Predicting Protein-Protein Interactions from Multimodal Biological Data Sources via Nonnegative Matrix Tri-Factorization

Hua Wang, Heng Huang*, Chris Ding, and Feiping Nie

Department of Computer Science and Engineering,
University of Texas at Arlington, Arlington, TX 76019
{huawangcs,feipingnie}@gmail.com, {heng,chqding}@uta.edu

Abstract. Due to the high false positive rate in the high-throughput experimental methods to discover protein interactions, computational methods are necessary and crucial to complete the interactome expeditiously. However, when building classification models to identify putative protein interactions, compared to the obvious choice of positive samples from truly interacting protein pairs, it is usually very hard to select negative samples, because non-interacting protein pairs refer to those currently without experimental or computational evidence to support a physical interaction or a functional association, which, though, could interact in reality. To tackle this difficulty, instead of using heuristics as in many existing works, in this paper we solve it in a principled way by formulating the protein interaction prediction problem from a new mathematical perspective of view — sparse matrix completion, and propose a novel Nonnegative Matrix Tri-Factorization (NMTF) based matrix completion approach to predict new protein interactions from existing protein interaction networks. Because matrix completion only requires positive samples but not use negative samples, the challenge in existing classification based methods for protein interaction prediction is circumvented. Through using manifold regularization, we further develop our method to integrate different biological data sources, such as protein sequences, gene expressions, protein structure information, *etc*. Extensive experimental results on *Saccharomyces cerevisiae* genome show that our new methods outperform related state-of-the-art protein interaction prediction methods.

Keywords: Protein-Protein Interaction, Multimodal Biological Data, Nonnegative Matrix Factorization.

1 Introduction

Proteins play an essential role in nearly all cellular functions such as promoting biochemical reactions and composing cellular structures. The multiplicity of functions that proteins execute in most cellular processes and biochemical events

* Corresponding author.

B. Chor (Ed.): RECOMB 2012, LNBI 7262, pp. 314–325, 2012.

is attributed to their interactions with other proteins. As a result, it is critical to understand protein-protein interactions (PPIs) in both scientific research and practical applications such as new drug development. A variety of techniques are now available to experimental biologists for discovering protein-protein interactions, such as yeast two-hybrid systems [14], mass spectrometry [13], and many others as surveyed in [27]. Although these high-throughput experimental methods have accumulated a large amount of data, interactomes of many organisms are far from complete [27]. The low interaction coverage along with the experimental biases toward certain protein types and cellular localizations reported by most experimental techniques call for the development of computational methods that are able to predict more reliable putative PPI for further experimental screening [28].

Recently, machine learning techniques, such as Bayesian networks [15], decision trees [32,7], random forest [22,6], and support vector machines (SVM) with different kernels [20,2,25,19], have been successfully applied to predict PPIs. These methods used various data sources to train a classifier to distinguish positive examples of truly interacting protein pairs from the negative examples of non-interacting pairs. However, all these methods suffer from a fundamental difficulty — how to choose the negative samples. Compared to the obvious choice of positive samples from truly interacting protein pairs, the selection of negative samples typically is not easy. First, non-interacting protein pairs refer to those currently without experimental or computational evidence to support a physical interaction or functional association. In reality, however, such protein pairs could interact. Second, the number of non-interacting protein pairs is much larger than the number of the interacting ones, therefore unbalanced training data often cause skewed prediction models that lead to unsatisfactory prediction results. In most existing classification based methods, heuristics are often employed to tackle these problems. By recognizing these difficulties, instead of considering PPI prediction as a classification problem, we approach it from a new perspective by using matrix completion, which is an important mathematical topic to address the problem to recover a matrix from what appears to be incomplete, or even corrupted [5]. Because matrix completion only uses truly interacting protein pairs without requiring negative training samples, the difficulty in classification based PPI prediction methods is circumvented.

In this paper, we propose a novel Nonnegative Matrix Tri-Factorization (NMTF) [16,11] based matrix completion approach to predict protein-protein interactions. NMTF focuses on the analysis of data matrices whose elements are nonnegative, such as the adjacency matrix of a PPI graph, and decomposes the input matrix into three nonnegative factor matrices that approximate the input matrix by a low-rank nonnegative representation [9,10]. We first employ NMTF approach to predict putative protein interactions, which only makes use of PPI network data. After that, we extend the standard NMTF framework by adding manifold regularization [12], such that additional biological data, *e.g.*, protein sequences data, protein structures information, and gene expressions, can be incorporated to achieve enhanced PPI prediction performance.

