}, the binary variable of drug i and protein j interaction in reality;
$ZD_{mn}^{(ij)}$, the binary interaction variable of substructure m and domain n.

Probabilistic Model : SUDO

- **Reliance**: The drug protein interact if and only if at least one pair of drug substructures and protein domains interact.
- **Independence**: The interactions between drug substructures and protein domains are independent.
- Under the assumptions above,we can get:

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

YP_{ij}, the binary variable of drug i and protein j interaction in reality;
$ZD_{mn}^{(ij)}$, the binary interaction variable of substructure m and domain n.

Probabilistic Model : SUDO

- **Reliance**: The drug protein interact if and only if at least one pair of drug substructures and protein domains interact.
- **Independence**: The interactions between drug substructures and protein domains are independent.
- Under the assumptions above,we can get:

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

YP_{ij}, the binary variable of drug i and protein j interaction in reality;
$ZD_{mn}^{(ij)}$, the binary interaction variable of substructure m and domain n.

Maximum Likelihood Estimation by EM algorithm

- Here we use EM algorithm to estimate θ from incomplete data according to Deng *et al.*, *Genome Research* 2002.

$$\theta_{mn}^{(t)} = \frac{1}{N_{mn}} \sum_{i,j:Zm \in Y_i, Dn \in P_j} E(D_{mn}^{(ij)} | O_{ij}, \theta^{(t-1)}) \tag{2}$$

- Comparison with the **Association Method**

$$\theta_{mn} = \frac{I_{mn}}{N_{mn}} \tag{3}$$

I_{mn} is the number of interacting drug-protein pairs containing the pair of substructure m and domain n; N_{mn} is the number of total drug protein pairs containing this pair.

Maximum Likelihood Estimation by EM algorithm

- Here we use EM algorithm to estimate θ from incomplete data according to Deng *et al.*, *Genome Research* 2002.

$$\theta_{mn}^{(t)} = \frac{1}{N_{mn}} \sum_{i,j:Zm\in Y_i, Dn\in P_j} E(D_{mn}^{(ij)} | O_{ij}, \theta^{(t-1)}) \tag{2}$$

- Comparison with the **Association Method**

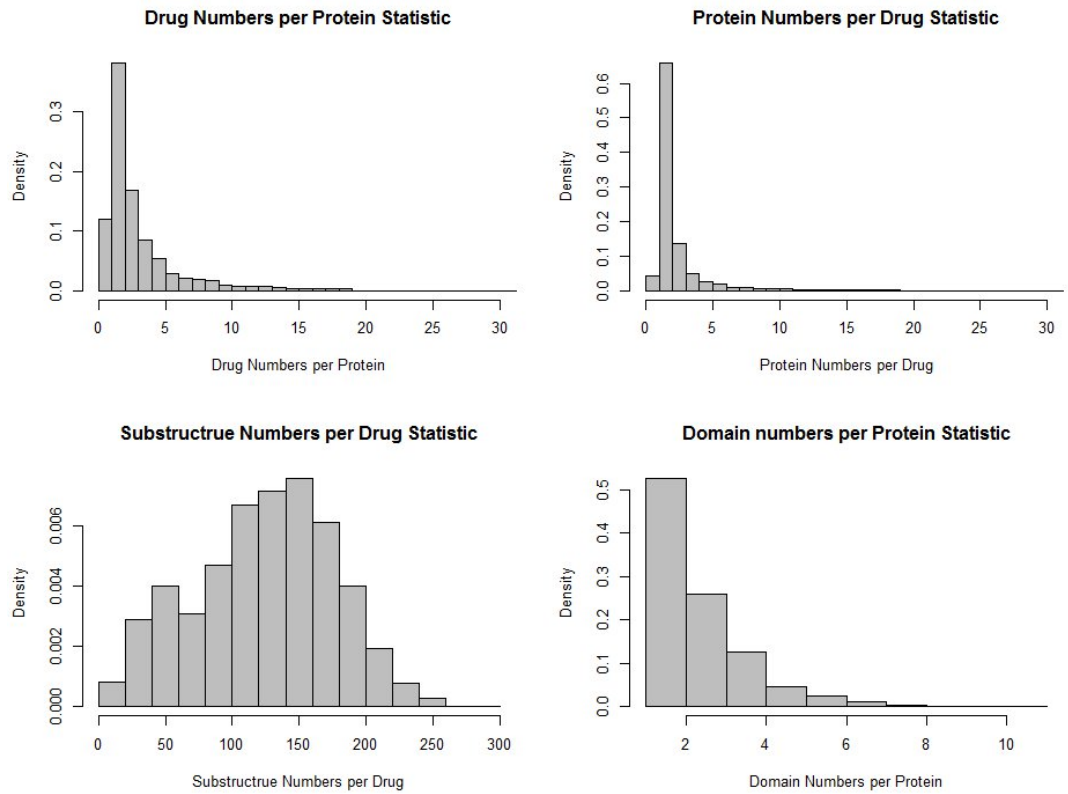
$$\theta_{mn} = \frac{I_{mn}}{N_{mn}} \tag{3}$$

I_{mn} is the number of interacting drug-protein pairs containing the pair of substructure m and domain n; N_{mn} is the number of total drug protein pairs containing this pair.

