Protein-protein Interaction Prediction via Collective Matrix Factorization

Qian Xu
Bioengineering Program
HKUST
Hong Kong, China
fleurxq@ust.hk

Evan Wei Xiang
Dept. of Computer Science and Engineering
HKUST
Hong Kong, China
wxiang@cse.ust.hk

Qiang Yang
Dept. of Computer Science and Engineering
HKUST
Hong Kong, China
qyang@cse.ust.hk

***Abstract*—Protein-protein interactions (PPI) play an important role in cellular processes and metabolic processes within a cell. An important task is to determine the existence of interactions among proteins. Unfortunately, existing biological experimental techniques are expensive, time-consuming and labor-intensive. The network structures of many such networks are sparse, incomplete and noisy, containing many false positive and false negatives. Thus, state-of-the-art methods for link prediction in these networks often cannot give satisfactory prediction results, especially when some networks are extremely sparse. Noticing that we typically have more than one PPI network available, we naturally wonder whether it is possible to 'transfer' the linkage knowledge from some existing, relatively dense networks to a sparse network, to improve the prediction performance. Noticing that a network structure can be modeled using a matrix model, in this paper, we introduce the well-known Collective Matrix Factorization (CMF) technique to 'transfer' usable linkage knowledge from relatively dense interaction network to a sparse target network. Our approach is to establish the correspondence between a source and a target network via network similarities. We test this method on two real protein-protein interaction networks, *Helicobacter pylori* (as a target network) and *Human* (as a source network). Our experimental results show that our method can achieve higher and more robust performance as compared to some baseline methods.**

***Keywords*-protein-protein interactions; transfer learning; Collective Matrix Factorization**

I. Introduction

Protein-protein interactions (PPIs) can reveal insights on biological regulatory pathways and metabolic processes. A complete and reliable protein interaction map provide us with an opportunity to understand the basic biological processes within a cell. Global interaction patterns among proteins, for example, can suggest new drug targets and aid the design of new drugs by providing a clear picture of biological pathways in the neighborhoods of the drug targets [1]. Therefore, considerable attention has been spent on the issue of protein-protein interaction prediction [2].

In biological research, various experimental approaches have been developed to map the interactions among the proteins, such as mass spectrometry, yeast two-hybrid, tandem affinity purification and co-immunoprecipitation. As a result, large-scale maps of PPIs become available. Unfortunately, interactions among proteins detected are largely incomplete because wet-lab experiments are time-consuming and labor-intensive, and the known number of possible protein-protein interactions is small. Take the proteins encoded by human genome as an example. There are about 312 million possible interactions for 25,000 proteins [1], but the actual known interactions are actually far less than this number. Moreover, the currently known interactions are noisy in that there exist high false positive and false negative rates [3]. As a result, the study of protein-protein interaction is still a very challenging topic. To solve this problem, researchers have investigated a large number of computational methods including many in machine learning area, in order to make accurate and robust link prediction.

A main problem with the link prediction problem is that network sparsity problem in the PPI networks. Traditional supervised classification methods may fail since we can not accumulate enough training data to build a steady and effective classifier. When the network is sparse, overfitting can easily happen, which causes significant performance degradation. To solve this problem, our approach is to invoke the knowledge we have about network linkage in other related PPI networks and transfer the shared knowledge to a target sparse network. In the 'transfer learning' setting, a relatively dense interaction network is assume to exist as the source network, and the objective is to infer the PPI network links in a relatively sparse target network. Noticing that the network structure can be captured using a matrix model, we exploit the Collective Matrix Factorization (CMF) methods [4] using similarities of proteins between two interaction networks as the correspondence knowledge. We show that when the source matrix is sufficient dense and similar to the target PPI network, transfer learning is effective for predicting protein-protein interactions in a sparse network. In this method, similarities between protein entities are computed by taking both protein sequences and topological structures of interaction networks into account. To evaluate our method, *Human protein interaction network*, which is relatively dense, is used as the source network and *Helico protein interaction network*, which is very sparse, is used as the target network. Our experimental results demonstrate that our proposed method indeed leads to significant performance improvement of protein interaction prediction when applied

on real-world protein-protein interaction datasets.