Extensive empirical evaluations on *Saccharomyces cerevisiae* genome have shown encouraging performance, which demonstrate the effectiveness the proposed methods.

2 Methods

We first briefly formalize the problem of PPI prediction. Given a PPI network, we may construct a graph $G = (V, \Omega, \mathbf{X})$, with V corresponding to $n = |V|$ proteins and $\Omega \subseteq V \times V$ corresponding to known PPIs. $\mathbf{X} \in \{0,1\}^{n \times n}$ is the adjacency matrix, such that $\mathbf{X}_{ij} = 1$ if $(i,j) \in \Omega$, *i.e.*, there exists a PPI between protein i and protein j, and $\mathbf{X}_{ij} = 0$ otherwise. Our task is to identify a subset of non-interacting protein pairs $M \subseteq (V \times V) \setminus \Omega$ which tend to interact and can be served as potential targets for further experimental screening.

Throughout this paper, we denote matrices as boldface upper case characters. Given a matrix $\mathbf{M}$, its Frobenius norm and trace are denoted as $\|\mathbf{M}\|$ and $\mathbf{tr}(\mathbf{M})$ respectively. For convenience, given an index set M of a matrix $\mathbf{X}$, we define $\mathbf{X}_M$ as following:

$$(\mathbf{X}_M)_{ij} = \begin{cases} \mathbf{X}_{ij}, & \forall\ (i,j) \in M, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

2.1 Predict New Protein Interactions via PPI Networks

Objective to Predict PPIs. We first predict protein interactions only using PPI network data. Considering the protein interaction prediction as a matrix completion problem, where the input PPI adjacency matrix $\mathbf{X}$ contains missing entries (pairs of proteins whose interactions are yet to be determined), we wish to predict $\mathbf{Y}$ which has full entries, *i.e.*, every elements of $\mathbf{Y}$ is filled with computed values. $\mathbf{Y}$ completes $\mathbf{X}$ in the sense that $\mathbf{Y}_\Omega = \mathbf{X}_\Omega$, or more explicitly, $\mathbf{Y}_{ij} = \mathbf{X}_{ij}, \forall (i,j) \in \Omega$, where Ω denotes the set of edges where the input adjacency matrix $\mathbf{X}$ has known values (the set of interacting edges). Mathematically, the PPI prediction problem can be solved as the following optimization problem:

$$\min_{\mathbf{Y}}\ J_1 = \|\mathbf{X} - \mathbf{Y}\|_\Omega^2 = \sum_{(i,j)\in\Omega} (\mathbf{X} - \mathbf{Y})_{ij}^2\ . \tag{2}$$

Due to the low-rank nature of the adjacency matrix of an input PPI network as discussed earlier, the completed matrix $\mathbf{Y}$ can be factorized and written as $\mathbf{Y} = \mathbf{HSH}^T$, where $\mathbf{H} \in \mathbb{R}_+^{n \times k}$ and $\mathbf{S} \in \mathbb{R}_+^{k \times k}$ are the factor matrices with nonnegative elements. As a result, $\mathbf{Y} = \mathbf{HSH}^T$ can be seen as a low-rank representation of the input matrix $\mathbf{X}$ with rank of $k \ll n$. Thus we can rewrite Eq. (2) as following:

$$\min_{\mathbf{H}\geq 0,\ \mathbf{S}\geq 0}\ J_2 = \|\mathbf{X} - \mathbf{HSH}^T\|_\Omega^2\ . \tag{3}$$

Note that, although other low-rank matrix approximation methods, *e.g.*, singular value decomposition (SVD), exist, using NMTF as in Eq. (3) to constrain

the factor matrices $\mathbf{H}$ and $\mathbf{S}$ to be nonnegative is a natural choice as all the entries of the adjacency matrix $\mathbf{X}$ of the input PPI network are positive by definition. Moreover, because of the clustering interpretation of NMTF [9,11], other biological data sources can be easily incorporated via manifold regularization as introduced later.

