Predicting Biological Process Membership of Proteins from Protein-Protein Interaction Using Latent Mixed Membership Model

Xiao (Katrina) Liu
Carnegie Mellon University
xiaol3@andrew.cmu.edu

November 2, 2022

Abstract

The latent mixed membership (LMM) model has been used to predict protein functions based on the protein-protein interaction (PPI) data. Previously, the latent mixed membership model was used to predict protein functions on the MIPS dataset. However, the overall success rate to recover the protein functions is only around 40% to 45%. Intuitively, we believe that this model might be more accurate in predicting the biological process membership of proteins from the PPI data. We implement an Estimation-Maximization (EM) algorithm to approximate the model to calculate the degree of membership of each protein in each biological process group. Our result proves a noticeable improvement in prediction accuracy of latent mixed membership model to mixed membership stochastic blockmodel.

1 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. Throughout the last two decades, both experimental approaches and theoretical models on PPIs have been studied widely. It provides plenty of information on not only the protein themselves but also intra-organism activities. It is a common assumption that PPIs to some extent resemble the memberships of biological processes of the involved proteins.

The relationship between protein and their corresponding features can resemble a mix-membership relation: where each protein can possess multiple features. In other words, protein elements can belong to multiple feature groups. Information of such membership relation can be represented in the protein-protein interaction data as proteins possessing the same feature will be more likely to interact with each other. In this case, features can be referring to biological functions or membership/involvement of biological processes.

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 for predicting proteins and their associated functions based on the binary adjacency interaction matrix [1]. They tested their model on a fraction of yeast protein-protein interaction data collected from the MIPS database[6] 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 what the performance of the model on predicting protein membership in biological processes from PPIs in other scopes would be like and if there can be an improvement in the prediction accuracy than that in predicting protein functions.

2 Statistical Model

We will briefly introduce the latent mixed-membership statistical model[1][1] behind our implementation and the involved notations that will also be used in section 3.

Latent Dirichlet Allocation (LDA) is a common model in topic discovery applications[5]. It has been widely used in categorizing documents based on the content in natural language processing. The latent mixed-membership model will share a similar generative process with LDA, with a few changes in restricting the latent variables.

Suppose we have N protein element and K biological processes to be identified. Denote the binary interaction matrix as R with dimension $N \times N$ where each entry $R_{i,j}$ indicates if protein i interacts with protein j. The model attempts to account for different scenarios of an interaction between two proteins. The generative process is described in Algorithm 1. The generative process samples the interactions from a Bernoulli distribution for each protein based on each pair of latent classes. Let α of dimension K represent the Dirichlet parameter of the model and η of dimension $K \times K$ represents the Bernoulli parameters between each pair of groups. θ's are sampled only once throughout the generative process as Dirichlet random variables that indicate each proteins degree of membership into each feature group; $z_{i,j,1}$ and $z_{i,j,2}$'s are sampled for each interaction pair of proteins as indicators into the feature groups.

Algorithm 1 Generative Process of LMM[1]

for $i = 1...N$ **do**
  Sample $\theta_i \sim Dirichlet(\alpha)$
end for
for $i, j \in N \times N$ **do**
  Sample $z_{i,j,1} \sim Multinomial(\theta_i, 1)$
  Sample $z_{i,j,2} \sim Multinomial(\theta_j, 1)$
  Sample $R_{i,j} \sim Bernoulli(\eta_{z_{i,j,1}, z_{i,j,2}})$
end for

[1]This section is largely adopted from [1] with minor modifications and additional clarifications.

This model lead to a probability distribution of

$$p(R, z, \theta|\alpha, \eta) = \prod_{i=1}^{N} p(\theta_i|\alpha) \times$$

$$\times \prod_{j=1}^{N} p(z_{i,j,1}|\theta_i) p(z_{i,j,2}|\theta_j) p(R_{i,j}|\eta, z_{i,j})$$

The derived marginal probability is

$$p(R|\alpha, \eta) = \int_Z \int_\Theta p(R, z, \theta|\alpha, \eta) d\theta dz$$

However, the above marginal probability is not tractable to compute, so we will need to carry out a process of variational inference and parameter estimation.