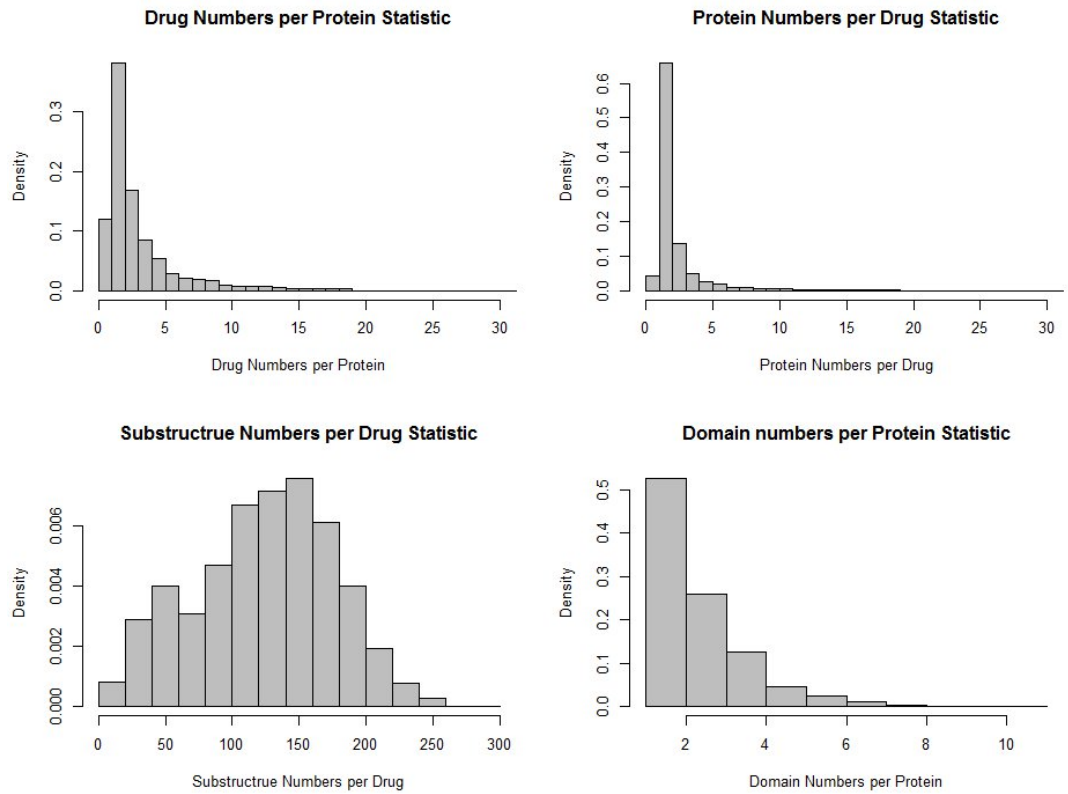
Content

Content

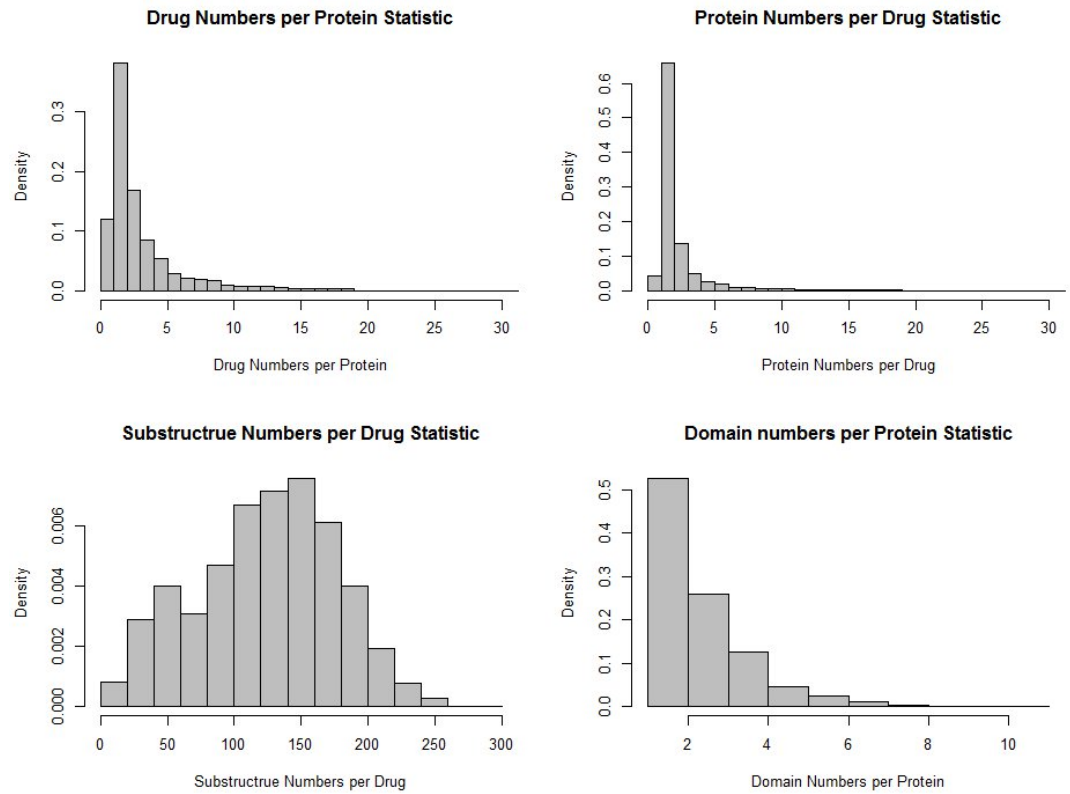
- Drugs : 440 dimensional binary vector of substructures.
- Proteins : 1071 dimensional binary vector of domains.
- Each drug has 3 targets n average, and each proteins has 2 drugs against it on average.



- Drugs : 440 dimensional binary vector of substructures.
- Proteins : 1071 dimensional binary vector of domains.
- Each drug has 3 targets n average, and each proteins has 2 drugs against it on average.



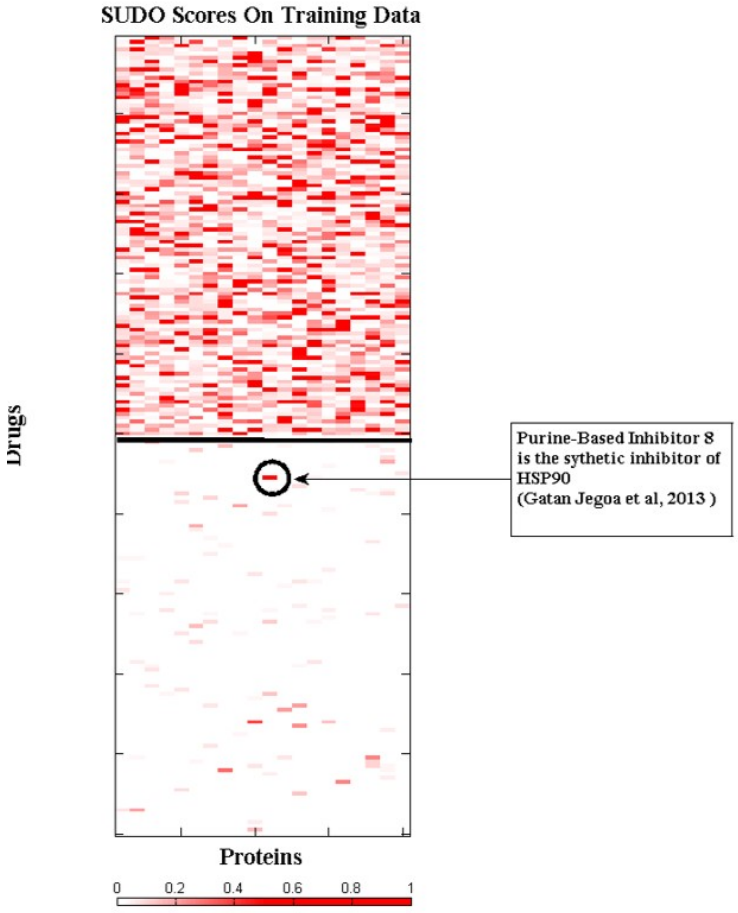
- Drugs : 440 dimensional binary vector of substructures.
- Proteins : 1071 dimensional binary vector of domains.
- Each drug has 3 targets n average, and each proteins has 2 drugs against it on average.

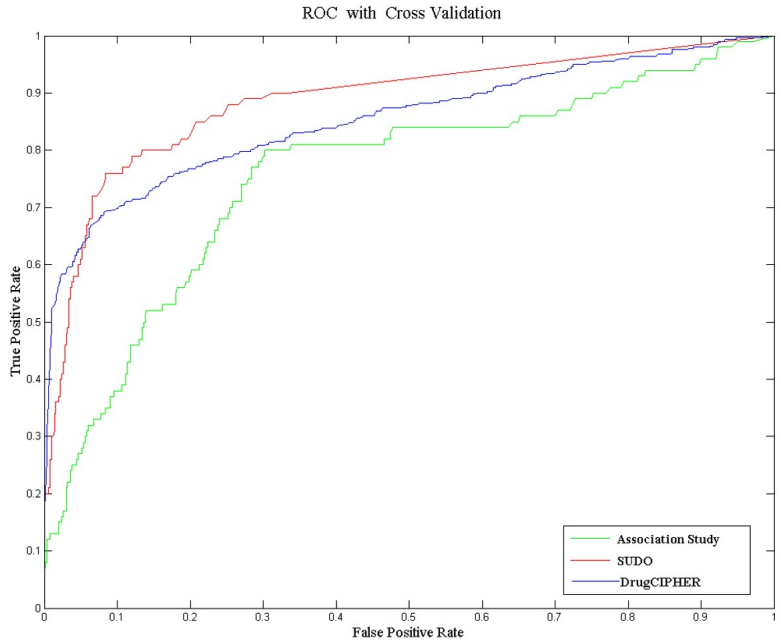


Content

ROC Curve with Cross Validation

The associations of drug substructures and protein domains from SUDO can predict drug targets well.





Content

Substructures and Domains Association Network from SUDO

Substructures or domains with High Degrees In Substructure Domain Association Network

Substructures In PubChem	Pfam Domains
N-P	Thrombin light chain
>=4 Cl	Acetylcholinesterase tetramerisation domain
>= 2 saturated or aromatic heteroatom-containing ring size 4	Oestrogen receptor
>= 4 P	Hsp90 protein
>= 3 unsaturated non-aromatic carbon-only ring size 6	Cytochrome c

The Pairs of Substructures and Domains with high scores From SUDO

Substructure In PubChem	Pfam Domain
>= 1 K	K-Cl Co-transporter type 1 (KCC1)
>= 1 Mg	Neuronal voltage-dependent calcium channel alpha 2acd
>= 2 P	Polyprenyl synthetase
>= 2 P	Nucleoside diphosphate kinase
>= 3 any ring size 5	short chain dehydrogenase
S(~C)(~H)	Glutathione peroxidase
>= 3 any ring size 5	Heme oxygenase

