A Novel Stochastic Block Model for Network-Based Prediction of Protein-Protein Interactions

Xiaojuan Wang[1], Pengwei Hu[2], and Lun Hu[1,3](✉)

[1] School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
[2] IBM Research, Beijing, China
[3] Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi, China
hulun@ms.xjb.ac.cn

Abstract. Proteins interact with each other to play critical roles in many biological processes in cells. Although promising, laboratory experiments usually suffer from the disadvantages of being time-consuming and labor-intensive. Hence, a variety of computational approaches have been proposed to predict protein-protein interactions from an alternative view. However, most of them heavily rest on the biological information of proteins while ignoring the latent structural features in protein interaction networks. In this paper, we propose a novel stochastic block model for network-based prediction of protein interactions. By simulating the generative process of a protein interaction network, our approach can capture the latent structural features of proteins from the perspective of forming protein complexes, thus verifying whether two proteins interact with each other or not. To evaluate the performance of the proposed prediction approach, a series of extensive experiments have been conducted and we have also compared our approach with state-of-the-art network-based prediction model. The experiment results show that our approach has a promising performance when applied to predict protein-protein interactions.

Keywords: Protein interaction network · Protein interaction prediction · Stochastic block model

1 Introduction

Proteins are important component of cells, they provide the material basis for life and undertake various functions in living organisms. Instead of functioning alone, protein interact with each other to form complexes, thus performing their functions. Since protein-protein interactions (PPIs) constitute most of biological processes, such as transcription, replication and translation of DNA, it is of great significance to understand the mechanisms of PPIs. Furthermore, knowing how proteins interact with other not only allows us to get more familiar with cellular mechanisms but also facilitates the design of novel drugs.

D.-S. Huang and K.-H. Jo (Eds.): ICIC 2020, LNCS 12464, pp. 621–632, 2020.
https://doi.org/10.1007/978-3-030-60802-6_54

A lot of efforts have been made to the prediction of PPIs. Several laboratory-based approaches have been developed at early stage, such as two-hybrid [13], TAP-tagging [8, 19], protein chips [25], synthetic lethal analysis [26] and correlated mRNA expression profile [5]. Although promising, laboratory-based approaches are both time-consuming and labor-intensive. Hence, to overcome these disadvantages, a variety of computational approaches have been proposed recently and they are mainly classified into three categories including sequence-based approaches [9–12], evolutionary-based approaches [21] and network-based approaches [1, 17]. Unlike the first two categories that heavily rest on biological information of proteins, network-based approaches are preferred as they only perform their tasks based on protein interaction network data.

As a recent attempt in network-based approaches, L3 [17] follows the intuition that proteins interact not if they are similar to each other, but if one of them is similar to the other's partners. In contrast to the traditional triadic closure principle [16] based on the similarity of neighborhoods [2, 7, 23], L3 relies on network paths of length three and introduces a degree-normalized score to measure the likelihood of being interacting for two proteins. Although the experiment results show that L3 significantly outperforms all existing link prediction methods when applied to predict PPIs, we believe that its performance could be constrained by the assumption of network paths with length three. To address this concern, we intend to propose a more flexible network-based prediction model.

In fact, proteins in the same protein complex are intensively interacted with each other, thus forming a denser structure than those from different complexes. Motivated by this observation, a novel stochastic block model, modified from mixed membership stochastic block model (MMSB), is proposed for network-based prediction of protein interactions. As a well-known community detection model, MMSB analyzes latent structural features by taking into account the memberships of entities. It captures different communities in the network and allows each entity has its own community distribution, which indicates that each node in the network has a K dimensional mixed membership to represent its weight for each community. The introduction of K dimensional community membership allows each node to be assigned to multiple communities instead of only one community, it is reasonable for protein network, because the same protein can participate multiple biological processes under different conditions in real biological system. In generative process, MMSB uses a binary adjacency matrix to represent interaction data and the mixed membership for each node follows Dirichlet distribution. Based on the community membership, there are two interaction indicators for each protein pair to indicate the community they belong to which follows Multinomial distribution. Combining these indicators with a $K \times K$ matrix of Bernoulli parameters which represent the probability of connection between K communities, the pairwise proteins can be classified as interacting or non-interacting.

