TSINGHUA UNIVERSITY

Predicting Compound-Protein Interactions by Multi-Task Learning

Songpeng Zu
January 28, 2015

It is a key problem to evaluate the compound-protein interactions (CPIs) 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. 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.

CONTENTS

1 Introduction

- New chemical scaffold

2 METHOD AND MATERIALS

2.1 MATERIALS

2.2 METHOD

Suppose we have m groups of data. In each group, we have the data $\mathscr{D}_j = \left\{\mathbf{X}_j, \mathbf{y}_j\right\}, j = 1, ..., m$, and $\mathbf{X}_j \in \mathbb{R}^{d \times n_j}$. Then we have

$$\mathbf{y}_j \sim \mathscr{N}\left(\mathbf{X}_j^T \omega_j, \sigma_y^2 \mathbf{I}\right) \tag{2.1}$$

Since different groups data may share similar or common structures, we assume the parameters of ω_j have the same mean on the prior distribution.

$$\omega_j \sim \mathscr{N}\left(\omega_*, \sigma_j^2 \mathbf{I}\right) \tag{2.2}$$

In which,

$$\omega_* \sim \mathscr{N}\left(\mu, \sigma_*^2 \mathbf{I}\right) \tag{2.3}$$

Suppose, for simplicity, that $\mu = \mathbf{0}, p(\sigma_y^2) \propto 1$, and that σ_j^2 and σ_* are fixed. Let $\Theta = \left\{\omega_j, j = 1, ..., m, \omega_*, \sigma_y^2\right\}$. We have

$$\begin{aligned}
\mathscr{L}_{hier}(\mathscr{D}; \Theta) &= \mathscr{L}_{orig}(\mathscr{D}|\Theta) + logp(\Theta) \\
&= \sum_j \left(logp(\mathscr{D}_j|\omega_j) - \frac{\| \omega_j - \omega_* \|^2}{2\sigma_j^2} \right) - \frac{\| \omega_* \|^2}{2\sigma_*^2} - \overbrace{\sum_j \frac{d}{2} log(2\pi\sigma_j^2) - \frac{d}{2} log(2\pi\sigma_*^2)}^{Const} \\
&= \sum_j \left(-\frac{\| \mathbf{y}_j - \mathbf{X}_j^T \omega_j \|^2}{2\sigma_y^2} - \frac{\| \omega_j - \omega_* \|^2}{2\sigma_j^2} \right) - \frac{\| \omega_* \|^2}{2\sigma_*^2} - \sum_j \frac{n_j}{2} log(2\pi\sigma_y^2) \\
&\quad - \overbrace{\sum_j \frac{d}{2} log(2\pi\sigma_j^2) - \frac{d}{2} log(2\pi\sigma_*^2)}^{Const}
\end{aligned} \tag{2.4}$$

Then we can get

$$\begin{aligned}
\frac{\partial \mathscr{L}_{hier}(\mathscr{D}; \Theta)}{\partial \omega_j} &= -\frac{1}{2\sigma_y^2} \frac{\| \mathbf{y}_j - \mathbf{X}_j^T \omega_j \|^2}{\partial \omega_j} - \frac{1}{2\sigma_j^2} \frac{\| \omega_j - \omega_* \|^2}{\partial \omega_j} \\
&= \frac{\mathbf{X}_j \mathbf{y}_j}{\sigma_y^2} + \frac{\omega_*}{\sigma_j^2} - \left(\frac{\mathbf{X}_j \mathbf{X}_j^T}{\sigma_y^2} + \frac{1}{\sigma_j^2} \mathbf{I} \right) \omega_j
\end{aligned} \tag{2.5}$$

$$\frac{\partial \mathscr{L}_{hier}(\mathscr{D}; \Theta)}{\partial \omega_*} = -\sum_j \frac{\omega_* - \omega_j}{\sigma_j^2} - \frac{\omega_*}{\sigma_*^2} \tag{2.6}$$

$$\frac{\partial \mathscr{L}_{hier}(\mathscr{D}; \Theta)}{\partial \sigma_y^2} = \frac{\sum_j \| \mathbf{y}_j - \mathbf{X}_j^T \omega_j \|^2}{2(\sigma_y^2)^2} - \frac{n}{2\sigma_y^2} \tag{2.7}$$

where n is the total number of samples in all the groups.

In this way, L-BFGS-B method is used to get the MAP estimations of these parameters following the function of $\mathscr{L}_{hier}(\mathscr{D};\omega)$ and their derivations [1].