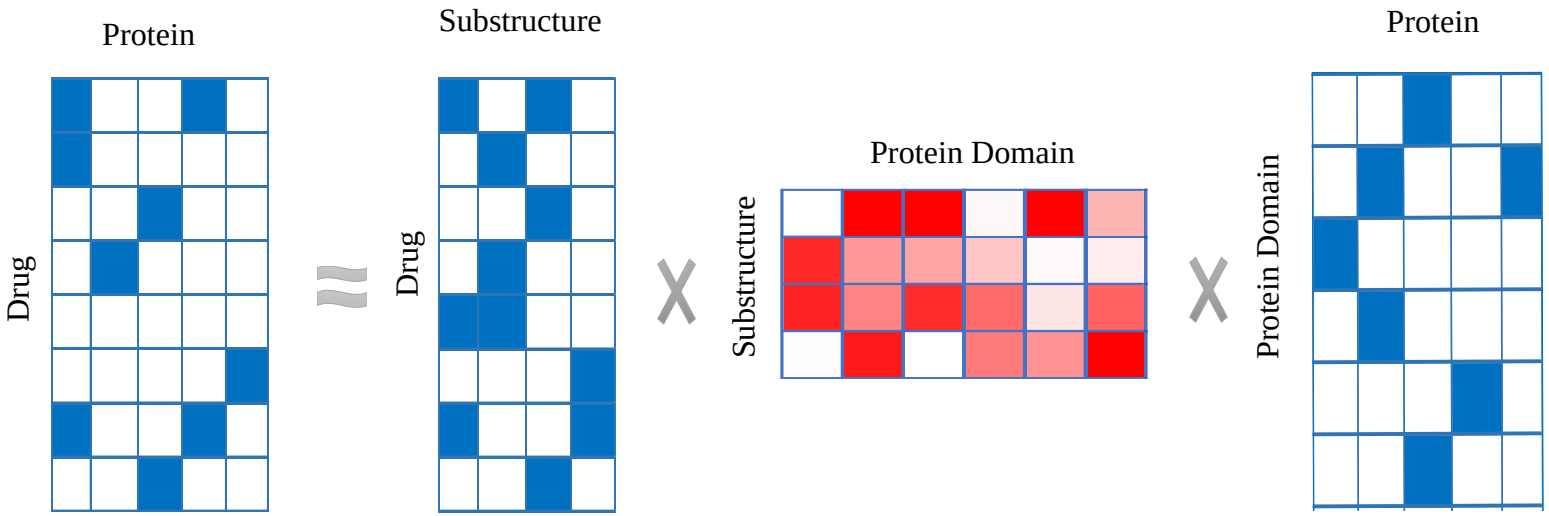
Drug Substructure

Protein Domain

Association Edege From SUDO Scores

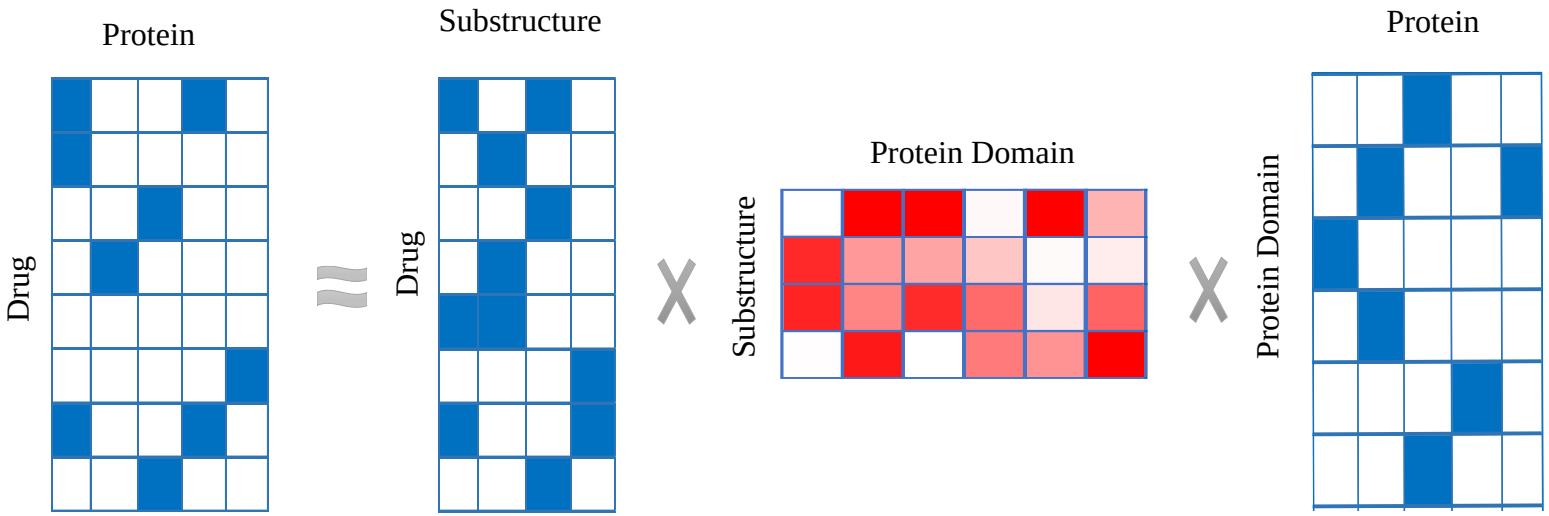
Content

- Here we proposed a MLE approach to extract the pairs of substructures and domains that can predict drug targets well.



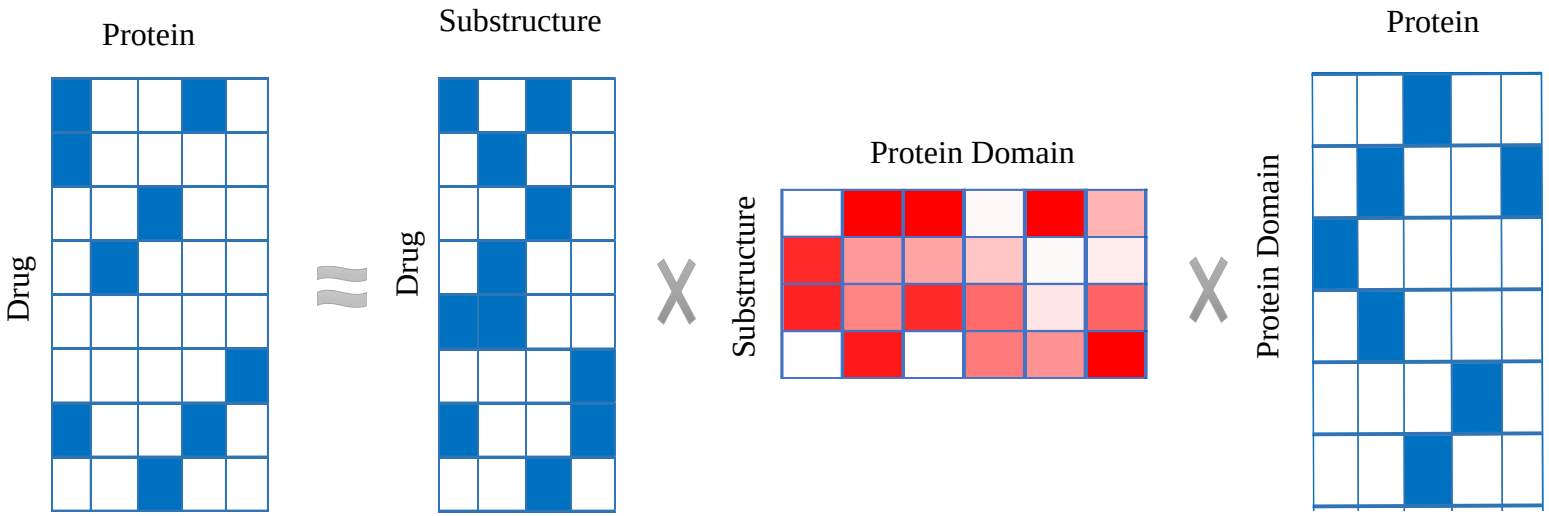
- Incompleteness of data is considered.
- Ignored the combination of substructures.

- Here we proposed a MLE approach to extract the pairs of substructures and domains that can predict drug targets well.



- Incompleteness of data is considered.
- Ignored the combination of substructures.

- Here we proposed a MLE approach to extract the pairs of substructures and domains that can predict drug targets well.



- Incompleteness of data is considered.
- Ignored the combination of substructures.

Part II

Incorporation of Ligand-based and Network-based Approaches for Prediction of Drug Target Interactions

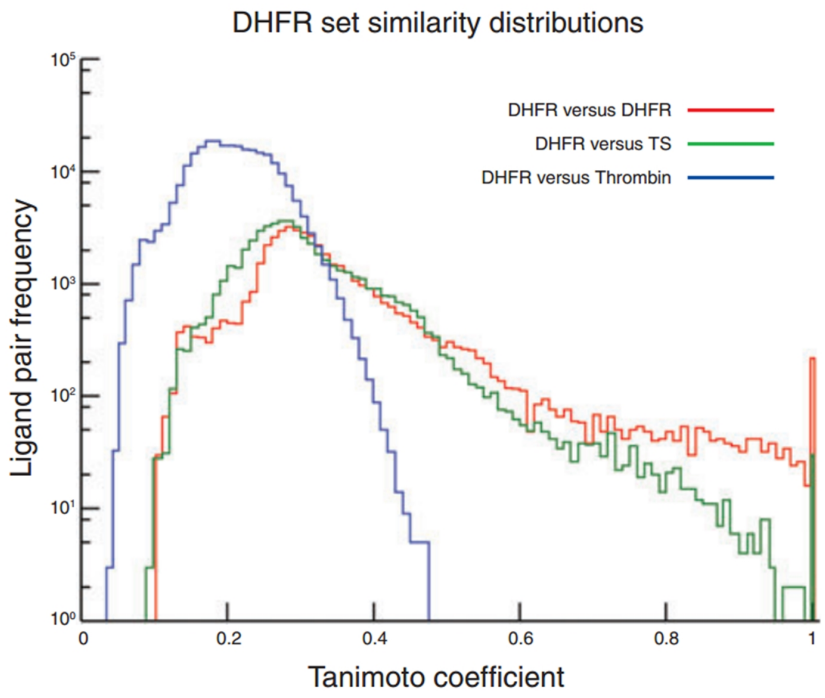
Content

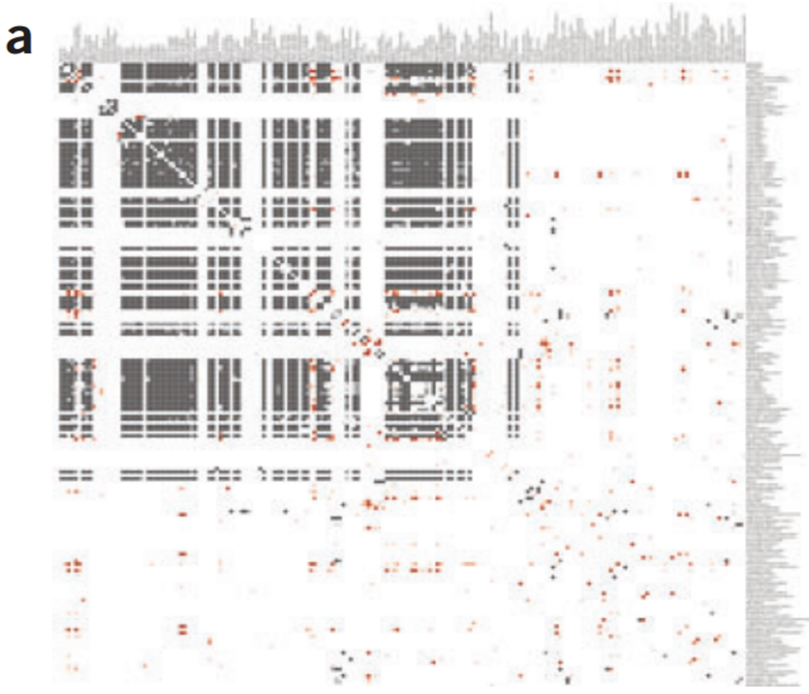
Content

Ligand-based prediction

Relating protein pharmacology by ligand chemistry.