Considering the huge data of PPIs, we follow the scheme of assortative mixed membership blockmodel (aMMSB) [6] to solve the inference problem of the proposed stochastic block model. Moreover, when determining the complex for each of proteins, we introduce a weighted similarity measure based on the strengths of complexes as well as the distance between the memberships of two proteins. In order to evaluate the performance of our model, we have applied it to three independent PPI datasets and compared its performance with L3. The experiment results show that our prediction approach has a promising performance in terms of several different metrics.

The rest of the paper is organized as follows. The details of our approach are presented in Sect. 2, following which we compare the proposed approach with L3 and also discuss the experiment results in Sect. 3. Finally, the paper ends with a conclusion in Sect. 4.

2 Methods

In aMMSB, protein networks are as input to the model. They are divided into K communities to identify different protein groups and proteins in each group have their mixed memberships, interaction between two proteins are identified by the two elements. To improve the performance of aMMSB, we changed the criterion of original aMMSB and proposed our approach. In this section, we introduce the generative process of aMMSB and the details of our approach.

2.1 The Generative Process of PPI Network

Though aMMSB can capture latent communities and assign each node a mixed membership to describe its weight for each community, the performance of original aMMSB is not good. The interaction indicator z is a K dimensional vector where only one element equals to one, the index of which indicates its corresponding community, and all the other elements equal to zero. In view of the fact that multiple proteins can interact with others in different protein complexes under different conditions, it means one protein may belong to several groups. As a criterion of interacting probability between two proteins, the indicate vectors with only one element equals to one is too strict, a number of interacting protein pairs are predicted to not interact and it leads to many false negative results thus. Due to this severe criterion of original aMMSB, we proposed an improved approach to relax standard and promote performance.

To weaken the impact of harsh criterion, we quantify the difference between two community memberships by Euclidean distance and combine it with community strength to calculate interacting probability. In training set, a protein network is put into aMMSB as input, community strength and community memberships are obtained as evaluation criterion of interacting by the model. The probability that a protein pair can interact is determined by the two elements with normalized function.

Suppose we have observed the links between proteins, each protein can be seen as a node and the link between them can be seen as edge represented by y. $N \times N$ binary matrix is completed by N proteins where y_{ij} is equal to 1 if there is a link between protein i and protein j, y_{ij} is equal to 0 otherwise. For each protein, there is a community membership π to describe its degree of membership for each community which obeys Dirichlet distribution. For the whole network, there is community strength β_k to capture latent communities, β_k is a K dimensional vector and ranges from 0 to 1, the elements in it represent the tightness of each community. Specific process is as follows:

- For each protein i, sample community membership $\pi_i \sim Dirichelet(\alpha)$. The specific equation is: $Dir(\pi_i|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_{i,k}^{\alpha_k - 1}$

- For each community k, sample community strength $\beta_k \sim Beat(\eta)$. The specific equation is: $Beta(\beta_k|\eta_0,\eta_1) = \frac{\Gamma(\eta_0+\eta_1)}{\Gamma(\eta_0)\Gamma(\eta_1)}\beta_k^{\eta_0-1}(1-\beta_k)^{\eta_1-1}$
- For each pair of node i and node j:

  Sample link $y_{ij} = \frac{\gamma-\gamma_{min}}{\gamma_{max}-\gamma_{min}}$, where $\gamma = \sqrt{\sum_{k=1}^{K}\beta_k(\pi_{i,k}-\pi_{j,k})^2}$

2.2 Stochastic Variable Inference

Our approach is a subclass of MMSB model with stochastic variable inference, it is a statistical model that allows nodes to participant in multiple communities. Figure 1 shows corresponding joint distribution of variables applied. The goal of the model is to compute the posterior distribution $p(\pi_{1:N},\beta_{1:K},z|y,\alpha,\eta)$ and the strategy adopted is variable inference (VI) [14]. However, traditional variable inference in MMSB deals with all the node pairs each iteration which requires a lot of time. Therefore, stochastic variable inference (SVI) is applied in aMMSB to save time and improve efficiency, aMMSB with SVI can also get comparable results to MMSB with VI.

