Predicting Compound-Protein Interactions By Transferring Information.

Songpeng Sam Zu[[1]](#footnote-1)

2014-04-30

## *Abstract*

It is a key problem to evaluate the compounds’ protein targets in the early stage of drug discovery and drug design. Many *in silico* methods have been used to predict compound-protein interactions due to the high cost and high risk experiments.

In this paper, we develop a Bayesian Hierarchical model to predict compound-protein interactions (CPIs).The main innovations are : a) we integrate both chemoinformation and genomic information, i.e., large-scale compound-protein interactions information, compounds 2D structure fingerprints, chemo-physics information and several levels protein information, such as proteins family information, proteins sequence information, functional information and protein-protein interactions network information. b) instead of predicting whether or not they interact like most papers do, here we can predict the CPIs potencies. c) last but most importantly, we predict compound-protein interactions by ligand-based approach, but not independently. That means we incorporate other proteins chemoinformation into the current learning procedure, namely which kind of compounds can interact with a specific proteins, by considering proteins similarities from different levels.

From the technical view, our method can reduce the model complexity and avoid over-fitting problems by sharing different proteins chemoinformation, which largely constrain the parameters’ degrees of freedom. Also our model can be used for proteins which have limited numbers of compounds targeting them.

From the chemical view, our method can put the structure scaffold diversity into consideration by both including compounds chemo-physics information, structures fingerprints information and other proteins chemoinformation.

We demonstrate our method through simulation results, comparison with SEA[[2]](#footnote-2), known drugs target prediction analysis and large-scale target prediction for compounds from traditional Chinese herbs.

## Method

***Model 1: Bayesian Hierarchical model***

Suppose we have M proteins, and for every protein g (1), compounds targets it. represents the potency of the i compound against the g protein. , a binary vector, represents the i compound’s 2D chemical structure fingerprint in protein g. The dimension of is D, and in this paper, we assume that all the have the same dimensions. Here we use a linear model to describe the relationship between the potency and chemical structure fingerprint given a compound in a specific protein, i.e.

is the average effect in protein g. is the parameter vector for different fingerprints in protein g and is the systematic random effect, which are independent standard normal distributions.

In which, follows Gamma distribution with hyper parameters .

And

Considering the fact that similar[[3]](#footnote-3) proteins should have similar values of , here we construct a hierarchical model to fit this phenomenon. Let represent the dth parameter () in different proteins. We assume that

In which, follows a non-informative prior distribution, i.e.,

can be treated as a kind of covariance matrix. In our model,

In which, is a similarity matrix among proteins defined in the nth level or way, and in total we have N levels or ways to describe the similarities between different proteins, such as sequence similarity, GO functional similarity and the distances in the protein-protein interactions networks and so on.

is the weight for . And

Here, we assume the prior distribution of follows Dirichlet distribution[[4]](#footnote-4),

Then we can derive the posterior distribution as follows, let ,

Then, we can derive the parameters conditional posterior distribution, which can be used in Gibbs sampling, by the Bayesian rule.

The conditional posterior distribution of follows Gamma distribution,

The conditional posterior distribution of follows *Inv*- distribution,/

The conditional posterior distribution of follows multivariable normal distribution,

Let

Since

Where

Also

Where

1. Email: zsp07@mails.tsinghua.edu.cn [↑](#footnote-ref-1)
2. One of the well-known ligand-based approaches. [↑](#footnote-ref-2)
3. We mention the “ proteins’ similarity ” a lot. The only similarity we care for here, is the similarity of pharmacological property. However it is hard for us to evaluate this similarity directly. But we know that proteins with similar pharmacological character share similar sequence similarities, functional similarities and so on. [↑](#footnote-ref-3)
4. In the current version, we don’t treat as random variables. [↑](#footnote-ref-4)