- SEA(Keiser *et al, Nature* 2009;Lounkine *et al, Nature* 2012)

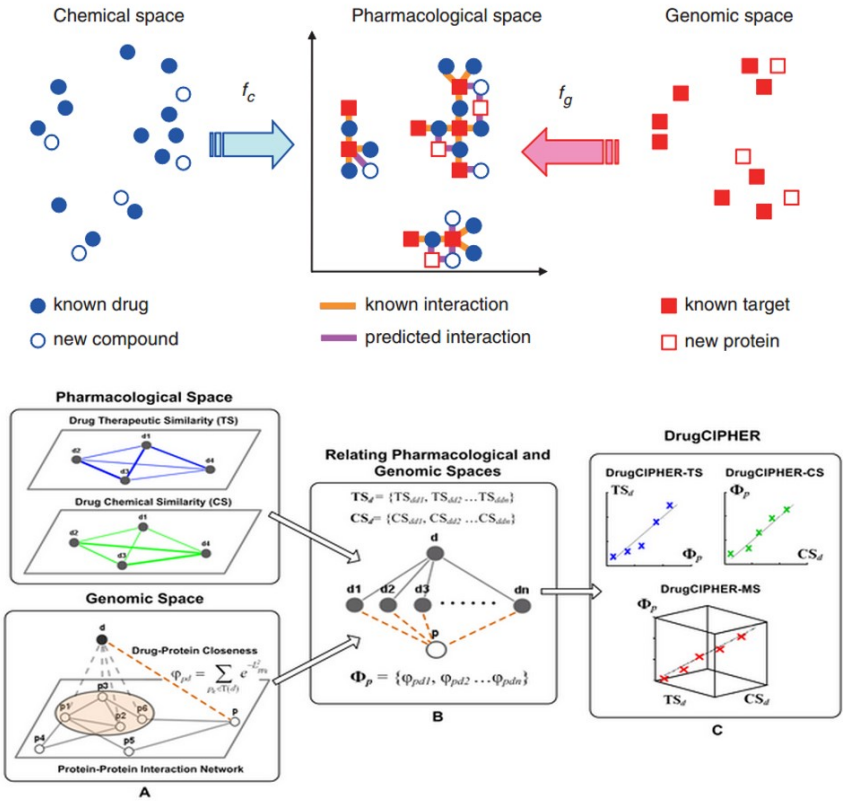




Network-based prediction

Integrating genomic and chemical information into a heterogenous network.

- Yamanishi, Yoshihiro *et al, Bioinformatics* 2008 integrated of chemical and genomic spaces.
- Zhao, Shiwen *et al, Plos One* 2010 integration of chemical and pharmacological spaces.
- Cheng, Feixiong *et al, Plos Computational Biology* 2012 prediction directly on drug targets bipartite newtork.



Content

Limitations of Current Approaches

Network-based approaches

- The known drug-protein pairs are limited.
- Low sensitivity and lack of experimental evaluation.

Ligand-based approach

- Only learning from themselves.
- Limited by the number of active ligand against the protein.
- Difficulty in identifying novel structural scaffolds.

Limitations of Current Approaches

Network-based approaches

- The known drug-protein pairs are limited.
- Low sensitivity and lack of experimental evaluation.

Ligand-based approach

- Only learning from themselves.
- Limited by the number of active ligand against the protein.
- Difficulty in identifying novel structural scaffolds.

Limitations of Current Approaches

Network-based approaches

- The known drug-protein pairs are limited.
- Low sensitivity and lack of experimental evaluation.

Ligand-based approach

- Only learning from themselves.
- Limited by the number of active ligand against the protein.
- Difficulty in identifying novel structural scaffolds.

Limitations of Current Approaches

Network-based approaches

- The known drug-protein pairs are limited.
- Low sensitivity and lack of experimental evaluation.

Ligand-based approach

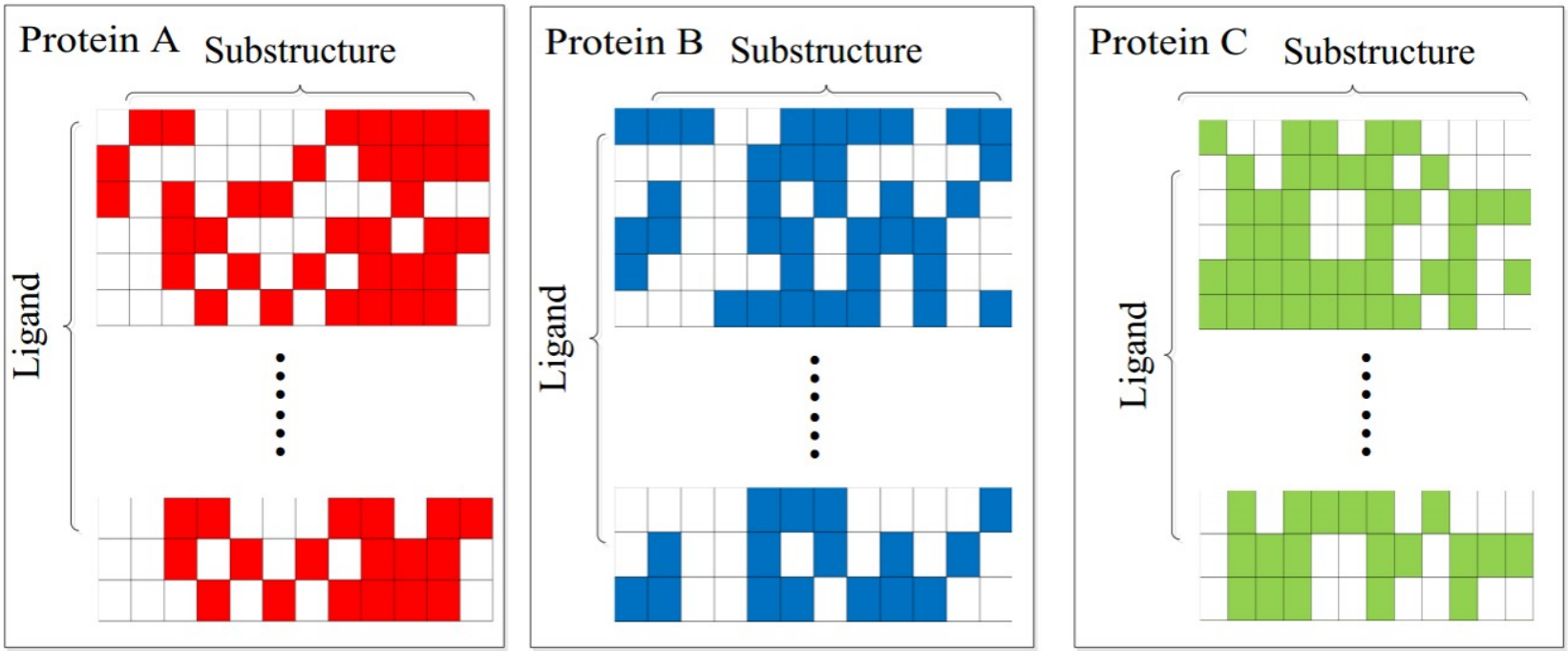
- Only learning from themselves.
- Limited by the number of active ligand against the protein.
- Difficulty in identifying novel structural scaffolds.