SVI is a coordinate ascent algorithm that iteratively updates local and global parameters. For each iteration, we first subsample the network and get a subset S, local parameters are optimized given current global parameters and we then update global parameters using stochastic natural gradient with subset S and local parameters. The first step is called local step (L-step) and the second step is called global step (G-step). The specific generative process is described in Algorithm 1.

Algorithm 1 stochastic aMMSB

Initialize variables randomly: $\gamma = (\gamma_n)_{n=1}^{N}, \lambda = (\lambda_k)_{k=1}^{K}$
while not converging **do**
  Sample a subset S of node pairs
  L-step: $\forall(i,j) \in S$, optimize$(\phi_{i\to j}, \phi_{i\leftarrow j})$
  Compute natural gradients: $\forall n, \partial\gamma_n^t$; $\forall k, \partial\lambda_k^t$
  G-step: update γ and λ, $\gamma \leftarrow \gamma + \rho_t\,\partial\gamma^t$; $\lambda \leftarrow \lambda + \rho_t\,\partial\lambda^t$
  Set $\rho_t = (\tau_0 + t)^{-\kappa}$; $t \leftarrow t+1$
end while

The Global Step

The global step updates global parameters community strengths λ and community memberships γ. For a network with N nodes, there is $M = N(N-1)/2$ node pairs, we extract a node pair (i,j) at random. In t-th iteration, the stochastic natural gradients of global parameters are:

$$\partial\gamma_{i,k}^t = \alpha_k + \frac{1}{g(i,j)}\phi_{i\to j,k}^t - \gamma_{i,k}^{t-1}$$

$$\partial\lambda_{k,m}^t - \eta_{k,m} + \frac{1}{g(i,j)}\phi_{i\to j,k}\cdot\phi_{i\leftarrow j,k}\cdot y_{ij,m} - \lambda_{k,m}^{t-1}$$

where $y_{ij,0} = y_{ij}$ and $y_{ij,1} = 1 - y_{ij}$, we require that $\sum_t \rho_t^2 < \infty$ and $\sum_t \rho_t = \infty$, we set $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$, where κ is learning rate and $\kappa \in (0.5, 1]$, τ_0 downweights early iterations and $\tau_0 \geq 0$. The time for a G-step is O(NK) and the memory required is O(NK).

The Local Step

The local step updates local parameters ϕ which represents the posterior approximation of which communities are important in determining whether there is a link. The time for L-step is O(nK) where n is the number of node pairs sampled in each iteration.

$$\phi^t_{i\to j,k}|y = 0 \propto \exp\left\{E_q\left[log\pi_{i,k}\right] + \phi^{t-1}_{i\to j,k} E_q[\log(1-\beta_k)]\right\}$$

$$\phi^t_{i\to j,k}|y = 1 \propto \exp\left\{E_q\left[log\pi_{i,k}\right] + \phi^{t-1}_{i\to j,k} E_q[\log\beta_k] + \left(1 - \phi^{t-1}_{i\to j,k}\right) log \in\right\}$$

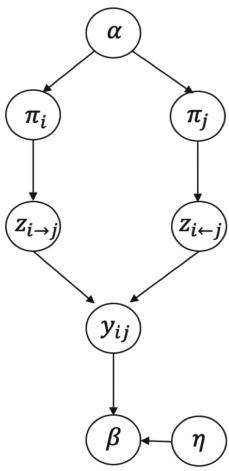

Fig. 1. Graphical model of our approach

3 Results

To evaluate the performance of aMMSB and its improved version experimentally, three protein networks including *yeast* [27], *krogan* [18] and *human* [17, 22] are selected where the human dataset is composed of three human networks containing HI-II-14 [22], HI-III [22] and HI-tested [17]. We adopt 5-fold cross-validation to get more comprehensive results. Except computational experiments in our model, we also applied these datasets to L3 model and compared their results. It shows that aMMSB with improved version acquires remarkable effect and outperforms L3 in numerous evaluation measures.