The Solution Algorithm. Different from standard NMTF based objectives as in [16,9,11,31], which are defined over the entire input nonnegative matrix, the objective in Eq. (3) for PPI prediction is defined over a subset of the entries that correspond to known PPIs. Therefore, the solution algorithms to standard NMTF cannot be directly applied to solve Eq. (3). To this end, we present an iterative algorithm in Algorithm 1 to solve Eq. (3). The main computational load of Algorithm 1 is step 4 to solve a symmetric NMF problem, whose numerical solution was just proposed in our recent work in [31].

Algorithm 1. Algorithm to solve Eq. (3).

Input: Input PPI adjacency matrix $\mathbf{X}$;
Index set of known PPIs Ω.
begin
1. $t = 0$;
2. Initialize $\mathbf{Z}^{(0)} = \mathbf{X}_{\Omega}$;
while *not converge* **do**
3. $t = t + 1$;
4. Solve

$$\arg\min_{\mathbf{H}\geq 0\ \mathbf{S}\geq 0} \left\| \mathbf{Z}^{(t-1)} - \mathbf{H}^{(t)}\mathbf{S}^{(t)}\left(\mathbf{H}^{(t)}\right)^T \right\|^2 \tag{4}$$

to obtain $\mathbf{H}^{(t)}$ and $\mathbf{S}^{(t)}$;
5. Compute $\mathbf{Y}^{(t)} = \mathbf{H}^{(t)}\mathbf{S}^{(t)}\left(\mathbf{H}^{(t)}\right)^T$;
6. Compute $\mathbf{Z}^{(t)} = \mathbf{X}_{\Omega} + \mathbf{Y}_M^{(t)}$;
end
end
Output: Output matrix with filled missing entries $\mathbf{Y}$.

Solving Eq. (3) by Algorithm 1 for matrix completion, our NMTF approach to predict PPIs is proposed.

2.2 Predict New Protein Interactions from Multimodal Biological Data

In last subsection, we infer putative protein interactions only from PPI network data, while in practice we may also have other biological data, such as protein sequence data [20,25,19] and 3D protein structures [2,7,23], and so on. To exploit these useful information, in this subsection, we further develop the proposed NMTF based matrix completion approach.

An important reason of the popularity of NMTF in statistical learning lies in its close connection to k-means clustering [9,11]. Specifically, given a symmetric nonnegative input matrix $\mathbf{W}$, the resulted factor matrix $\mathbf{H}$ can be seen as the clustering indications of the vertices [18]. Therefore, if we have biological data other than PPI networks appearing in form of pairwise similarity, we can incorporate them through manifold regularization [8,26]. Specifically, let $\mathbf{W}_{(k)}\,(0 \leq k \leq K)$ be a set of pairwise similarities constructed from different biological data, an integrated similarity among proteins can be constructed as $\mathbf{W} = \sum_k \eta_k \mathbf{W}_{(k)}$ $(\eta_k \geq 0, \sum_k \eta_k = 1)$, where η_k are parameters to balance the data from different sources. With $\mathbf{W}$ we further develop the objective in Eq. (3) as following:

$$\begin{aligned} \min_{\mathbf{H}\geq 0,\ \mathbf{S}\geq 0} \quad & J_4 = \|\mathbf{X} - \mathbf{HSH}^T\|_{\Omega}^2 + 2\lambda\, \mathbf{tr}\left(\mathbf{H}^T\left(\mathbf{D} - \mathbf{W}\right)\mathbf{H}\right), \\ & s.t.\ \ \mathbf{H}^T\mathbf{DH} = I, \end{aligned} \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal entries $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ are the degree of the corresponding data points on $\mathbf{W}$, and λ is a parameter to balance the relative importance of the regularization term which is empirically selected as $\lambda = 0.01$ in all our experimental evaluations. Because $\mathbf{H}$ can be seen as the "soft" clustering labels [9], the second term in Eq. (5) enforces the smoothness over the variation of the clustering labels with respect to the underlying manifold described by $\mathbf{W}$ [4,12], by which additional biological data sources are incorporated.

