Supplementary Materials

Songpeng Zu1[[1]](#footnote-1), Ting Chen2, Shao Li1\*

1Bioinformatics Division, TNLIST and Department of Automation, Tsinghua University, Beijing 10084, China

2Molecular and Computational Biology Program, Department of Biological Science, University of Southern California, Los Angeles, California 90089, USA

Table of Contents

[Supplementary Materials 1](#_Toc411819690)

[Data Sources 3](#_Toc411819691)

[Methods 4](#_Toc411819692)

[The EM framework 4](#_Toc411819693)

[Derivation of EM algorithm in our model 5](#_Toc411819694)

[The Association Method 6](#_Toc411819695)

[Variance Estimation of the EM results. 7](#_Toc411819696)

[Combinations of drug chemical substructures 9](#_Toc411819697)

[Estimation of fn and fp. 10](#_Toc411819698)

[Results of the predicted drug-domain interactions 10](#_Toc411819699)

[Reference 12](#_Toc411819700)

## Data Sources

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[2], and 1554 proteins are represented by 876-dimensional protein domain binary vectors from the Pfam database[3]. 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.

## Methods

### The EM framework

Here we propose a probabilistic model to infer the substructure-domain interactions. It is inspired by the work [7], and follows the assumptions below:

1. The interactions between drug chemical substructures and protein domains in the pair of drug-target interactions are independent.
2. A drug and a protein interact if and only if at least one pair of the drug chemical substructures and the protein domains interact.

Let Yi represent the drug i, Pj represent the protein j, Zm represent the chemical substructure m, and Dn represent the domain n. Let denote whether the pair of the chemical substructure m from the drug i and the protein domain n from the protein j interact (the value is one) or not (zero otherwise). We use to present the interaction possibilities of the chemical substructure m and the protein domain n, that is,

Then our aim is to evaluate the interaction possibilities of the chemical substructures and the protein domains.

Under the Assumption 1 and 2, we can get

In which, denote whether or not the drug Yi and the protein Pj interact or not. If equals to 1, then they interact, and 0 otherwise.

Since we know many drug-protein interactions remain unknown, which means YP cannot directly be used to denote the observed drug-protein interactions data, we then use the O = {Oij} (where Oij denote whether the drug i and the protein j interact or not, one for interaction, zero otherwise) to represent the given drug-protein interactions. In addition, in order to connect YP with O, we introduce two parameters, namely, the false negative rate fn and the false positive rate fp defined below:

Then, we can get

Moreover, the log likelihood function, i.e., the total probability of the observed drug-protein interactions data is

In which , where fn and fp are predefined.

Then our aim is to estimate based on the maximum likelihood estimation (MLE). However, because we don’t know whether (which means the chemical substructure m from drug i interact with the protein domain n from protein j or not), this is a missing data problem. It is naturally to solve the problem by EM algorithm [6]. It follows:

* The E step is :
* The M step is :

Note that is the total number of drug-protein pairs that contain the chemical substructure m and the protein domain n.

### Derivation of EM algorithm in our model

Here we would show how to derive EM algorithm in our model. In general, two steps are involved in EM algorithm:

* E step:
* M step:

In which, X represents the observed and incomplete data. (X,Y) then are the complete data, while Y is the latent data.

In our model,

Not that the first summation has nothing to do with , while the last summation can be rewritten as followed:

Then,

Let the formula above equals to zero, we can finally get our EM procedure.

### The Association Method

One of the problem of EM algorithm is that it converge to a local minimum and different initial values usually result in different local minimums. Instead of randomly choosing the initial values many times, here we used the association model to choose the initial values, which is a local way to evaluate the possibilities of drug substructures and protein domains interactions.

It follows,

in which Imn is the number of interacting pairs of drug-protein pairs containing the pair of chemical substructure Zm and protein domain Dn and Nmn is the number of total drug-protein pairs containing the pair ofchemical substructure Zm and protein domain Dn.

This method has two limitations.

* Firstly, it computes the chemical substructure-protein domain interactions locally, which means it ignores other interactions between the chemical substructures and protein domains in the same drug-protein pairs. For example, drug Yi containing substructures {Zm ,Zy } interacts with both protein Pj containing domains {Dn ,Dy} and protein Pk containing domains {Dn ,Dc }. Substructure Zm and protein domainDn do not appear in any other drugs and proteins, respectively. Then = 2/2 = 1. It obviously ignores other interactions between substructures and protein domains such as substructures Zn interacting with protein domain Dc. Therefore, to infer drug substructure and protein domain interactions, we should consider all the drug protein interactions and all the interactions between drug substructures and protein domains.
* Secondly, this method relies on the accuracy of observed data. However, current drug-protein data are largely incomplete.