3.1 Data Structure

Considering the diversity of structure in datasets, the experimental results may be influenced by different protein networks. The datasets in different species are used to get more persuasive results and the detailed analysis of three datasets applied is shown in Table 1.

Table 1. Analysis of three datasets and their main characteristics

Dataset	N	E	k_{av}	ρ	C	SIPs
Yeast	964	3846	7.979	0.008	0.148	0
Krogan	2708	7123	5.261	0.002	0.188	0
Human	6657	32307	9.521	0.001	0.069	616

3.2 Computational Experiments

In experiments, all the datasets are divided into training set and test set, training set is as input network and test set is used to access predictive power of the model, 5-fold cross-validation is used to be more convincing. As for negative dataset, we keep consistent with the strategy used in L3 to facilitate comparison. 244 non-interacting protein pairs are selected including 100 pairs where at least one of the proteins are in the top 500 L3 predictions.

Sensitivity of *K* Value

To explore the relationship between number of communities and performance of aMMSB, we take the method of traversal to get best K value. Evaluation measures included are F1-score, AUC and PR-AUC, the results are shown in Fig. 2. Form the figure, the best K value in *yeast* dataset is 9, best K value in *krogan* dataset is 10 and best K value in *human* dataset is 9 too.

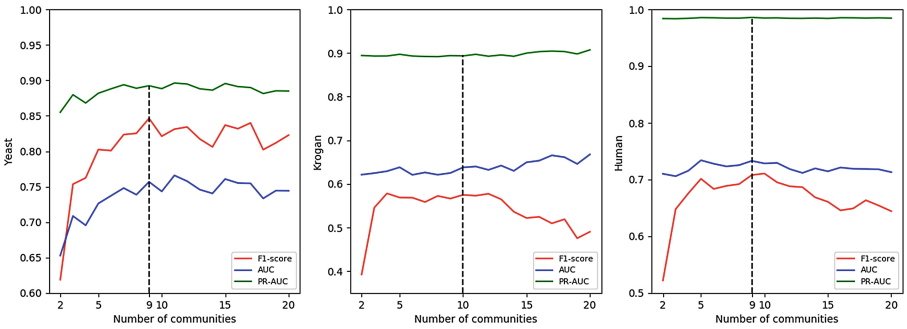

Fig. 2. The performance of different K values

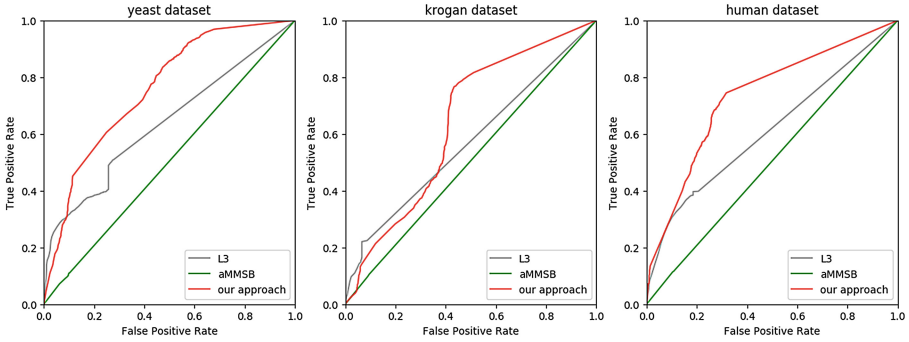

Fig. 3. The performance of AUC

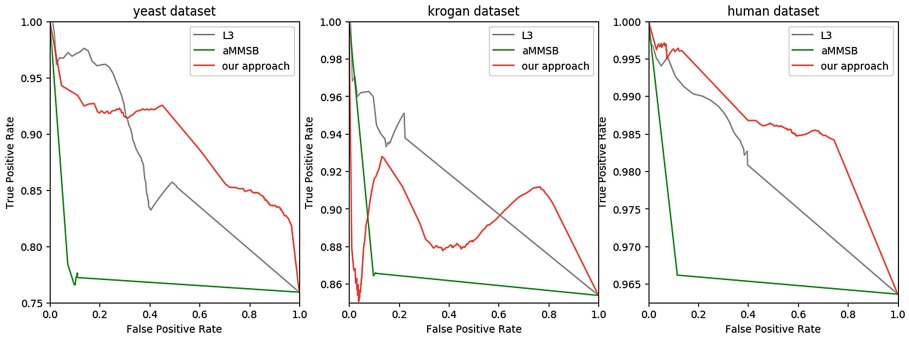

Fig. 4. The performance of PR-AUC

5-Fold Cross-Validation

In 5-fold cross-validation, we split the dataset into five subsets, each of them is as test set and the other four subsets remained are as training set, the test set and training set are rotated five times to access performance. Precision, Recall, F1-score, AUC, PR-AUC, AUC(L3) are used as evaluating measures where AUC(L3) is the measure used in L3 to calculate AUC but there is a little different between it and general AUC calculating method. AUC(L3) follows Eq. 4 