Content

Content

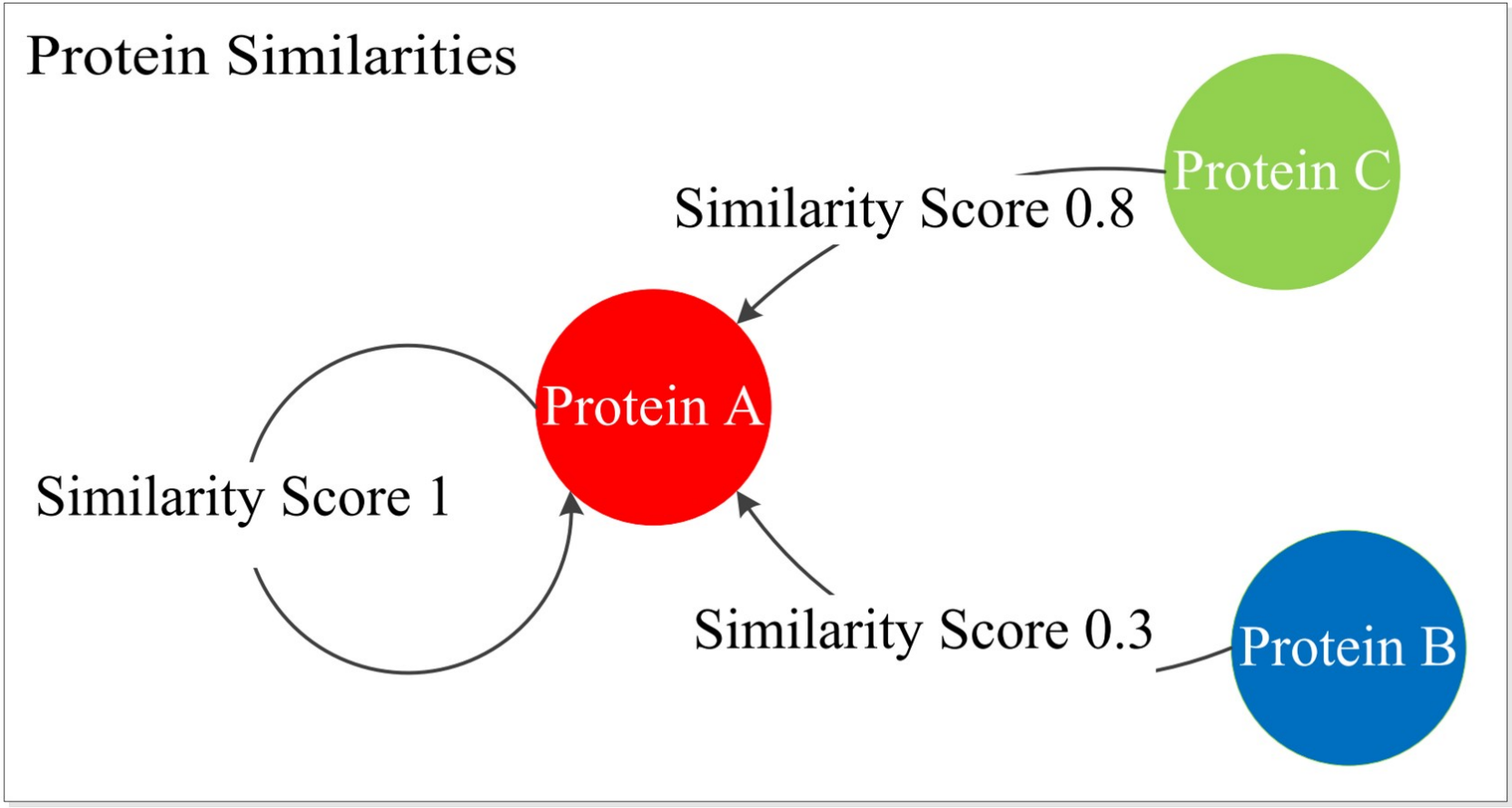
Data from the ligand-based view

- Each protein has lots of ligands that bind to it.

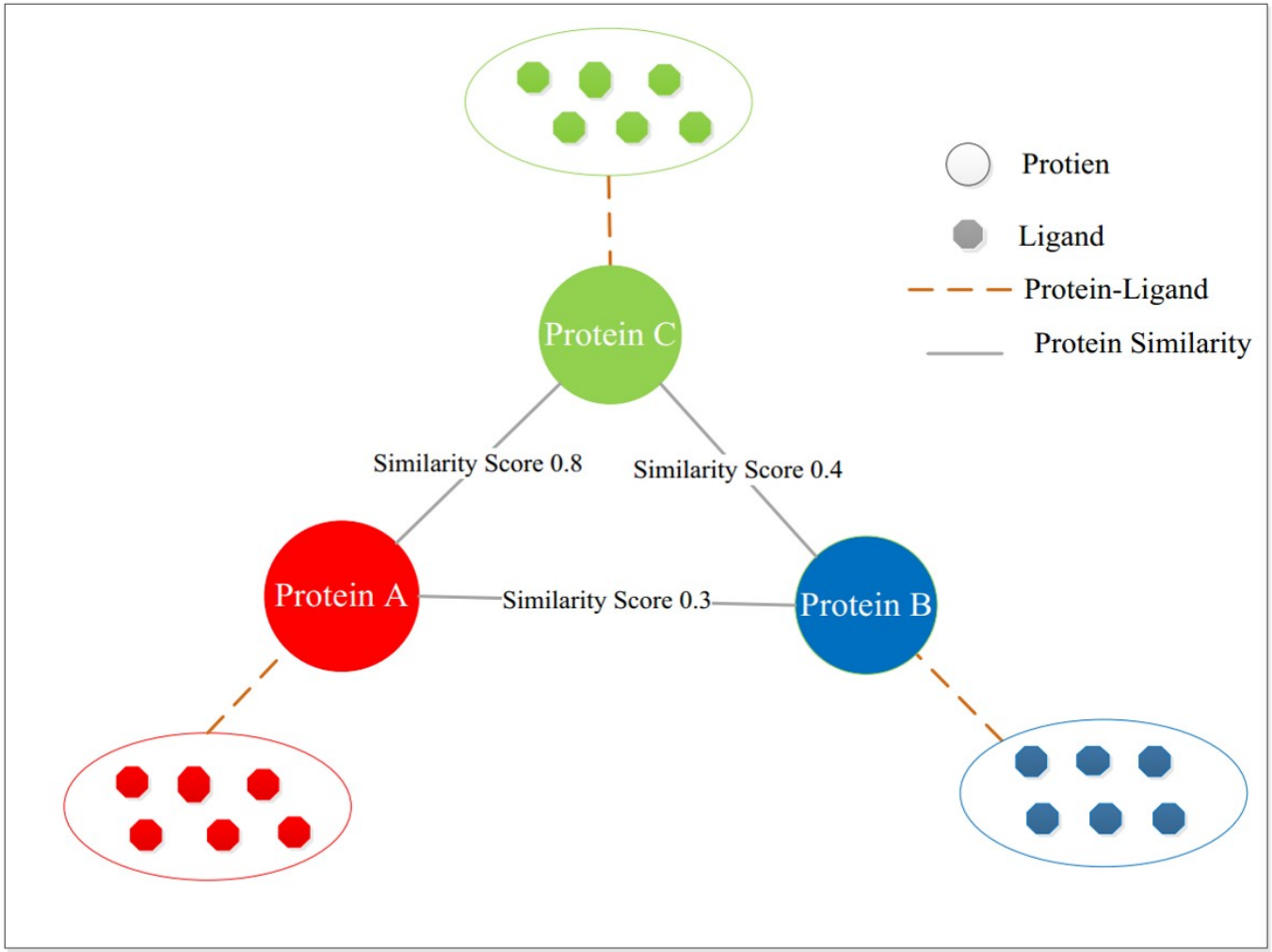


Data from the network-based view

- Proteins also show sequence or structural similarities.