Equation (5) takes a similar form to Eq. (3), which, again, is not a standard NMTF problem. We use Algorithm 1 to solve it by replacing step 4 to minimize the following objective:

$$\begin{aligned} J_4 = & \|\mathbf{Z} - \mathbf{HSH}^T\|^2 + 2\lambda\, \mathbf{tr}\left(\mathbf{H}^T\left(\mathbf{D} - \mathbf{W}\right)\mathbf{H}\right), \\ & s.t.\ \mathbf{H} \geq 0, \mathbf{S} \geq 0, \mathbf{H}^T\mathbf{DH} = I. \end{aligned} \tag{6}$$

Solving Eq. (5) for matrix completion, our Regularized Non-negative Matrix Tri-Factorization (R-NMTF) approach for PPI prediction is proposed, which is able to utilize both PPI network data as well as other biological data.

3 Experimental Results and Discussions

3.1 Materials and Data Sources

Protein Interaction Networks. We construct PPI graphs using the protein interaction networks compiled by BioGRID database [29]. We evaluate our methods on the *Saccharomyces cerevisiae* genome, for which an undirected graph is constructed, with vertices representing proteins and edges representing observed physical interactions. When constructing the graph, we only consider the largest connected component of the physical interaction map from BioGRID database

of version 2.0.56. The details of the PPI graph of *S. cerevisiae* genome are listed in Table 1, where "coverage" stands for the percentage of known PPIs against the total number of protein pairs $(n \times (n-1)/2)$.

Table 1. PPI graphs of the *S. cerevisiae* genome constructed using BioGRID database of version 2.0.56.

Number of proteins	5056
Edges (number/coverage)	9439/0.738%

Protein Sequence Data. We downloaded protein sequence data from GenBank [3] and computed the sequence based similarity using the mismatch kernel [17]. A protein sequence s_i is first mapped to a feature vector $\Phi_{k,m}(s_i) = \{\phi_\beta(\alpha)\}_{\beta \in \mathcal{A}^k}$, where $\mathcal{A}$ is the alphabet of 20 amino acids. The neighborhood $\mathcal{N}_{k,m}(\alpha)$ of a k-mer α is the set of k-mers that differs in at most m positions. The feature vector encodes all the k-mers in the neighborhood for $\phi_\beta(\alpha) = 1$ if $\alpha \in \mathcal{N}_{k,m}(\beta)$, and 0 otherwise. Then the mismatch kernel, thereby the induced pairwise similarity between two protein sequences s_i and s_j, is computed as $W_{ij}^{(1)} = \mathcal{K}(s_i, s_j) = \langle \Phi_{k,m}(s_i), \Phi_{k,m}(s_j) \rangle$. In our empirical studies, we set $k = 6$ and $m = 1$, which is the same as in [19]. We use protein sequence data as the additional biological data source, *i.e.*, $W = W^{(1)}$.

Protein Annotation Data. We use the functional annotations defined by Gene Ontology (GO) Consortium [1], which is a set of structured vocabularies organized in a rooted directed acyclic graph (DAG), describing attributes of gene products (proteins or RNA) in three categories of "cellular component", "molecular function" and "biological process".

3.2 Improved Prediction Capability in Cross-Validation

We first evaluate the proposed methods and compare their prediction capabilities against three most recent PPI prediction methods:
(1) Tensor product pairwise kernel (TPPK) method [20]: This method builds a kernel for pairwise objects. In order for a fair comparison, protein sequence and protein interaction network topology are used for kernel construction. PPI prediction is then carried out by the ranking scores for non-interacting protein pairs yielded by a SVM on the resulted score.
(2) Metric learning pairwise kernel (MLPK) method [23]: This method represents a pair of objects as the difference between its members, such that the resulted kernel is invariant with respect to the order of the proteins. Again, SVM is used to compute the ranking score for putative PPIs.

(3) Nearest neighbor (NN) [19] method: NN is the simplest classification method in machine learning. In [19], a ranking score for each non-interacting protein pair is computed as

$$f_{NN}(x_i) = \sum_{x_j \in (\mathcal{N}_k(x_i) \cap E)} d(x_i, x_j) - \sum_{x_j \in (\mathcal{N}_k(x_i) \cap ((V \times V) \setminus E))} d(x_i, x_j), \quad (7)$$

where $\mathcal{N}_k(x_i)$ is the set of k-nearest neighbors of x_i, and $d(\cdot, \cdot)$ is distance function built from a kernel by $d(x_i, x_j) = \sqrt{\mathcal{K}(x_i, x_i) - 2\mathcal{K}(x_i, x_j) + \mathcal{K}(x_j, x_j)}$. In our evaluations, we use the mismatch kernel for protein sequence data.