where by randomly chooses n pairs of a positive link and negative link, larger score n′ times and equal score n″ times for the positive link are obtained. While AUC is the area under ROC curve [4, 20] which is plotted by the true positive rate against false negative rate and PR-AUC is the area under P-R curve [3] which is plotted by the precision against recall. Several elements used in other evaluation measures are explained as follows:

- TP (True Positive): the number of interacting protein pairs predicted correctly
- TN (True Negative): the number of non-interacting protein pairs predicted correctly
- FP (False Positive): the number of non-interacting protein pairs predicted as interacting protein pairs

- FN (False Negative): the number of interacting protein pairs predicted as non-interacting protein pairs

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - score = \frac{2PR}{P + R} \tag{3}$$

$$AUC(L3) = \frac{n' + 0.5n''}{n} \tag{4}$$

Among all the models, aMMSB with improved version is the most excellent one. Numerous measures are promoted by aMMSB with improved version in contrast with original aMMSB model. By relaxing the criterion, aMMSB achieves F1-score of 0.846 with AUC of 0.757 and PR-AUC of 0.892 in *yeast*, F1-score of 0.575 with AUC of 0.638 and PR-AUC of 0.690 in *krogan* and F1-score of 0.708 with AUC of 0.733 and PR-AUC of 0.986 in *human*, which outperforms original aMMSB. And our model is also superior to L3 obviously. The detailed results are shown in Fig. 3 and 4 and Table 2.

Table 2. The performance on three dataset

Dataset	Model	F-measure			AUC	PR-AUC	AUC(L3)
		Precision	Recall	F1-score			
Yeast	L3	0.842	0.393	0.536	0.637	0.866	0.600
	aMMSB	0.773	0.111	0.194	0.504	0.775	0.480
	Our approach	**0.854**	**0.847**	**0.846**	**0.757**	**0.892**	**0.810**
Krogan	L3	**0.942**	0.185	0.309	0.570	**0.909**	0.520
	aMMSB	0.866	0.108	0.191	0.505	0.867	0.505
	Our approach	0.880	**0.427**	**0.575**	**0.638**	0.893	**0.690**
Human	L3	0.982	0.388	0.556	0.612	0.979	0.605
	aMMSB	0.966	0.116	0.207	0.504	0.967	0.505
	Our approach	**0.986**	**0.552**	**0.708**	**0.733**	**0.986**	0.795