The variational method will posit a variational distribution on the latent variable $q(\theta, z)$. By Jensen's inequality, we can acquire a lower bound for the marginal log-likelihood:

$$\begin{aligned}\log p(R|\alpha, \eta) \geq & \sum_{i=1}^{N} \mathbb{E}[\log p(\theta_i|\alpha)] + \\ & + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[\log p(z_{i,j,1}|\theta_i)] + \\ & + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[\log p(z_{i,j,2}|\theta_j)] + \\ & + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}[\log p(R_{i,j}|z_{i,j}, \eta)] - \\ & - \mathbb{E}[\log q(\theta, z)]\end{aligned}$$

Here, the variational distribution is defined to be

$$q(\theta, z|\gamma, \phi) = \prod_{i=1}^{N} Dirichlet(\theta_i|\gamma_i) \times$$

$$\times \prod_{j=1}^{N} Mult(z_{i,j,1}|\phi_{i,j,1}) Mult(z_{i,j,2}|\phi_{i,j,2}),$$

from which we can derive the derivative for updating variational parameters in section 3.1.

3 Methodology

3.1 Algorithm

The overall procedure is designed to be an Estimation-Maximization (EM) algorithm with coordinate ascend. We start with initialization:

$$\alpha_g = \frac{1}{K} \forall g = 1...K$$

$$\gamma_i = Dirichlet(\alpha) \forall i = 1...N$$

$$\phi_{i,j,1} = Mult(\alpha), \phi_{i,j,2} = Mult(\alpha) \forall i, j = 1...N.$$

The parameter estimation step is updating Bernoulli parameter matrix η as described in Algorithm 2 and the variational inference step is updating variational parameters γ and ϕ as described in Algorithm 3. Here, Ψ refers to the digamma function where $\Psi(x) = \ln x - \frac{1}{2x}$. The update is conducted according to the method described in [7]. The two steps iteratively take turns to be conducted until the Dirichlet random variable γ and ϕ converges so that the lower bound is optimized. However, because of the limitation of computation, instead of waiting for perfect convergence, we set a maximum number of iteration of 500 and also a error threshold of $1.0e-5$ to allow minimal differences for each value of the parameters.

Algorithm 2 Parameter Estimation[1]

for $g = 1...K$, $h = 1...K$ **do**

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

end for

3.2 Experiment Data

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 an adjacency/interaction matrix for conducting mixed membership clustering. We

Algorithm 3 Variational Inference[1]

for $i = 1...N$, $j = 1...N$, $g = 1...K$ **do**

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

$$\times \prod_{h=1}^{K}(1-\eta_{g,h})^{(1-R_{i,j})\phi_{i,j,2,h}}$$

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

$$\times \prod_{g=1}^{K}(1-\eta_{g,h})^{(1-R_{i,j})\phi_{i,j,1,g}}$$

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

end for

then used the GO API to retrieve all of the biological process terms in that each gene is involved [4]. From the data, we collect the membership groups of the genes to establish the model.

We have experimented with datasets with different threshold values of the number of occurrences for each biological process to select a subset of GO terms of biological processes. Here we present the information on the dataset with a threshold value of 800 and select the six most prominent biological process groups in Table 1. The six processes produce a subset of 5370 genes and 23017 interactions to experiment with. Because of the limitation of computation, we selectively choose a smaller subset of 100 genes and 200 genes to test the accuracy of the algorithm.

