Exploring Latent Mixed Membership Models for Predicting Membership of Biological Processes of Proteins Based on Protein-Protein Interaction (PPI) Data

Xiao (Katrina) Liu
Carnegie Mellon Univerisity

Introduction

Protein-protein interactions (PPI) have been a crucial part of the research on biological pathways and processes and demonstrate the physical basis of the reactions involved in compound constructions. It provides plenty of information not only on intra-organism activities but also on the proteins themselves. One assumption is that PPIs can be used to predict the functionalities of proteins. Throughout the last two decades, both experimental approaches and theoretical models on PPIs have been studied widely.

In the context of protein function prediction, mixed-membership models have been frequently used as it is common where one protein can have multiple distinct functions in different biological processes. Airoldi et al. proposed a Bayesian latent mixed-membership model on predicting proteins and their associated functions based on the binary adjacency interaction matrix [2]. They tested their model on a fraction of yeast protein-protein interaction data collected from the MIPS database[4] and achieved an accuracy of about 47%. Although their method outperformed naive spectral clustering, there is space for improvement in the prediction accuracy. However, as more information has been collected since, we wonder how the performance of the model on predicting protein functions from PPIs in other scopes would be like and if there can be an improvement in the accuracy in predicting protein functions.

Data Source

We retrieved the complete dataset of intra-species protein-protein interactions of Homosapiens that are recorded by the Agile Protein Interactomes Dataserver(APID) platform[3]. We use the interaction data to create adjacency/interaction matrix for conducting mixed membership clustering. We then used the GO API to retrieve all of the biological process terms that each gene is involved in. From the data, we collect the membership groups of the genes to establish the model.

Statistical Model

To explain our algorithm, we will first introduce the statistical model for interaction matrix for membership prediction of biological process[2]. The input of this model includes the number of proteins involved in the interactions N, the number of membership groups K, and the actual observed N x N interaction /adjacency matrix R, where R_{ij} indicates whether there is an observed interaction between protein i and protein j.

For statistical model, the observed interaction matrix is being modeled as a collection of Bernoulli random variables, where for each pair of proteins, the presence/absence of an interaction is drawn by two features: (1) choosing a latent class for each protein from a protein specific distribution(initiator); (2) drawing from a Bernoulli distribution with parameter associated with the pair of latent classes(receiver).

Therefore, the generative process with latent Dirichlet allocation can be described as given a K-dimensional Dirichlet parameter α, and a K x K matrix of Bernoulli parameters η, we will be able to generate a NxN adjacency matrix by sampling N samples θ_i from Dirichlet distribution with parameter α, sampling N x N $z_{i,j,1}$, $z_{i,j,2}$ feature random variable by Multinomial distribution with θ_i, θ_j, and lastly sampling the entries of R by Bernoulli distribution using η and z.

Algorithm[2]

The marginal probability based on the model is represented as:

$$p(R|\alpha,\eta) = \int_\theta \int_z = \prod_{i=1}^{N} p(\theta_i|\alpha) \prod_{j=i}^{N} p(z_{i,j,1}|\theta_i) p(z_{i,j,2}|\theta_j) p(R_{ij}|z_{i,j},\eta) dz d\theta$$

However, the above objective is not tractable to calculate, but we can obtain a lower bound of log-likelihood of the objective, where

$$\log p(R|\alpha,\eta) \geq \mathbb{L}(\eta,\phi;\alpha,\eta)$$

$$= \sum_{i=1}^{N} \mathbb{E}[\log p(\theta_i|\alpha_i)] + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[\log p(z_{i,j,1}|\theta_i) + \log p(z_{i,j,2}|\theta_j + \log(p(R_{ij}|z_{i,j},\eta))]$$

$$-\mathbb{E}[\log q(\theta, z)].$$

$$q(\theta, z|\gamma,\phi) = \prod_{i=1}^{N} Dirichlet(\theta_i|\gamma_i) \prod_{j=1}^{N} Mult(z_{i,j,1}|\phi_{i,j,1}) Mult(z_{i,j,2}|\phi_{i,j,2})$$

We implemented an EM-like algorithm to iteratively conduct variational inference and parameter estimation.