Also we can derived the conditional posterior distribution of these parameters.

$$p(\omega_j|\omega_{[-j]},\sigma_y^2) \sim \mathscr{N}\left(\mu_j,\Sigma_j\right) \tag{2.8}$$

$$\mu_j = \Sigma_j\eta_j,\ \eta_j = \frac{\mathbf{X}_j\mathbf{y}_j}{\sigma_y^2} + \frac{\omega_*}{\sigma_j^2},\ \Sigma_j^{-1} = \frac{\mathbf{I}}{\sigma_j^2} + \frac{\mathbf{X}_j^T\mathbf{X}_j}{\sigma_y^2}$$

$$p(\omega_*|\omega_{[-*]},\sigma_y^2) \sim \mathscr{N}\left(\mu_*,\Sigma_*\right) \tag{2.9}$$

$$\mu_* = \Sigma_*\eta_*,\ \eta_* = \sum_j \frac{\omega_j}{\sigma_j^2},\ \Sigma_*^{-1} = \sum_j \frac{\mathbf{I}}{\sigma_j^2} + \frac{\mathbf{I}}{\sigma_*^2}$$

$$p(\sigma_y^2|\omega) \sim Inv-\chi^2(\nu,s^2) \tag{2.10}$$

$$\nu = n-2,\ s^2 = \frac{\sum_j \parallel \mathbf{y}_j - \mathbf{X}_j^T\omega_\mathbf{j} \parallel^2}{n-2}$$

Then we can also use Gibbs Samplers method to get the final results.

If m equals to 1, our model is reduced into the ridge regression model.

$$\omega_j \sim \mathscr{N}\left(\mathbf{0},(\sigma_j^2+\sigma_*^2)\mathbf{I}\right) \tag{2.11}$$

This can be derived based on the relationship

$$E(x) = E\left(E(x|y)\right),\ Var(x) = E\left(Var(x|y)\right) + Var\left(E(x|y)\right) \tag{2.12}$$

and the marginal distribution is normal distribution if the hyperprior of mean and the conditional distribution is normal.

[1] Note that $\sigma_y^2 > 0$, this is an bound-constrained optimization. L-BFGS-B is memory-limited ,especially used in bound-constrained, and suitable for quite large-dimensional optimization.

5 APPEND

Suppose there are M different groups of data, i.e.,

$$\left\{(y_i^{(g)}, \boldsymbol{x}_i^{(g)})\right\}_{i=1}^{n^{(g)}}; 1 \le g \le M$$

In our work, we have M proteins, and for each protein g ($1 \le g \le M$),there are n_g compounds target it. $y_i^{(g)}$ represents the potency of the compound i against the protein g [2]. $\mathbf{x}_i^{(g)}$, a binary vector which has D dimensions, represents the compound i's two-dimensional chemical structure fingerprints in the protein g. Here we use a linear model to describe the relationship between the potency and chemical structure fingerprints [3]. All the vectors, if no special explanations, are column vectors.

$$y_i^{(g)} = \mu^{(g)} + \mathbf{x}_i^{(g)^T} \cdot \boldsymbol{\beta}^{(g)} + \varepsilon \tag{5.1}$$

In which, $\mu^{(g)}$ is the average effect[4], $\boldsymbol{\beta}^{(g)}$ is the coefficients and ε is the systematic random error.

$$\varepsilon \sim N(0, \tau^{-1}) \tag{5.2}$$

$$\tau \sim Gamma(\kappa_1, \kappa_2) \tag{5.3}$$

And

$$\mu^{(g)}|\tau \sim N(0, \sigma_\mu^2 \tau^{-1}) \tag{5.4}$$

Considering the fact that proteins having similar pharmacological properties tend to interact with structure similar compounds, we construct a hierarchical model to fit this phenomenon.

Let

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1D} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{M1} & \beta_{M2} & \cdots & \beta_{MD} \end{pmatrix}$$

Then $\boldsymbol{\beta}^{(g)}$ corresponds to the gth row of $\boldsymbol{\beta}$ matrix and let $\boldsymbol{\beta}_{(d)}$ corresponds to the dth column of $\boldsymbol{\beta}$ matrix. We assume that

$$\boldsymbol{\beta}_{(d)} \sim MVN\left(\mathbf{0}, \sigma^2 \tau^{-1} \Sigma_\beta\right) \tag{5.5}$$

In which, σ^2 is a scale parameter, and follows a non-informative prior,

$$Pr(\sigma^2) \propto \frac{1}{\sigma^2} \tag{5.6}$$

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

$$\Sigma_\beta = \sum_{1 \le n \le N} \omega_n \Omega^{(n)}. \tag{5.7}$$

[2] The variance of potencies might be too large, should take log.