II. Related Work

Due to the importance of understanding the protein-protein interactions, a large number of computational methods have been developed. In these methods, supervised learning is a dominant approach. The state-of-the-art supervised learning methods include K-nearest neighbor (KNN), support vector machines (SVMs), random forest and so on. Supervised learning aims at training a classifier using positive examples of truly interacting protein pairs and negative examples of non-interacting protein pairs, to predict an unobserved relationship between two proteins. Each protein pair is encoded as a feature vector in the data. Thus, much effort has been spent on developing informative and effective feature representation methods for PPI prediction. Feature vectors may be extracted based on protein sequences directly or may involve indirect evidences, including domain compositions, motif pairs and related mRNA expression [5], [6], [7], [8], [9], [10], among others. Bock and Gough [13] used SVM method based on compositions of amino acids and physiochemical descriptors. Urquiza et al. [3] extracted 26 genomic or proteomic features of yeast from diverse databases for each pair, such as information of protein domains, domain-domain interactions in proteins whose 3D-structures are known, and high quality annotation of gene ontology. Espadaler et al. [14] considered protein structural similarities among domains found in the databases of interacting proteins, combining conservation of pairs of sequence patches based on the observation that structural evidence has shown that usually interacting pairs of close homologs physically interact in the same way. Several methods help to infer protein interactions based on the conservation of gene neighborhood, conservation of gene order, gene fusion events, and the co-evolution of interacting protein pair sequences [15][16]. Qi et al. [18] split indirect features into roughly homogeneous sets of feature experts, who employ logistic regression to estimate prediction values and combine the prediction results of individual experts.Comprehensive reviews of these methods can be found in [19], [20].

There has also been work on exploiting useful knowledge of protein interaction networks across organisms. An intuitive idea is to use the interaction map of one organism as a template to predict interactions in another [1]. Wojcik and Schachter [21] applied the 'Interaction Domain Profile Pairs' method (IDPP) to protein interaction datasets of *Escherichia coli* and *Helicobater pylori*. They first converted the source dataset *Helicobater pylori* into an abstract interaction map linking clusters of interaction domains. They then inferred unobserved interactions by establishing the correspondence between this abstract map and the target *Escherichia coli* proteome. Unfortunately, the IDPP method required a high-quality protein interaction map to exist. However, in the real world, protein interaction datasets are often sparse, incomplete and noisy, which motivates our research.

In machine learning, researchers have begun to apply matrix factorization based methods for transfer learning. An important application is recommendation systems, where collaborative filtering is modeled using matrix factorization. Similar to our motivation, they also face the problem of network sparsity. To solve the problem, various matrix based transfer learning methods have been developed [22], [23], [24]. A main difference from our work is that these research works are aimed at improving the 'rating' of a product by a user. In our case, we are interested in predicting links between nodes, which corresponds to binary connectivity, where the proteins have different semantics from users and products.

III. Methods

Our objective is to predict the interaction links between nodes in a protein-protein interaction network G_t, where the subscript t stands for 'target.' To apply a supervised learning method, we first train a classification model based on known protein pairs, each of which being represented by a feature vector $v = {x_1}^v, {x_2}^v, ..., {x_n}^v$. We then make predictions on unobserved interaction links among protein pairs. As mentioned above, we will apply Collective Matrix Factorization (CMF) [4] and exploit a known and relatively dense PPI network to help improve the target-domain prediction.