Data from the integrative view



Content

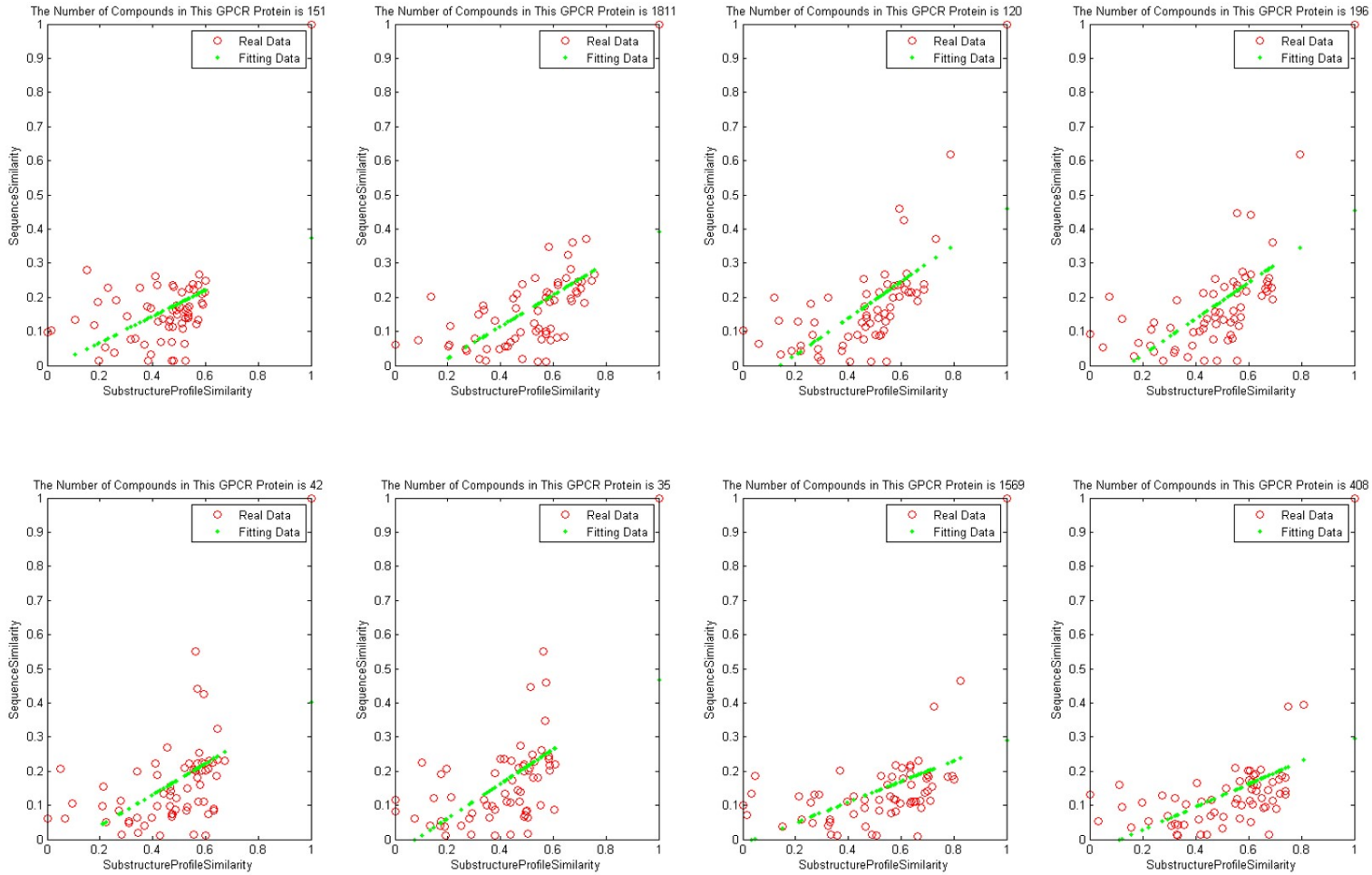
Assumptions

- Targets similar in sequence or structure should show similar ligand substructure patterns.

$$Distance^2_{(ij)} = \sum_n \left(\frac{\sum_m X_n^{(m)}}{\#\{m\}} - \frac{\sum_{m'} X_n^{(m')}}{\#\{m'\}}\right)^2 \tag{4}$$

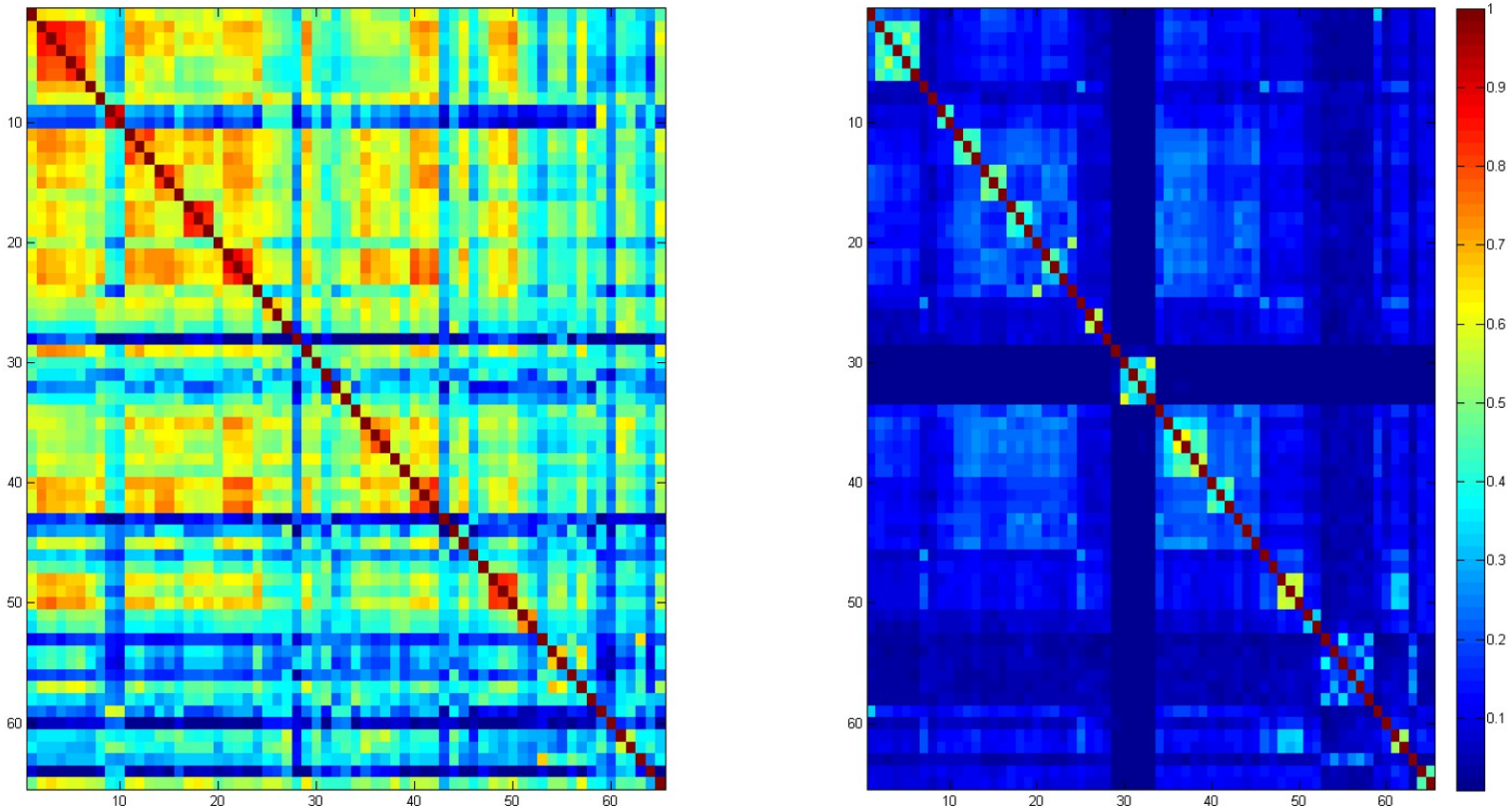
$X_n^{(m)}$: the binary value of substructure n in ligand m;
m: ligand m from target i;
m' :ligand m' from target j;
n: the chemical substructure n.

- We can see a significant correlation between ligand-based similarities and sequence-based similarities.



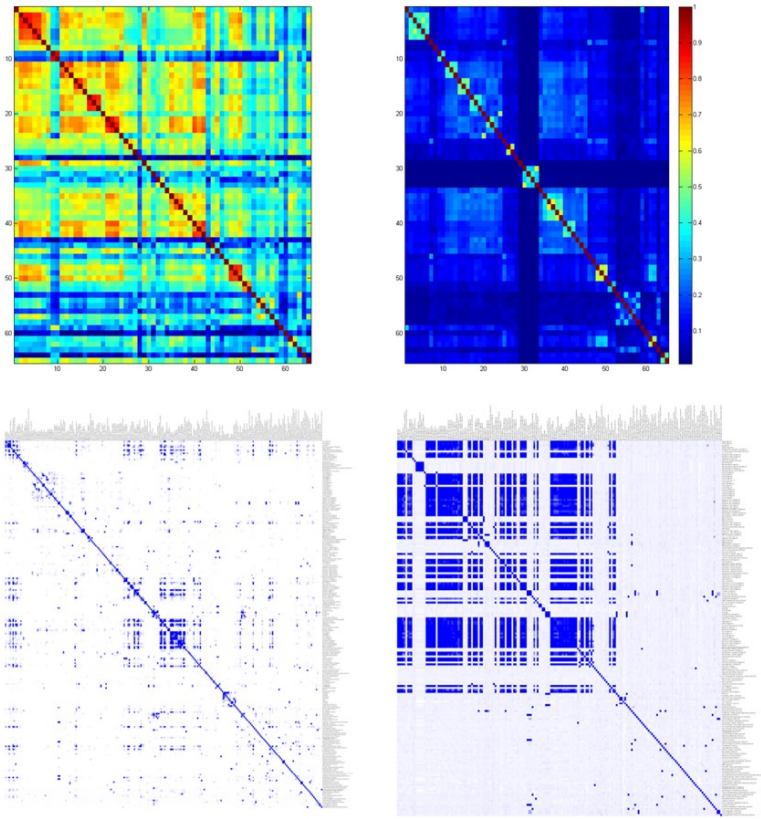
Heat Map of Similarity Matrix from GPCR Proteins

- The left matrix is ligand-based similarity matrix and the right one is sequence-based similarity matrix.
- It shows similar patterns.