Experimental Procedures. For each method, we perform 20-fold cross-validation as following. For each trail, we remove 5% known edges (PPIs) from the input graph and try to recover them using the remained graph, which is repeated by 10 times. The average results over the 10 trials on the *S. cerevisiae* genome are reported in Fig. 1. During each trial, an internal 5-fold cross validation is performed for parameter selection. For our NMTF and R-NMTF methods, the parameter is the rank k of the factor matrices H and S. For TPPK and MLPK methods, we use the Gaussian kernel, therefore the parameters are the two regularization parameters. For NN method, we select the k of NN from $\{1, 2, 3, 5, 10, 15\}$, which is the same as in [19]. We fine tune the parameters for best prediction precision for all the compared methods.

Results. Because all compared methods produce a list of ranking scores for non-interacting protein pairs, we employ precision-recall curves to measure the prediction performance. We compute the precisions and recalls when picking up a range of top k non-interacting protein pairs as predictions, and average them over the 10 trials. The resulted precision-recall curves on the *S. cerevisiae* genome are reported in Fig. 1. From the results, we can see that the both proposed methods, NMTF and R-NMTF, consistently outperform the compared methods, sometimes very significantly. In addition, the prediction performances of R-NMTF method are always better than those of NMTF method, which is consistent with our previous theoretical analysis in that multimodal biological data, *i.e.*, protein interaction network plus protein sequence data, offer enhanced prediction performance.

A more careful analysis on the prediction results shows that the non-interacting protein pairs (including the non-interacting protein paris in the original PPI graphs and those removed due to cross-validation) with high ranking scores identified by the proposed methods typically exhibit high similarities in their functional roles. In Table 2, we list the predicted protein pairs with top 5 highest ranking scores by R-NMTF method on *S. cerevisiae* species, in which the biological functions of all protein pairs are very similar to each other. For example, "PHO91" works as "Low-affinity phosphate transporter of the vacuolar membrane", which is also the main functional role of its putative interacting partner "PHO90". Moreover, both of them have functionalities of "transcription independent of Pi and Pho4p activity" and "overexpression results in vigorous

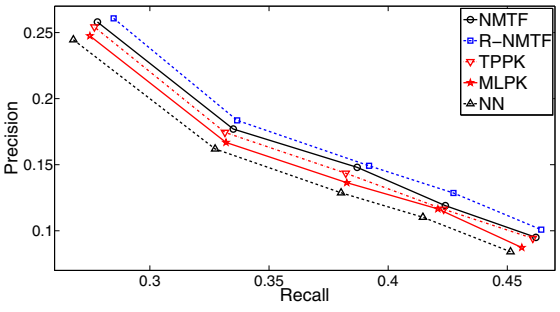

Fig. 1. Precision-recall curves by 20-fold cross-validation on *S. cerevisiae* genome by the compared PPI prediction methods

growth". These observations clearly demonstrate that these two proteins are functionally related and tend to interact with each other, which provide a concrete evidence to support that the predicted protein interactions by the proposed R-NMTF are biologically meaningful.

Note that, in our empirical studies we only use one additional data source, *i.e.*, protein sequence data, for the purpose of demonstration. In practice, more biological data could be also incorporated through a proper kernel construction under the R-NMTF prediction framework to achieve better prediction results.

3.3 Improved Protein Function Prediction Using Predicted PPI Networks

Protein interaction networks are broadly used in various biological applications, whose performances are inevitably determined by the quality input PPI graphs. Therefore, we assess the quality of predicted PPI networks in protein function prediction on *S. cerevisiae* species.

We predict protein functions on the original PPI graph constructed from the BioGRID database, the PPI graph filled by the top 1000 putative interacting protein pairs predicted by NMTF method, and that by R-NMTF method. We make predictions using the following three benchmark graph-based protein function prediction methods:

(1) Majority voting (MV) [24] method: This method assigns functions to a protein via its connecting neighbors in certain ranges.

(2) Iterative majority voting (IMV) [30] method: This method is same as MV method, but iteratively repeats the function assignment process until certain conditions are satisfied.