From Table 2, we can figure out that recall in *krogan* and *human* is in a low value compared to *yeast*. That is because *krogan* and *human* networks are larger than *yeast* network, when optimize parameters, mixed memberships and community strength in small network are tend to converge and be closer to true value. Thus, there are multiple false negative results in *krogan* and *human*, model in *yeast* dataset performs better. Besides, we assign mixed memberships at random to nodes which appear in test set and

not appear in train set, the predicted results involving these nodes are related to a random parameter which is uncontrollable and may leads to false negative results, such uncontrollable nodes in *krogan* accounts for a larger proportion than yeast and human which leads to lower recall value in *krogan*.

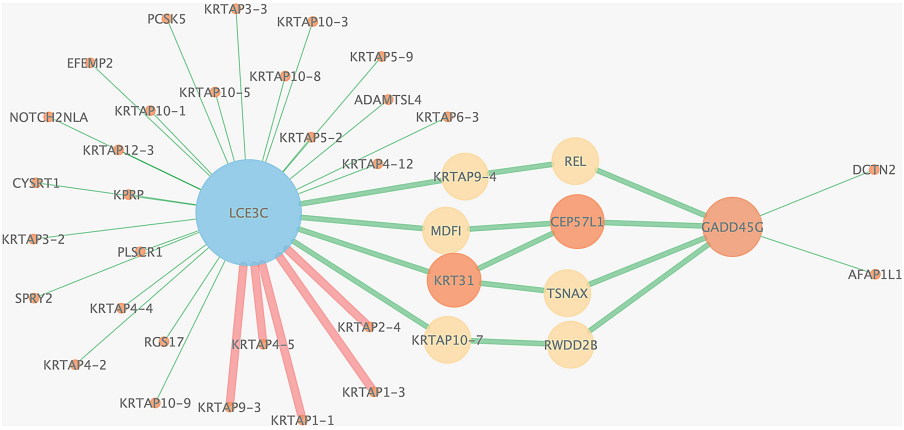

Fig. 5. A part of PPIs identified by our approach

L3 Limitations

L3 is a network-based approach to predict PPIs, it opposes the traditional idea that if two proteins share multiple neighbors, they tend to interact with each other. The author thinks similar proteins are not necessarily interact and interacting proteins are not necessarily similar. Through the study of a variety of protein pairs, the author draws a conclusion that proteins are likely to interact not if they are similar to each other, but if one of them is similar to the other's partners and the probability that two proteins interact is determined by their L3 score which is calculated by Eq. 5. For each protein pair, X and Y:

$$p_{XY} = \sum_{U,V} \frac{a_{XU} a_{UV} a_{VY}}{\sqrt{k_U k_V}} \tag{5}$$

where k_U is the degree of node U and a_{XU} equals to 1 if protein X and protein U interact, and 0 otherwise. In contrast to numerous traditional models based on the similarity of proteins, L3 achieves better results. However, L3 is still not strong enough and have some limitations for PPIs prediction.

L3 dependents on the known interacting protein pairs strongly, it supports the idea that if protein A interact with protein B and protein A is similar with protein C, protein B can interact with protein C, too. This idea requires a complete database, incomplete PPIs dataset limits the performance of L3 vastly. Supposing there is a protein D which appears in test set immensely, many protein pairs containing protein D need to be predicted, but it never appears in training set. Obviously, the probability that protein

pairs involving protein D interact can't be predicted. Because we don't know which proteins are able to interact with protein D and which proteins are similar with protein D according to training set. If there are many proteins like protein D in test set, the performance of L3 is greatly compromised. Nevertheless, we describe each protein with a mixed membership in aMMSB model to represent its degree membership for each community even when the node appears in test set and not appears in training set, which allows us to predict interacting probability of all the protein pairs involved in test set. In additional, for more comprehensive results, the author checked the performance of paths up to $l = 8$, and $l = 3$ is proved to be the best path in predictive power. Consistent with this, our model is flexible in the number of community. The community strength for PPIs network and the community membership for each node are both K dimensional vector where K is available to set. Given different number K, aMMSB can get different performance.