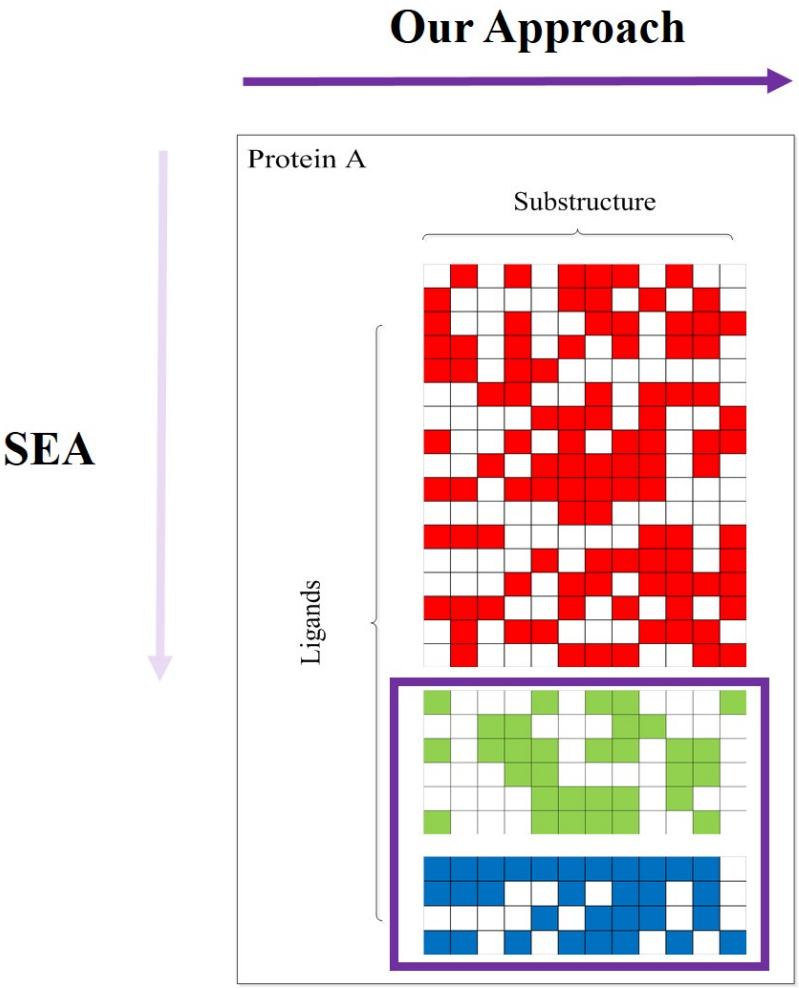


Compared with SEA

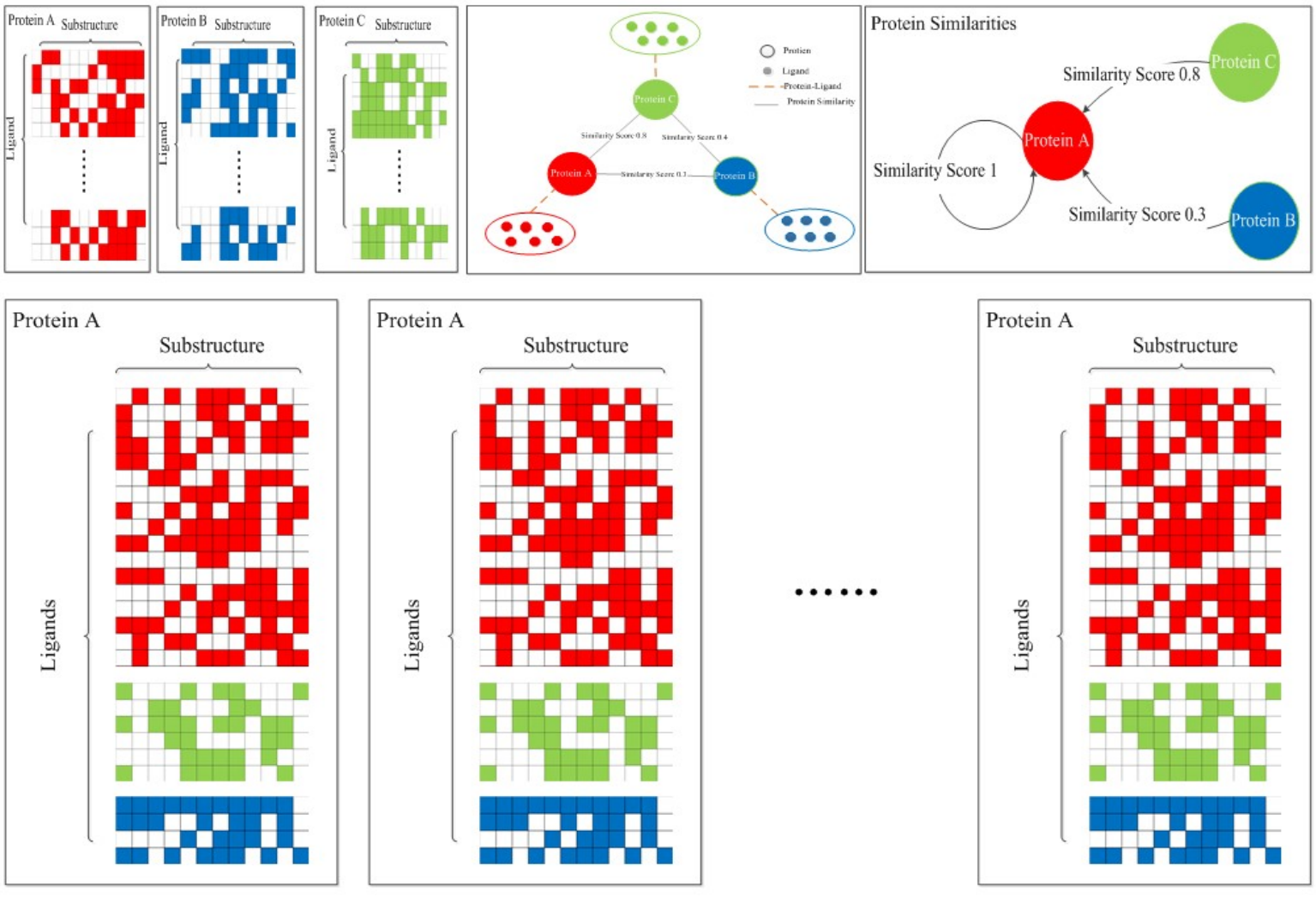
- In SEA (Lounkine *et al*, *Nature* 2012), it shows no significant correlation between ligand-based similarity matrix and sequence-based similarity matrix among GPCR proteins



- Compared with SEA, we see the ligand-based similarity from a different way.
- Another reason may be that SEA treats the similarity as the binary variable.



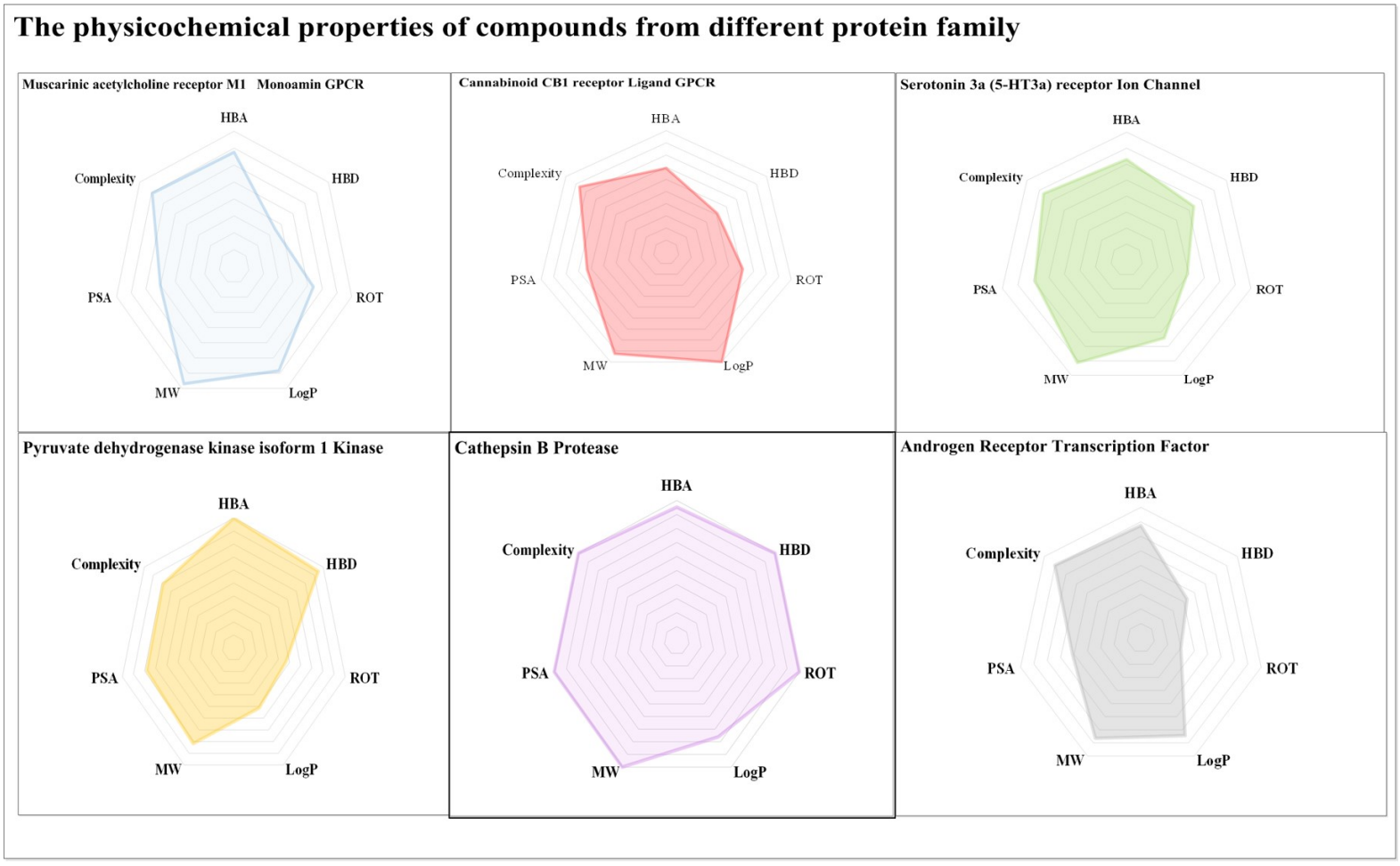
Weighted Resample



Content

- Compound promiscuity is mostly connected to related targets. (Ye Hu *et al, Drug Discovery Today* 2013).
- Compound molecular properties can be used to classify targets class. (Yusof, Iskander *et al, Drug Discovery Today* 2013)