- Initialization: We initialize $\gamma_i^0 = Dirichlet([\frac{2N}{K}] * K)$ for all $[i]_0^N$, $\phi_{i,j,v} = Mult([\frac{1}{K}] * K)$ for all $[i]_0^N, [j]_0^N, v = 1, 2$.
- Variational Inference: We obey the following rules to update γ, ϕ_1, ϕ_2, with digamma function Ψ:

$$\gamma_{i,g}^* = \alpha_g + \sum_{j=1}^{N} \phi_{i,j,1,g} + \sum_{j=1}^{N} \phi_{i,j,2,g}$$

$$\phi_{i,j,1,g}^* \propto e^{\Psi(\gamma_{i,g}) - \Psi(\sum_{g=1}^{K} \gamma_{i,g})} \prod_{h=1}^{K} \eta_{g,h}^{R_{ij}\phi_{i,j,2,h}} \prod_{h=1}^{K} (1-\eta_{g,h})^{(1-R_{ij})\phi_{i,j,2,h}}$$

$$\phi_{i,j,2,h}^* \propto e^{\Psi(\gamma_{i,g}) - \Psi(\sum_{g=1}^{K} \gamma_{i,g})} \prod_{g=1}^{K} \eta_{g,h}^{R_{ij}\phi_{i,j,1,g}} \prod_{g=1}^{K} (1-\eta_{g,h})^{(1-R_{ij})\phi_{i,j,1,g}}$$

  and scale ϕ_1, ϕ_2 to the extent where the sums of elements of each are equal to 1.
- Parameter Estimation: We obey the following rules to update η:

$$\eta_{g,h}^* = \frac{\sum_{i=1}^{N}\sum_{i=1}^{N} \phi_{i,j,1,g}\phi_{i,j,2,h} R_{ij}}{\sum_{i=1}^{N}\sum_{i=1}^{N} \phi_{i,j,1,g}\phi_{i,j,2,h}}$$

We iterate the variational inference step and parameter estimation step until the values for γ, ϕ_1, ϕ_2 to converge.

Experiment Data Selection

We have experimented datasets with different count threshold values to select a subset of GO terms of biological processes. Here we present the information on the dataset with count threshold of 800:

GO Term ID	Biological Process Name
GO:0006357	Regulation of transcription by RNA polymerase II
GO:0006355	Regulation of transcription, DNA-templated
GO:0045944	Positive regulation of transcription by RNA polymerase II
GO:0015031	Protein transport
GO:0007165	Signal transduction
GO:0030154	Cell differentiation

Table 1. The 6 selected GO biological process terms that most frequently appeared in the complete dataset.

Evaluating Performances

We first introduce the representation of the membership matrix, which is of dimension N x K. Each row represents a gene and each column represent a group. We constructed the reference membership matrix by

$$ref_{i,j} = \mathbb{I}[\text{gene i is involved in process j}].$$

To evaluate the accuracy of the predicted membership model, we scale the γ values along each column since each column represents a group by the factor of the inverse of the maximum value of each column. Notice that we still need to match up the clusters from predicted results to the actual process. We resolve this concern by computing the minimum sum of the absolute difference between the reference membership and every possible permutation of columns of the scaled predicted membership. The scaling is represented by

$$\gamma'_{i,j} = \frac{\gamma_{i,j}}{\max_{l=1}^{N} \gamma_{l,j}}.$$

The difference score is then computed by

$$d(ref, \gamma') = \min_{p \in perm([1,\ldots,K])} \frac{\sum_{i=0}^{N}\sum_{j=0}^{K} |ref_{i,j} - \gamma'_{i,p_j}|}{NK}$$

The closer the difference score is to 0, the more similar the predicted membership is to the reference membership.

Result and Conclusion

We used the naive mixed membership stochastic block model [1] implementation as a baseline method to compare with the results of latent mixed membership algorithm. For the latent mixed membership algorithm, we have a difference score of 0.35 and for the baseline method we have a difference score of 0.72. We can see a significant improvement on the prediction accuracy. One possible reason for such improvement can be the fact that the latent model included two latent variables z_1 and z_2 to account for the hidden features in the process of clustering. For future work, we can explore the performances of the two models with larger datasets with a larger number of genes, interactions, and/or function groups.

References

[1] Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing.
Mixed membership stochastic blockmodels.
Advances in neural information processing systems, 21, 2008.

[2] Edoardo Airoldi, David Blei, Eric Xing, and Stephen Fienberg.
A latent mixed membership model for relational data.
In *Proceedings of the 3rd international workshop on Link discovery*, pages 82–89, 2005.

[3] Diego Alonso-Lopez, Miguel A Gutiérrez, Katia P Lopes, Carlos Prieto, Rodrigo Santamaría, and Javier De Las Rivas.
Apid interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks.
Nucleic acids research, 44(W1):W529–W535, 2016.

[4] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, et al.
The mips mammalian protein–protein interaction database.
Bioinformatics, 21(6):832–834, 2005.

[5] Hanhuai Shan and Arindam Banerjee.
Mixed-membership naive bayes models.
Data Mining and Knowledge Discovery, 23(1):1–62, 2011.

xiaol3@andrew.cmu.edu ntiyal@andrew.cmu.edu