Beyond that, the hypothesis of L3 is not always correct, we show the superiority of aMMSB over L3 with a specific example in Fig. 5. Based on the hypothesis of L3 with the thick green lines, L3 score of LCE3C and GADD45G is not equal to zero obviously. In fact, they are predicted to interact by L3 finally. However, the truth is that LCE3C and GADD45G are in negative test set, they don't interact with each other, which is correctly predicted by aMMSB. Except predictive power, aMMSB also provides new mechanistic insights into protein function. With the thick red lines in Fig. 5, LCE3C belongs to LCE family and it is a structural component of the cornified envelope of the stratum corneum involved in innate cutaneous host defense. KRTAP belongs to KRTAP family where KRTAP 1-1 and KRTAP 1-3 belongs to KRTAP type 1 family, KRTAP 2-4 belongs to KRTAP type 2 family, KRTAP 4-5 belongs to KRTAP type 4 family and KRTAP 9-3 belongs to KRTAP type 9 family. Keratin-associated proteins (KRTAP) consists hair keratin intermediate filaments in the hair cortex, they are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins. The interactions between LCE3C and KRTAP 1-1, KRTAP 1-3, KRTAP 2-4, KRTAP 4-5, KRTAP 9-3 have been proved in IntAct [15], all of them can participant the biological process of keratinization (GO: 0031424). Their interactions are correctly predicted by aMMSB, but they are predicted to not interact in L3.

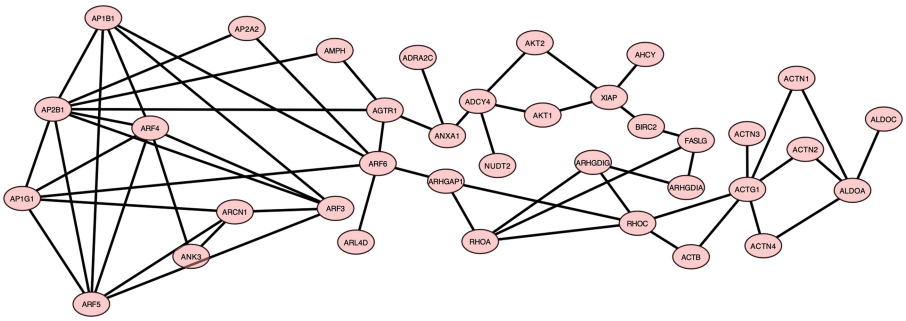

Fig. 6. A part of PPIs predicted by our approach and not included in our dataset

Predicted Protein Pairs Outside the Dataset
Besides the test set, our model can also predict the protein pairs out of the dataset. aMMSB allows each protein to own a mixed membership and assign community strength to each community. With only the two elements, the interacting probability of any protein pairs can be predicted. That is, aMMSB can predict any combinations of protein pairs in dataset only with their mixed memberships and community strength. If there is N proteins in dataset, instead of being limited to test set, $N(N-1)/2$ protein pairs can be predicted by our model. Figure 6 shows a part of protein pairs which is predicted to interact and not belongs to the dataset. All the interactions in Fig. 6 have been proved in STRING [24].

4 Conclusion

We apply a probabilistic model aMMSB to predict PPIs, this model is able to classify protein pairs into interacting pairs and non-interacting pairs with high accuracy and high precision, it is also available to calculate the exact probability of interacting pairs. The basic idea of our approach is that a node may belongs to several communities, it is common with the reality of proteins in multiple biological processes. Besides, the great performance of our approach also provides us a new sight to cellular mechanisms research. The success of the method is based on its ability to capture the structure of whole network and mixed membership of each node. The experimental results show that our approach obtains excellent performance in numerous evaluating measures and outperforms L3 model greatly.