### Variance Estimation of the EM results.

The natural way to estimate the variance of the maximum likelihood estimation is followed [9]:

is the observed information.

In our situation, we derive the observed information below.

Since

And

Where

Then,

Also, let

We can get

Then

Note that,

Besides,

Finally, we have

### Combinations of drug chemical substructures

Different drug chemical substructures may take functions as one unit in the drug-protein interactions. We try to estimate the combination behaviors between two drug chemical substructures on drug-protein interactions through adding the pairs of drug chemical substructures as the new “drug chemical substructures”, we handle this problem by our probabilistic model.

Instead of considering all the pairs of drug chemical substructures, we firstly use a filter method to select those pairs of drug chemical substructures that significantly appear in the drug-protein interactions. The filter method follows two steps: we use hypergeometric distribution to detect whether the co-appearing times of two drug chemical substructures are significant or not. Then we move out the pairs of drug chemical substructures that also significantly appear in the randomly selected compounds.

The reason why we follow this procedure is that we are only interested in the combinations of drug chemical substructures that can interact with proteins but not often co-exists in the compound chemical space. The randomly selected compounds are from CHEMBL database, and in total, there are over 9,000 compounds representing the compound chemical space. We use the Bonferroni adjustment for the multiple test corrections here. Note that due to the PubChem substructures definition, we only consider the pairs with SMARTS records. Finally, we select 1870 pairs of drug chemical substructures for learning.

## Estimation of fn and fp.

In our model, two parameters, i.e., fn and fp, should be predefined. According to our model,

It is shown that on average the number of target proteins per drug is about 6.3[8]. Then we can get,

We can estimate fp, which equals to , in the similar way.

In order to analyze our model robustness to these parameters, we used five folds cross validation to detect the recoveries of drug-protein interactions on different combinations of fn and fp. It showed that the performances of recovering drug-protein interactions kept stable on different combinations of fn and fp.

The procedure are followed: (i) split the original drug-target interactions equally to five fold. (ii)Each time, we select one of them as the test data set and use the others as the training set in our model. (iii) Estimate the test set after learning by the area under the operating characteristic curve (AUC). The curve is generated by plotting the false positive rate in the x-axis versus true positive rate. Note that the negative samples are randomly selected from the known non-interacted drug protein pairs, since we do not have the real negative samples.

## Results of the predicted drug-domain interactions

Table 1. Results of the prediction on drug-domain interactions. k value is the proportion of the number of the binding site residues lying within the protein domain over the total number of the binding site residues. When k value was larger than 0.5, we treated the drug interacted with the domain. TRUE means the drug was predicted to interact with domain by GIFT.

## Reference

1. Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K., & Yamanishi, Y. (2012). Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*, *28*(18), i487-i494.
2. [Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN *Annual Reports in Computational Chemistry*, Volume 4, Elsevier: Oxford, UK; 2008, pp. 217-240.](http://oldwww.acscomp.org/Publications/ARCC/volume4/chapter12.html)
3. [The Pfam protein families database](http://nar.oxfordjournals.org/content/40/D1/D290.full):M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn *Nucleic Acids Research* (2012)  Database Issue 40:D290-D301
4. Finn, R. D., Miller, B. L., Clements, J., & Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic acids research*, *42*(D1), D364-D373.
5. Kruger, F. A., Rostom, R., & Overington, J. P. (2012). Mapping small molecule binding data to structural domains. *BMC bioinformatics*, *13*(Suppl 17), S11.
6. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, *39*(1), 1-38.
7. Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring domain–domain interactions from protein–protein interactions. *Genome research*, *12*(10), 1540-1548.
8. Mestres, J., Gregori-Puigjane, E., Valverde, S., & Sole, R. V. (2008). Data completeness—the Achilles heel of drug-target networks. *Nature biotechnology*, *26*(9), 983-984.
9. Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information.*Biometrika*, *65*(3), 457-483.

1. Contact: zsp07@mails.tsinghua.edu.cn [↑](#footnote-ref-1)