3.3 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

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

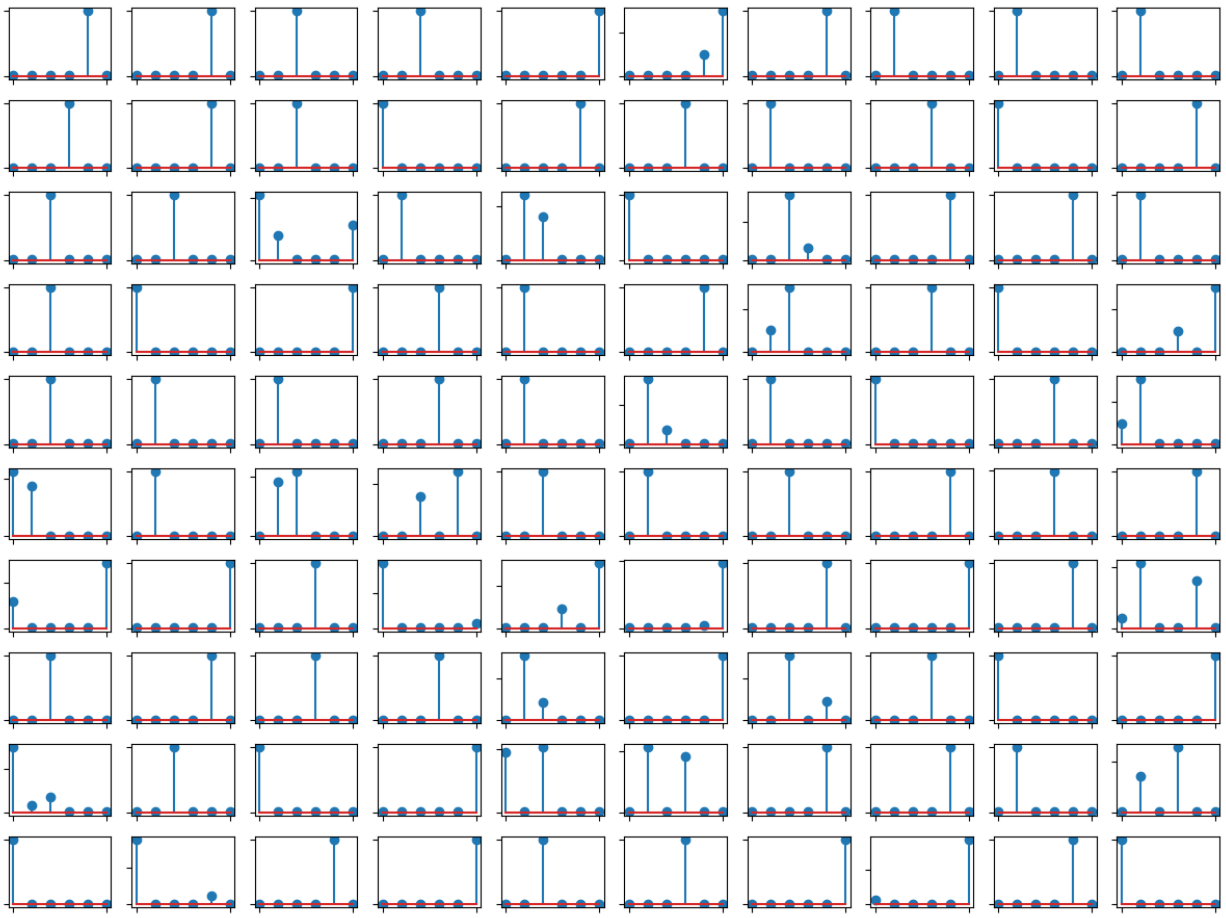

Figure 1: Predicted degrees of memberships for the 100 genes with 100 iterations.

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

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

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

4 Results

A demonstration of predicted degrees of membership is included in Figure 1 from the experiment with 100

GO Term ID	Count	Biological Process Name
GO:0006357	1633	Regulation of transcription by RNA polymerase II
GO:0006355	1518	Regulation of transcription, DNA-templated
GO:0045944	941	Positive regulation of transcription by RNA polymerase II
GO:0015031	810	Protein transport
GO:0007165	1475	Signal transduction
GO:0030154	895	Cell differentiation

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

genes and 100 as the max iteration threshold. The x-axis refers to biological processes in the order they appeared in Table 1.

We used the naive mixed membership stochastic block model [2] implementation as a baseline method to compare with the results of the latent mixed membership algorithm.

When testing with 100 genes and 100 as the threshold of max iteration of convergence, 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 in prediction accuracy. In Figure 2, we plot the difference scores of each biological process group for the optimal alignment of clusters.

When testing with 200 genes and 500 as a threshold of max iteration of convergence, we have a difference score of 0.29 for LMM and a difference score of 0.74 for MMSB. Corresponding results for each biological processes are plotted in Figure 3.