Funding. This work has been supported by the National Natural Science Foundation of China [grant number 61602352], and the Pioneer Hundred Talents Program of Chinese Academy of Sciences.

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. **9**(Sep), 1981–2014 (2008)
2. Bass, J.I.F., Diallo, A., Nelson, J., Soto, J.M., Myers, C.L., Walhout, A.J.: Using networks to measure similarity between genes: association index selection. Nat. Methods **10**(12), 1169 (2013)
3. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)
4. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
5. Ge, H., Liu, Z., Church, G.M., Vidal, M.: Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. Nat. Genet. **29**(4), 482–486 (2001)
6. Gopalan, P.K., Gerrish, S., Freedman, M., Blei, D.M., Mimno, D.M.: Scalable inference of overlapping communities. In: Advances in Neural Information Processing Systems, pp. 2249–2257 (2012)

7. Granovetter, M.S.: The strength of weak ties. In: Social Networks, pp. 347–367. Elsevier (1977)
8. Ho, Y., et al.: Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature **415**(6868), 180–183 (2002)
9. Hu, L., Chan, K.C.: Discovering variable-length patterns in protein sequences for protein-protein interaction prediction. IEEE Trans. Nanobiosci. **14**(4), 409–416 (2015)
10. Hu, L., Chan, K.C.: Extracting coevolutionary features from protein sequences for predicting protein-protein interactions. IEEE/ACM Trans. Comput. Biol. Bioinf. **14**(1), 155–166 (2016)
11. Hu, L., Hu, P., Yuan, X., Luo, X., You, Z.H.: Incorporating the coevolving in- formation of substrates in predicting hiv-1 protease cleavage sites. IEEE/ACM Trans. Comput. Biol. Bioinf. (2019, early access)
12. Hu, L., Yuan, X., Hu, P., Chan, K.C.: Efficiently predicting large-scale protein- protein interactions using MapReduce. Comput. Biol. Chem. **69**, 202–206 (2017)
13. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehen- sive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. **98**(8), 4569–4574 (2001)
14. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Mach. Learn. **37**(2), 183–233 (1999)
15. Kerrien, S., et al.: The intact molecular interaction database in 2012. Nucleic Acids Res. **40** (D1), D841–D846 (2012)
16. Keskin, O., Tuncbag, N., Gursoy, A.: Predicting protein–protein interactions from the molecular to the proteome level. Chem. Rev. **116**(8), 4884–4909 (2016)
17. Kovács, I.A., et al.: Network-based prediction of protein interactions. Nature Commun. **10** (1), 1–8 (2019)
18. Krogan, N.J., et al.: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature **440**(7084), 637–643 (2006)
19. Mann, M., Pandey, A.: Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. Trends Biochem. Sci. **26**(1), 54–61 (2001)
20. Metz, C.E.: Basicprinciplesofrocanalysis. In: Seminarsinnuclearmedicine, vol.8, pp. 283–298. WB Saunders (1978)
21. Mirabello, C., Wallner, B.: InterPred: a pipeline to identify and model protein–protein interactions. Proteins: Struct., Funct., Bioinf. **85**(6), 1159–1170 (2017)
22. Rolland, T., et al.: A proteome-scale map of the human interactome network. Cell **159**(5), 1212–1226 (2014)
23. Simmel, G.: Soziologie: Untersuchungen u ̈ber die formen der vergesellschaftung. BoD–Books on Demand (2015)
24. Szklarczyk, D., et al.: The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. **45**, gkw937 (2016)
25. Tong, A.H.Y., et al.: A combined ex-perimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295**(5553), 321–324 (2002)
26. Tong, A.H.Y., et al.: Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science **294**(5550), 2364–2368 (2001)
27. Tong, A.H.Y., et al.: Global mapping of the yeast genetic interaction network. Science **303** (5659), 808–813 (2004)