We first introduce matrix factorization (MF) [25] methods for an individual network. Matrix factorization is increasingly popular in link prediction in many domains, including social networks. We consider a network $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where there are $m = |\mathbf{V}|$ nodes. The links between $\mathbf{V}$ can be represented by an $m \times n$ adjacency matrix X. Our aim is to seek a low-rank approximation for X with the form of $X \approx f(UV^T)$. In this way, the observed links in the matrix X can be approximated by a product of two low-rank matrices, $U \in R^{m \times k}$ and $V \in R^{n \times k}$, where $k > 0$ is the rank, and f is possibly a nonlinear link function. The information carried by the adjacency matrix X is encapsulated by the two factor matrices,and thus the missing values in X can also be recovered by $\hat{X} \approx f(UV^T)$. The goal of the matrix factorization process is to seek a factor matrix pair (U, V) minimizing the error measure $||X - UV^T||$. Different function f and definitions of error measure result different models. Unfortunately, X may not be factorized into U and V successfully when X is too sparse since the learned factors U and V might be biased towards the few observed entries in the sparse X, causing overfitting.

In this paper, we are consider two protein interaction networks G and P. G is our target protein interaction network and P, relatively dense, is our source network. We will try to improve the of predicting links in G with the aid of P. The links of network G and P are represented by an

$m \times m$ matrix $L_{m\times m}$ and an $n \times n$ matrix $L_{n\times n}$, respectively. The rows and columns of $L_{m\times m}$ and $L_{n\times n}$ correspond to protein entities in networks G and P, respectively. The elements in both matrices $L_{m\times m}$ and $L_{n\times n}$ indicate the existence of interaction links.

Now we can combine the two matrices $L_{m\times m}$ and $L_{n\times n}$ together to form a big matrix X^t with the size of $(m+n) \times (m+n)$:

$$X^t = \begin{bmatrix} L_{m\times m} & 0 \\ 0 & L_{n\times n} \end{bmatrix}$$

However, we will find that factorizing X^t is actually equivalent to factorizing $L_{m\times m}$ and $L_{n\times n}$ individually. Because we are not aware of the correspondence between nodes in G and P, the information carried by the dense matrix $L_{n\times n}$ cannot be successfully transferred to help with the factorization of the sparse matrix $L_{m\times m}$. Thus, we still need to involve some other information serving as an information bridge between G and P. This bridge is the similarity between the two networks.

Consider a similarity matrix $S_{m\times n}$ introduced as the correspondence between networks G and P. The rows and columns of $S_{m\times n}$ correspond to proteins in networks G and P, respectively, and the element S_{ij} of $S_{m\times n}$ represents similarity between node i in network G and node j in network P. The collective matrix factorization method reconstructs matrices $X^t \approx f_1(ZV^T)$ and $X^a \approx f_2(UV^T)$ together by sharing the common factor V. The objective of collective matrix factorization then is to minimize the regularized loss:

$$\begin{aligned} \mathcal{L}(X^t, X^a, U, V, Z) &= D(X^t || ZV^T) \\ &+ \lambda^a D(X^a || UV^T) \\ &+ \lambda^U ||U||_F^2 + \lambda^V ||V||_F^2 + \lambda^Z ||Z||_F^2 \end{aligned} \quad (1)$$

,
where

$$X^t = \begin{bmatrix} L_{m\times m} & 0 \\ 0 & L_{n\times n} \end{bmatrix}$$

and

$$X^a = \begin{bmatrix} 0 & S_{m\times n} \\ S'_{m\times n} & 0 \end{bmatrix}$$

In Equation 1, the first and second terms are loss functions that compute the distance between original matrix and its factorized results. The rest of the terms are regularization terms that are used to prevent overfitting. The parameters λ^s, λ^U, λ^V and λ^Z are used to control the weight of target network or model complexity. In our experimental setting, we focus on predicting existence of links. Therefore, a logistic loss function is adopted when computing D. Note that there are two underlying assumptions for our collective matrix factorization process: 1) the latent factors of two nodes in one network are similar if there exists a link between them, and 2) the latent factors of two nodes in different networks are similar if the similarity between them is high.