(3) Function Flow (FF) [21] method: This method formulates the annotation problem as a minimum multiway-cut problem, where the goal is to assign a unique function to all un-annotated proteins so as to minimize the cost of edges connecting proteins with different assignments.

Table 2. Comparisons of biological functionalities of the predicted protein interactions

LAT1	**PDX1**
Dihydrolipoamide acetyltransferase component (E2) of pyruvate dehydrogenase complex, which catalyzes the oxidative decarboxylation of pyruvate to acetyl-CoA—Lat1p; protein coding	Dihydrolipoamide dehydrogenase (E3)-binding protein (E3BP) of the mitochondrial pyruvate dehydrogenase (PDH) complex, plays a structural role in the complex by binding and positioning E3 to the dihydrolipoamide acetyltransferase (E2) core—Pdx1p; protein coding
PHO91	**PHO90**
Low-affinity phosphate transporter of the vacuolar membrane; deletion of pho84, pho87, pho89, pho90, and pho91 causes synthetic lethality; transcription independent of Pi and Pho4p activity; overexpression results in vigorous growth—Pho91p; protein coding	Low-affinity phosphate transporter; deletion of pho84, pho87, pho89, pho90, and pho91 causes synthetic lethality; transcription independent of Pi and Pho4p activity; overexpression results in vigorous growth—Pho90p; protein coding
PHO91	**PHO87**
Low-affinity phosphate transporter of the vacuolar membrane; deletion of pho84, pho87, pho89, pho90, and pho91 causes synthetic lethality; transcription independent of Pi and Pho4p activity; overexpression results in vigorous growth—Pho91p; protein coding	Low-affinity inorganic phosphate (Pi) transporter, involved in activation of PHO pathway; expression is independent of Pi concentration and Pho4p activity; contains 12 membrane-spanning segments—Pho87p; protein coding
PHO87	**PHO89**
Low-affinity inorganic phosphate (Pi) transporter, involved in activation of PHO pathway; expression is independent of Pi concentration and Pho4p activity; contains 12 membrane-spanning segments—Pho87p; protein coding	Na+/Pi cotransporter, active in early growth phase; similar to phosphate transporters of Neurospora crassa; transcription regulated by inorganic phosphate concentrations and Pho4p—Pho89p; protein coding
COY1	**SVP26**
Coy1p—Golgi membrane protein with similarity to mammalian CASP; genetic interactions with GOS1 (encoding a Golgi snare protein) suggest a role in Golgi function; protein coding	Integral membrane protein of the early Golgi apparatus and endoplasmic reticulum, involved in COP II vesicle transport; may also function to promote retention of proteins in the early Golgi compartment—Svp26p; protein coding

Table 3. Performance of protein function prediction by involved method on compared PPI graphs

	MV	IMV	FF
Original PPI graph	30.12%	30.92%	32.99%
Predicted PPI graph by NMTF method	34.85%	35.21%	36.02%
Original PPI graph by R-NMTF method	35.98%	36.33%	38.19%

We implement these methods following the details in the original literatures. Because FF method produces a ranking list of predicted protein functions, we select a threshold such that the prediction precision is maximized. 5-fold cross-validation is performed to predict the functions in "biological process" of GO. The average prediction precision over all test functions and 5 trials of cross-validation of the involved methods on different PPI graphs are reported in Table 3.

The results in Table 3 show that the function prediction performance for all three methods are improved when the predicted PPI graphs are used. Such results experimentally prove that the predicted PPI graphs have higher quality than the original one, which demonstrates that the filled putative protein interactions by the proposed methods are largely biological meaningful. Thus, we can tentatively conclude that the proposed NMTF and R-NMTF indeed can improve the protein interaction networks. Again, multimodal biological data sources based R-NMTF method is better than single data source based NMTF method.

4 Conclusions

In this paper, instead of considering protein-protein interaction prediction as a binary classification problem as in many existing works, we formulated it as a matrix completion problem. Taking this different perspective, the difficulty of selecting negative training samples required by classification based methods is averted. Moreover, because the number of protein interaction types is small, the recovery of missing PPIs from an incomplete observed protein interaction network can be suitably solved under the framework of matrix completion. We first proposed to use NMF approach to predict PPIs only from protein interaction network data, and then extended it through manifold regularization to incorporate multimodal biological data sources. We have conducted extensive empirical studies to evaluate different aspects of the proposed methods on *S. cerevisiae* genome. Promising results in the experiments validate our methods that are consistent with our theoretical analysis.