Compound molecular properties can be used to classify targets class



Part III

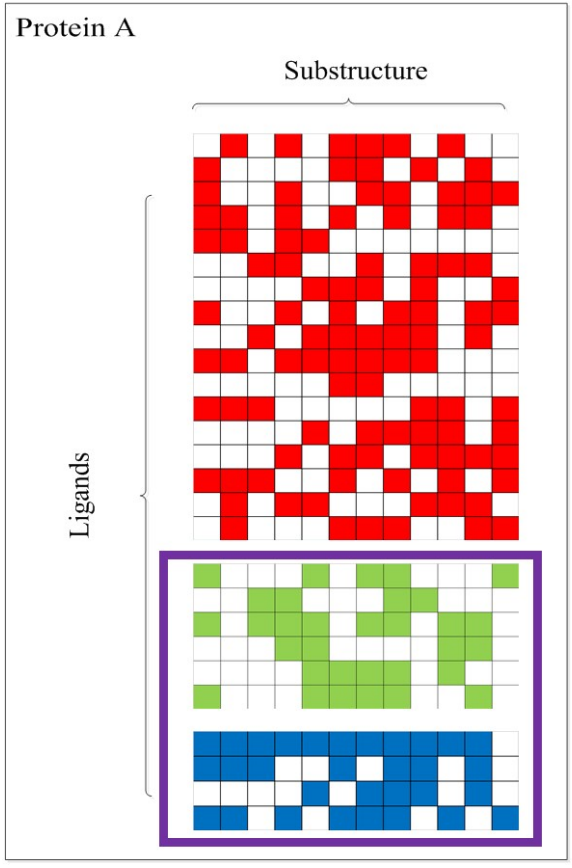
Summary and Future Plan

Summary

- Here we proposed two ways to analyze drug targets relation.
- Chemical substructures deepen the understanding of drug polypharmacology.

SEA

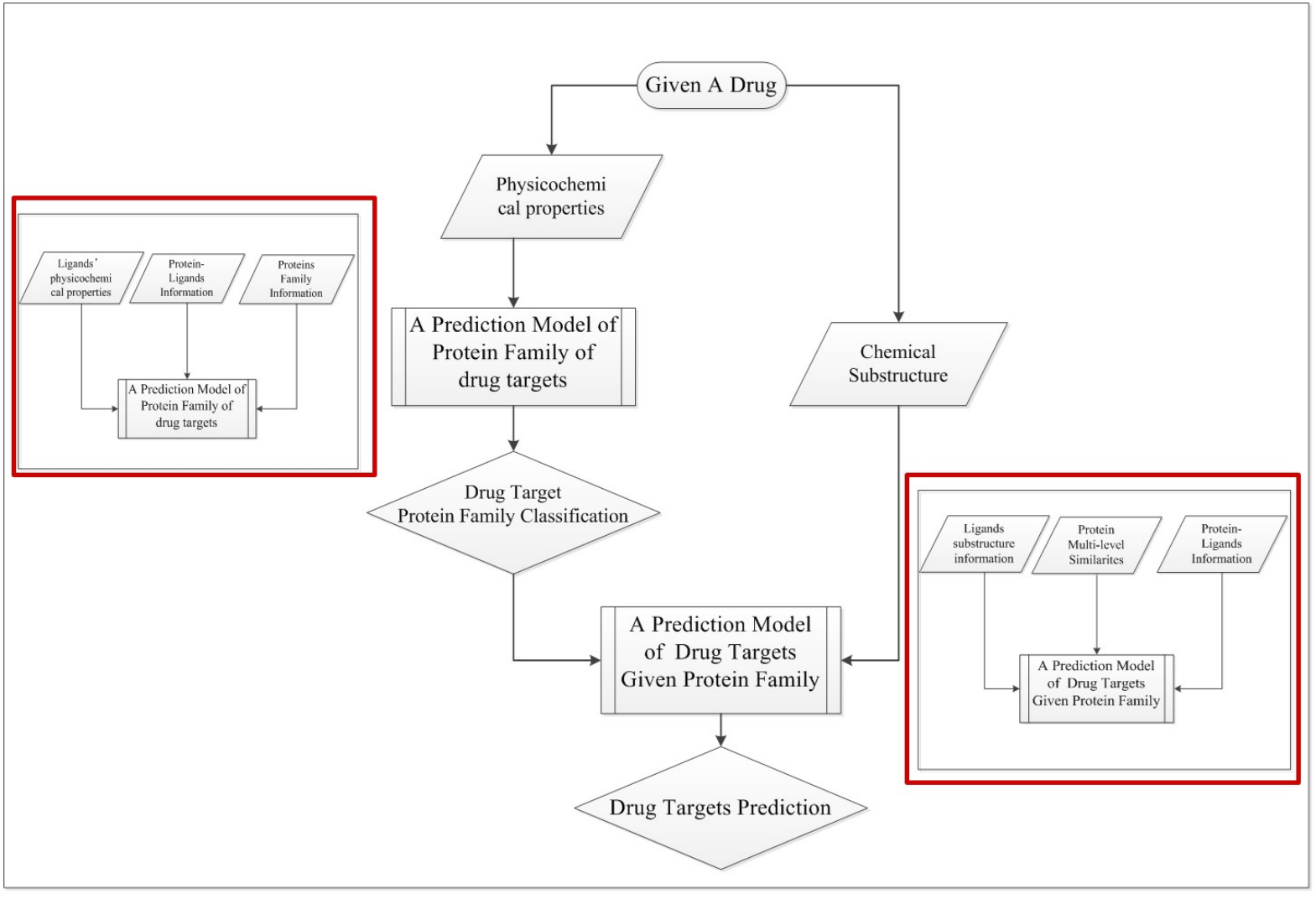
Our Approach



Future Plan

- Complete the algorithm of LigNet.
- Analyze the substructure profile features of different proteins.

Flow Chart of LigNet



Thank you !

Probabilistic Model: SUDO

- Then

$$
\begin{aligned}
&Pr(O_{ij} = 1|\theta) \\
&\quad = Pr(O_{ij} = 1|YP_{ij} = 1)Pr(YP_{ij} = 1|\theta) \\
&\qquad + Pr(O_{ij} = 1|YP_{ij} = 0)Pr(YP_{ij} = 0|\theta) \\
&\quad = (1 - fn)Pr(YP_{ij} = 1|\theta) + fp(1 - Pr(YP_{ij} = 1|\theta))
\end{aligned}
$$

- The log likelihood function of the observed data is

$$
\begin{aligned}
l(\theta) &= log(Pr(O|\theta)) \\
&= log(\prod_{all\ i,j} Pr(O_{ij} = 1|\theta)^{O_{ij}} Pr(O_{ij} = 0|\theta)^{1-O_{ij}}) \qquad (5)
\end{aligned}
$$

Estimation of fn and fp

- fp is set as 10^{-4} for the known drug-target pairs are credible.
- The estimation of each drug's average targets number is about 6(Mestres *et al*,,2008),then

$$
\begin{aligned}
fn &= Pr(O_{ij} = 0 | YP_{ij} = 1) \\
&= 1 - \frac{Pr(O_{ij} = 1, P_{ij} = 1)}{Pr(P_{ij} = 1)} \\
&\geq 1 - \frac{Pr(O_{ij} = 1)}{Pr(P_{ij} = 1)} \\
&\geq 1 - \frac{\text{number of observed interaction pairs}}{\text{number of real interaction pairs}} \\
&\geq 0.80
\end{aligned} \tag{6}
$$

fn is set as 0.85.

How to evaluate the performance of the algorithm

- The performance of target classification
  - Cross Validation
  - Compare with the results of prediction without classification.
- The performance of drug targets prediction by LigNet.
  - Compare with SEA with experiment data.

Ligand-based similarity in SEA

- Roughly,the ligand-based similarity in SEA is defined as followed:

$$Distance_{ij} = \sum_{m}\sum_{m'} \frac{\sum_n (X_n^{(m)} - X_n^{(m')})^2}{\sum_n (X_n^{(m)} + X_n^{(m')} - X_n^{(m)} \cdot X_n^{(m')})} \quad (7)$$

$X_n^{(m)}$: the binary value of substructure n in ligand m;
m: ligand m from target i;
m' :ligand m' from target j;
n: the chemical substructure n.