We exploit the IsoRank [26] method to construct our similarity matrix $S_{m\times n}$, which can be thought as a particular spectral method for global graph alignment. This is based on the assumption that a protein in one PPI network is a good matching for a protein in another network if their respective sequences and neighborhood topologies are a good match. Thus, IsoRank can give a good matching of protein nodes between two PPI networks by simultaneously considering node similarities and network similarities. The main idea of the IsoRank algorithm can be formalized as follows:

$$R(i,j) = \sum_{v\in N(i)} \sum_{u\in N(j)} \frac{1}{|N(u)||N(v)|} R(u,v) \quad (2)$$
$$i \in V_S, j \in V_T,$$

where $N(i)$ denotes the set of neighbors of i, V_S denotes the set of vertices of graph S and element $R(i,j)$ represents the similarity between a vertex i of graph S and a vertex j of graph T. In the case of PPI networks, $R(i,j)$ represents the similarity between proteins i and j. The intuitive idea behind this recursive formula is that the more i and j have similar neighbors, the greater the similarity measure between i and j will be.

Equation 2 can be rewritten from a matrix perspective:

$$R = AR, \quad (3)$$

where A is the $N^2 \times N^2$ matrix defined as:

$$A(i,j)(u,v) = \begin{cases} \frac{1}{|N(u)||N(v)|} & \text{if}(i,u) \in E_S, (j,v) \in E_T \\ 0 & \text{otherwise} \end{cases}$$

where E_S denotes the set of edges of graph T.

To estimate R, we observe that Equation 3 is an eigenvalue problem, where the value of R is the principle eigenvector of A. A is a stochastic matrix so that the principal eigenvalue is 1. In our case, A and R are both sparse although A is typically a very large matrix. Singh et al. [26] propose to update R efficiently using the power method with the form of:

$$R(k+1) \leftarrow \frac{AR(k)}{\|AR(k)\|}. \quad (4)$$

To use other information to improve our prediction, we can also take the information on protein sequence similarities into account. The eigenvalue equation 3 can be rewritten to a convex combination of network and sequence similarity scores, which can be solved by similar techniques as Equation 3:

$$R = \alpha AR + (1-\alpha)C, \quad (5)$$

In this equation, C is a *normalized* score matrix generated by pairwise blast alignment method, and $\alpha \in [0, 1]$ controls the trade-off between both objectives, e.g., $\alpha = 0$ implies no network data will be considered, whereas $\alpha = 1$ indicates only network data will be used. Tuning α allows us to find the optimal alignment.

For the pairwise alignment method, we adopt the well-known Smith-Waterman algorithm [27], which is based on dynamic programming and is implemented in a Matlab toolbox. The Smith-Waterman algorithm is a conventional local pairwise alignment method [27]. It attempts to align segments of all possible lengths and optimize the similarity measure instead of considering the total sequence for sequence alignment between two protein sequences. When applying the Smith-Waterman algorithm in the Matlab toolbox, we can set the parameters BLOSUM50, 8, 8 as the scoring matrix, gap open value and extend gap value, respectively. We input two proteins as a query and get a pair-wise score with respect to their similarity. As a result, we get a score matrix C, each element of which indicates the similarity of two proteins.

IV. Materials and Results

A. Benchmark datasets

In this work, we used the *Helicobacter pylori dataset* as the target dataset and the *Human dataset* as the source dataset, both of which are also used in [28], [29], [30], [31], [12]. The target data set *Helicobacter pylori dataset* consists of 1,458 positives (interacting pairs) and 1,458 negatives (non-interacting pairs). The *Human dataset* consists of 941 positive samples (interacting pairs) and 941 negative samples (non-interacting pairs).