Acknowledgments. This research is supported by NSF-IIS 1117965, NSFCCF-0830780, NSF-DMS-0915228, NSFCCF-0917274.

References

1. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al.: Gene ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
2. Ben-Hur, A., Noble, W.: Kernel methods for predicting protein–protein interactions. Bioinformatics 21(suppl. 1), i38 (2005)
3. Benson, D., Karsch-Mizrachi, I., Lipman, D.: GenBank. Nucleic Acids Res. 34, D16–D20 (2006)
4. Cai, D., He, X., Wu, X., Han, J.: Non-negative matrix factorization on manifold. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 63–72. IEEE (2008)
5. Candès, E., Plan, Y.: Matrix completion with noise. Proceedings of the IEEE (2009)
6. Chen, X., Jeong, J.: Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 25(5), 585 (2009)
7. Chen, X., Liu, M.: Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21(24), 4394 (2005)
8. Chung, F.: Spectral Graph Theory. Amer. Math. Society (1997)
9. Ding, C., He, X., Simon, H.: On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proc. SIAM Data Mining Conf., Citeseer, pp. 606–610 (2005)
10. Ding, C., Li, T., Jordan, M.: Convex and semi-nonnegative matrix factorizations for clustering and low-dimension representation. Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-60428 (2006)
11. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 126–135. ACM (2006)
12. Gu, Q., Zhou, J.: Co-clustering on manifolds. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 359–368. ACM (2009)
13. Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868), 180–183 (2002)
14. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., Sakaki, Y.: Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proceedings of the National Academy of Sciences of the United States of America 97(3), 1143 (2000)
15. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N., Chung, S., Emili, A., Snyder, M., Greenblatt, J., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302(5644), 449 (2003)
16. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
17. Leslie, C., Eskin, E., Weston, J., Noble, W.: Mismatch string kernels for SVM protein classification. In: Advances in Neural Information Processing Systems, pp. 1441–1448 (2003)

18. Luo, D., Ding, C., Huang, H., Li, T.: Non-negative laplacian embedding. In: 2009 Ninth IEEE International Conference on Data Mining, pp. 337–346. IEEE (2009)
19. Martial, H., Michael, R., Jean-Philippe, V., William, N.: Large-scale prediction of protein-protein interactions from structures. BMC Bioinformatics 11(114) (2010)
20. Martin, S., Roe, D., Faulon, J.L.: Predicting protein–protein interactions using signature products. Bioinformatics 221(2), 218 (2005)
21. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21(suppl. 1), i302 (2005)
22. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random forest similarity for protein-protein interaction prediction from multiple sources. In: Pac. Symp. Biocomput., vol. 10, pp. 531–542 (2005)
23. Qiu, J., Hue, M., Ben-Hur, A., Vert, J., Noble, W.: A structural alignment kernel for protein structures. Bioinformatics 23(9), 1090 (2007)
24. Schwikowski, B., Uetz, P., Fields, S.: A network of protein–protein interactions in yeast. Nature Biotechnology 18(12), 1257–1261 (2000)
25. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H.: Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences 104(11), 4337 (2007)
26. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE. Trans. on Pattern Analysis and Machine Intelligence 22, 888–905 (2000)
27. Shoemaker, B.A., Panchenko, A.R.: Deciphering protein-protein interactions. Part I. experimental techniques and databases. PLoS Computational Biology 3(3), 334–337 (2007)
28. Shoemaker, B., Panchenko, A.: Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Computational Biology 3(4), 595–601 (2007)
29. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34(database issue), D535 (2006)
30. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nature Biotechnology 21(6), 697–700 (2003)
31. Wang, H., Huang, H., Ding, C.: Simultaneous Clustering of Multi-Type Relational Data via Symmetric Nonnegative Matrix Tri-factorization. In: The 20th ACM Conference on Information and Knowledge Management. ACM (2011)
32. Zhang, L., Wong, S., King, O., Roth, F.: Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics 5(1), 38 (2004)