We can observe that the difference scores for both types exhibit a relatively even distribution and there was no dominant group with a significantly higher difference score than the others.

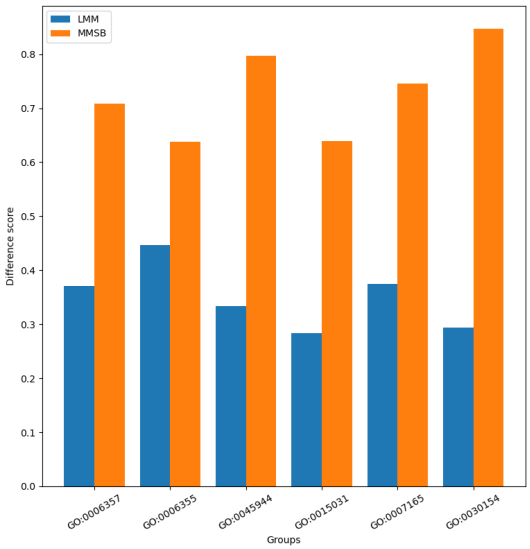

Figure 2: Difference scores of 100 genes with 100 iterations for the each selected biological process after aligning with the optimal permutation.

5 Discussion and Conclusion

The two experiments we conducted are to predict the membership of 100 randomly selected genes with max iterations of EM algorithm of 100 and 200 randomly selected genes with max iterations of EM algorithm of 500. Our assumption is that a higher number of genes will increase the difference score and therefore have a lower prediction accuracy. Also, with a higher value for the threshold of iterations, the degrees of memberships should be lower. However, due to our limitation of computation, we could only complete two experiments and the results show that the latent mixed membership model is more accurate with more iterations. The difference scores of each biological process are evenly distributed for both LMM and MMSB, with the differences scores of LMM around 20% 30% and those of MMSB around 70% 80%.

One possible reason for such difference can be the fact that the latent model included two latent vari-

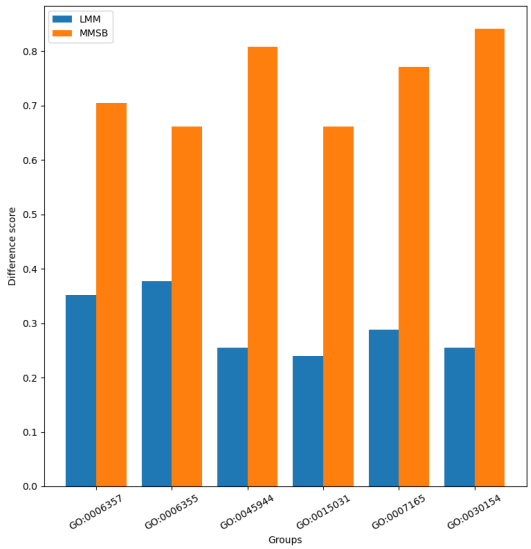

Figure 3: Difference scores of 200 genes with 500 iterations for the each selected biological process after aligning with the optimal permutation.

ables z_1 and z_2 to account for the hidden features in the process of clustering and it focuses on the effect of binary interaction matrix and weakens the effect of multi-value of the input matrix.

In conclusion, we see a significant improvement in LMM, and the difference scores it demonstrates are much better than naive random guesses, which proves that the model indeed reflects the statistical relation between the protein-protein interaction data and the protein's membership in biological processes.

6 Future Work

Because of our limitation of computation, we can only conduct an experiment on smaller subsets of the entire homo sapiens intra-species interactions, and the biological processes involved were not carefully chosen.

For future work, we can explore the performances of the more clustering models with larger datasets potentially with a larger number of genes, interactions, and/or biological process groups. Other potential extensions of our research project could be specifically targeting different types of tissues and experimenting with intra-tissue type protein-protein interactions.

References

[1] 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. 1, 2, 3

[2] Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008. 5

[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. 3

[4] David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'Donovan, and Rolf Apweiler. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, 25(22):3045–3046, 09 2009. 3

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. 2

[6] 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. 1

[7] Martin Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 01 2008. 3