The network density of these datasets are given below:

- The *Helicobacter pylori dataset*: 0.12%
- The *Human dataset*: 0.27%

We randomly sample 9/10, 5/10, 1/3, 2/10, 1/10 of total instances of *Helicobacter pylori dataset*, respectively, in order to test the performance of the approach under different target network sparsity. Under these settings, the density of the target networks becomes 0.12 times f, where $f = 9/10, 5/10, 1/3, 2/10, 1/10$, respectively. When varying the target network density, the source network *Human dataset* is kept constant. We repeat each experiment for ten trials and then report the average results.

B. Performance measurement

The area under receiver operating characteristic (ROC) curve (AUC) is a statistic used as our performance measurement. ROC figures, which have been used increasingly in machine learning and data mining community [32], plot the true positive rate (TPR or sensitivity) against the false positive rate (FPR or 1-specficity), where,

$$\begin{aligned} \mathrm{TPR} &= \frac{\text{Positives correctly classified}}{\text{Total positives}} \\ \mathrm{FPR} &= \frac{\text{Negatives incorrectly classified}}{\text{Total negatives}} \\ \text{Specificity} &= 1 - \mathrm{FPR} \\ \text{Sensitivity} &= \mathrm{TPR} \end{aligned}$$

TPR and FPR depend on the classifier function h and the threshold θ used to convert $h(x)$ to a binary prediction. Varying the threshold $theta$ results ROC curve. The area under the curve (AUC) indicates the performance of this classifier: the large the better [33].

C. Results

To evaluate our proposed methods, we chose four baseline methods. We describe these methods in details below and explain the reasons for us to choose these classifiers as baseline algorithms.

The first baseline method is low-rank matrix factorization on single network, which is a special case of Equation 1 with parameter $\lambda^s = 0$ and $\lambda^U = 0$. Comparing collective matrix factorization with single matrix factorization can help illustrate if transferring knowledge from a dense interaction network to a sparse interaction network can indeed help improve performance of inferring unobserved interactions. We use *MF* to represent this baseline method.

The second baseline method adopts the support vector machine (SVM) classifier, a state-of-the-art classification method for protein-protein interaction prediction. SVM requires the input samples to be represented by feature vectors and their corresponding labels. In our implmentation, 2-Grams amino acid compositions are extracted as feature vectors for protein instances to build SVM model. The method is denoted by *SVM2Gram*.

Our third baseline *Ensemble of k-local hyperplanes* was proposed by Nanni and Lumini [12] . This mehtod trained an ensemble of K-local hyperplane distance nearest neighbor classifiers based on the same datasets as ours and first generated feature vectors using a different physicochemical property of the amino acids, which combines the amino acid indices together with 2-Grams amino acid compositions.

We chose a hybrid method introduced in [11] as our last baseline. It first encodes protein pairs as the sum and absolute minus value of PseAAc composition pairs and then applies a hybrid feature selection system mRMR-KNNs-wrapper in order to obtain an optimized feature set by excluding poor-performed and redundant features. Based on the optimized feature subset, a prediction model was trained and tested in the k-nearest neighbors learning system. In the following, this method is represented by *hybridFSKNN* if applying feature selection and represented by *non-hybridFSKNN* if not applying feature selection.

In the experimental setting, we chose parameters achieving best results for baseline methods. More specifically, $\lambda^U = \lambda^V = \lambda^Z = 5$ was set for CMF; RBF kernel with $\gamma = 0.0004$ was chosen for SVM classifier; $K = 5$ was determined for KNN classifiers in the third and fourth baseline methods, leading to the best results. Note that we will report both performances of the fourth baseline with hybrid feature selection and without feature selection procedure. Performances of baseline methods and our proposed method are compared in Table I.

As shown in the above table and figure, our suggested method *Collective Matrix Factorization (CMF)*, which transfers knowledge from an auxiliary *Human interaction network* to a sparse target *Helicobacter pylori interaction network*, achieves 5-7% improvement as compared to the best baseline results under different parameter settings. To build a correspondence between the auxiliary network and target network, node similarities involving both protein sequences and network structures are computed using the graph matching method IsoRank. Most significantly, as the density of the target network *Helicobacter pylori dataset* reduces to 0.012%, we can still get promising results using our CMF method. This result illustrate the power of our approach in solving the network sparsity problem in PPI prediction.