[3] Further complex function can be involved, like Gauss Process

[4] If $\boldsymbol{y}$ is centered, we can ignore the μ.

In which, Σ^n is a similarity Matrix between different proteins. Since we have several ways to define their similarities, such as sequence similarity, GO functional similarity, the distances in the protein-protein interaction networks and so on. So n corresponds to the nth definition of the protein similarity and in total we have N ways. ω_n is the weight for Σ^n.[5]

Then we can derive the posterior distribution as followed, let $\boldsymbol{\Theta} = \{\tau, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\mu}, \}$. Then,

$$
\begin{aligned}
Pr(\boldsymbol{\Theta}|\boldsymbol{y},\boldsymbol{X}) &\propto Pr(\tau)Pr(\sigma^2)Pr(\boldsymbol{\beta}|\tau,\sigma^2)Pr(\boldsymbol{\mu}|\tau)Pr(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\Theta}) \\
&\propto \tau^{\kappa_1-1}e^{-\tau\kappa_2}\frac{1}{\sigma^2}\prod_{d=1}^{D}\frac{1}{\sqrt{(2\pi)^M\left|\sigma^2\Sigma_\beta\tau^{-1}\right|}}exp\left\{-\frac{1}{2}\beta_{(d)}^T\left(\sigma^2\tau^{-1}\Sigma_\beta\right)^{-1}\beta_{(d)}\right\} \\
&\cdot\prod_{g=1}^{M}\frac{1}{\sqrt{2\pi\sigma_\mu^2\tau^{-1}}}exp\left\{-\frac{1}{2}\frac{\mu^{(g)^2}}{\sigma_\mu^2\tau^{-1}}\right\}\cdot\prod_{g=1}^{M}\prod_{i=1}^{n^{(g)}}\frac{1}{\sqrt{2\pi\tau^{-1}}}exp\left\{-\frac{1}{2}\frac{\left(y_i^{(g)}-\mu^{(g)}-\boldsymbol{x}_i^{(g)^T}\boldsymbol{\beta}^{(g)}\right)^2}{\tau^{-1}}\right\}
\end{aligned}
\tag{5.8}
$$

Then, we can derive the conditional posterior distribution for these parameters.

The conditional posterior distribution of $\boldsymbol{\beta}_{(d)}$ follows multivariable normal distribution,

$$
\begin{aligned}
Pr(\boldsymbol{\beta_{(d)}}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta}_{[-\boldsymbol{d}]},\tau,\sigma^2,\boldsymbol{\mu}) &\propto Pr(\boldsymbol{\beta_{(d)}}|\sigma^2,\tau)Pr(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\Theta}) \\
&\propto exp\left\{-\frac{1}{2}\boldsymbol{\beta}_{(d)}^T(\sigma^2\tau^{-1}\Sigma_\beta^{-1})\boldsymbol{\beta}_{(d)}\right\} \\
&\cdot exp\left\{-\frac{\tau}{2}\boldsymbol{\beta}_d^T\boldsymbol{\Psi}_d\boldsymbol{\beta}_d+\tau\boldsymbol{\varphi}^T\boldsymbol{\beta}_{(d)}+Const\right\}
\end{aligned}
\tag{5.9}
$$

The conditional posterior distribution of τ follows Gamma distribution,

$$
\begin{aligned}
Pr(\tau|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\sigma^2},\boldsymbol{\mu}) &\propto Pr(\tau)Pr(\boldsymbol{\beta}|\tau,\sigma^2)Pr((\mu)|\tau)Pr(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\Theta}) \\
&\propto \tau^{\kappa_1+\frac{D\cdot M}{2}+\frac{M}{2}+M\cdot n^{(g)}/2}exp\left\{-\tau\kappa_2-\tau\sum_{d=1}^{D}\frac{1}{2}\boldsymbol{\beta}_{(d)}^T(\sigma^2\Sigma_\beta)^{-1}\boldsymbol{\beta}_{(d)}-\tau\sum_{g=1}^{M}\frac{1}{2}\frac{\mu^{(g)^2}}{\sigma_\mu^2}\right. \\
&\left.-\tau\sum_{g=1}^{M}\sum_{i=1}^{n^{(g)}}\frac{1}{2}\left(y_i^{(g)}-\mu^{(g)}-\boldsymbol{x}_i^{(g)^T}\boldsymbol{\beta}^{(g)}\right)^2\right\}
\end{aligned}
\tag{5.10}
$$

[5] We can also maximum similarity or the correlation among them.