V. Conclusion

In this paper, we proposed a Collective Matrix Factorization (CMF) solution to solving the network sparsity problem for PPI prediction. CMF is an extension of classical matrix factorization, which we exploited for use under a novel transfer learning framework. Our aim is to infer interactions in our target protein-protein interaction network by transferring network connectivity knowledge from a source network via the help of a similarity matrix. CMF achieves significant performance improvement through the correspondence between source network and target network. To compute the similarity of two networks, IsoRank is used recursively to compute similarities of protein entities between two interaction networks by considering both protein sequences and topology structures of interaction networks simultaneously. In our experiment, we use *Helicobacter pylori interaction dataset* as target network, which is sparse and *Human interaction dataset* as relatively dense source network. The experimental results show that the performance of uncovering interaction links in the target network can indeed be boosted with the aid of source network via transferring useful knowledge.

In the future, we will involve multiple source networks to aid link prediction in the target network, and investigate methods to improve the computation of similarity between networks.

References

[1] J. Yu and F. Fotouhi, "Computational approach for predicting protein-protein interactions: a survey," *Journal of Medical System*, vol. 30, no. 1, pp. 39–44, 2006.

[2] R. Mrowka, A. Patzak, and H. Herzel, "Is there a bias in proteome research?" *Genome Res.*, vol. 11, no. 12, pp. 1971–1973, 2001.

[3] J. M. Urquiza, I. Rojas, H. Pomares, J. P. Florido, G. Rubio, L. J. Herrera, J. C. Calvo, and J. Ortega, "Method for prediction of protein-protein interactions in yeast using genomics/proteomics information and feature selection," in *Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence*, 2009, pp. 853–860.

[4] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 650–658.

[5] S. Gomez, W. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[6] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain-domain interactions from protein-protein interactions," *Genome Res.*, vol. 12, no. 10, pp. 1540–1548, 2002.

[7] H. Wang, E. Segal, A. Ben-Hur, D. Koller, and D. Brutlag, "Identifying protein-protein interaction sites on a genome-wide scale," *In Advances in Neural Information Processing Systems*, vol. 17, no. 1, pp. 1465–1472, 2005.

[8] H. Wang, E. Segal, A. Ben-Hur, Q. Li, M. Vidal, and D. Koller, "Insite: A computational method for identifying protein-protein interaction binding sites on a proteome-wide scale," *Genome Biology*, vol. 8, no. 9, pp. 1–18, 2007.

[9] M. Li, L. Lin, X. Wang, and T. Liu, "Protein-protein interaction site prediction based on conditional random fields," *Bioinformatics*, vol. 23, pp. 597–604, 2007.

[10] S. Wu and Y. Zhang, "A comprehensive assessment of sequence-based and template-based methods for protein contact prediction," *Bioinformatics*, vol. 24, pp. 24–931, 2008.

[11] L. Liu, Y. Cai, W. Lu, K. Feng, C. Peng, and B. Niu, "Prediction of protein protein interactions based on pseaa composition and hybrid feature selection," *Biochemical and Biophysical Research Communications*, vol. 380, no. 2, pp. 318–322, 2009.

[12] L. Nanni and A. Lumini, "An ensemble of k-local hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 22, no. 10, pp. 1207–1210, 2006.

[13] J. Bock and D. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, pp. 455–460, 2001.

[14] J. Espadaler, O. Romero-Isart, R. M. Jackson, and B. Oliva, "Prediction of proteinprotein interactions using distant conservation of sequence patterns and structure relationships," *Bioinformatics*, vol. 21, no. 16, pp. 3360–3368, 2005.

Table I
PERFORMANCE COMPARISON(AUC %)

- 1 - sample sizes of Helicobacter pylori for training
- **Our CMF based Methods**:
- 3 - CMF with $\lambda^s = 0.2$ and $\lambda^U = \lambda^V = \lambda^Z = 5$
- 4 - CMF with $\lambda^s = 0.4$ and $\lambda^U = \lambda^V = \lambda^Z = 5$
- 5 - CMF with $\lambda^s = 0.6$ and $\lambda^U = \lambda^V = \lambda^Z = 5$
- 6 - CMF with $\lambda^s = 0.8$ and $\lambda^U = \lambda^V = \lambda^Z = 5$
- **Baselines**:
- 2 - MF
- 7 - SVM2Gram
- 8 - Ensemble of k-local hyperplanes
- 9 - non-hybridFSKNN
- 10 - hybridFSKNN(feature number)

1	2	3	4	5	6	7	8	9	10
1/10	80.81±0.0188	**82.63±0.0149**	82.50±0.0156	82.05±0.0161	76.55±0.0182	80.95	73.80	72.90	76.00 (82)
1/5	86.04±0.0126	89.27±0.0128	**89.39±0.0181**	87.27±0.0071	83.32±0.0118	83.36	80.10	76.30	75.90 (160)
1/3	88.20±0.0063	89.62±0.0085	**90.93±0.0092**	90.48±0.0090	88.35±0.0067	85.29	81.60	80.90	81.60 (151)
1/2	89.12±0.0112	90.32±0.0083	91.33±0.0059	**91.97±0.0060**	91.39±0.0109	86.62	85.40	84.50	85.40 (182)
9/10	89.31±0.0139	92.42±0.0154	94.47±0.0173	**96.72±0.0192**	95.41±0.0115	88.84	82.40	83.70	82.40 (162)

[15] B. Shoemaker and A. Panchenko, "Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners," *PLoS Comput Biol.*, vol. 3, no. 4, p. e43, 2007.

[16] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[17] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics (Proceedings of the Intelligent Systems for Molecular Biology Conference)*, vol. 21(Suppl 1), pp. 38–46, 2005.

[18] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "A mixture of feature experts approach for protein-protein interaction prediction," *BMC Bioinformatics*, vol. 8(Suppl 10), p. 6, 2007.

[19] A. Arkin, "Synthetic cell biology," *Current Opinion in Biotech*, vol. 12, no. 6, pp. 638–644, 2001.

[20] B. Shoemaker and A. Panchenko, "Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners," *PLoS Comput Biol*, vol. 3, no. 4, pp. 595–601, 2007.

[21] J. J. Wojcik and V. Schchter, "Protein-protein interaction map inference using interacting domain profile pairs," *Bioinformatics*, vol. 17, no. Suppl 1, pp. 296–305, 2001.

[22] B. Cao, N. Liu, and Q. Yang, "Transfer learning for collective link prediction in multiple heterogeneous domains," in *Proceedings of 27th International Conference on Machine Learning*, Haifa, Israel, June 2010.

[23] B. Li, Q. Yang, and X. Xue, "Transfer learning in collaborative filtering for sparsity reduction," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, Georgia, USA, July 2010.

[24] W. Pan, E. Xiang, N. Liu, and Q. Yang, "Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction," in *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, Pasadena, CA, USA, July 2009.

[25] A. P. Singh and G. J. Gordon, "A unified view of matrix factorization models," in *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pp. 358–373.

[26] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *PNAS*, vol. 105, no. 35, pp. 12 763–12 768, 2008.

[27] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, pp. 95–197, 1981.

[28] J. Bock and D. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–135, 2003.

[29] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[30] L. Nanni, "Fusion of classifiers for predicting protein-protein interactions," *Bioinformatics*, vol. 68, pp. 289–296, 2005.

[31] ——, "Hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 69, no. 1-3, pp. 257–263, 2005.

[32] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[33] M. G. Culver, "Active learning to maximize area under the roc curve," in *Proceedings of the 6th International Conference on Data Mining*, 2006, pp. 149–158.