Browse > Conferences> Bioinformatics and Biomedicine ...

A binary matrix factorization algorithm for protein complex prediction

Shikui Tu; Lei Xu; Runsheng Chen;
Dept. of Comput. Sci. & Eng., Chinese Univ. of Hong Kong, Hong Kong, China

This paper appears in: Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on
Issue Date: 18-18 Dec. 2010
On page(s): 113 - 118
Location: Hong, Kong
Print ISBN: 978-1-4244-8303-7
INSPEC Accession Number: 11777665
Digital Object Identifier: 10.1109/BIBMW.2010.5703783
Date of Current Version: 28 January 2011

ABSTRACT

We propose a binary matrix factorization (BMF) algorithm under the Bayesian Ying-Yang (BYY) harmony learning, to detect protein complexes by clustering the proteins which share similar interactions through factorizing the binary adjacent matrix of the protein-protein interaction (PPI) network. The proposed BYY-BMF algorithm automatically determines the cluster number while this number is usually specified for most existing BMF algorithms. Also, BYY-BMF's clustering results does not depend on any parameters or thresholds, unlike the Markov Cluster Algorithm (MCL) that relies on a so-called inflation parameter. On synthetic PPI networks, the predictions evaluated by the known annotated complexes indicate that BYY-BMF is more robust than MCL for most cases. Moreover, BYY-BMF obtains a better balanced prediction accuracies than MCL and a spectral analysis method, on real PPI networks from the MIPS and DIP databases.

INDEX TERMS

- **INSPEC**
  - **Controlled Indexing**
    Bayes methods , bioinformatics , learning (artificial intelligence) , matrix decomposition , molecular biophysics , proteins
  - **Non Controlled Indexing**
    BYY harmony learning , BYY-BMF algorithm , Bayesian Ying-Yang harmony learning , binary matrix factorization algorithm , protein complex prediction , protein-protein interaction network

Indexed by

A Binary Matrix Factorization Algorithm for Protein Complex Prediction

Shikui Tu and Lei Xu[†,*]
Department of Computer Science and Engineering,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong,
Email: {sktu,lxu}@cse.cuhk.edu.hk

Runsheng Chen[†]
Bioinformatics Laboratory and National Laboratory
of Biomacromolecules, Institute of Biophysics,
Chinese Academy of Sciences, Beijing 100101,
China, Email: crs@sun5.ibp.ac.cn

***Abstract*—We propose a binary matrix factorization (BMF) algorithm under the Bayesian Ying-Yang (BYY) harmony learning, to detect protein complexes by clustering the proteins which share similar interactions through factorizing the binary adjacent matrix of the protein-protein interaction (PPI) network. The proposed BYY-BMF algorithm automatically determines the cluster number while this number is usually specified for most existing BMF algorithms. Also, BYY-BMF's clustering results does not depend on any parameters or thresholds, unlike the Markov Cluster Algorithm (MCL) that relies on a so-called inflation parameter. On synthetic PPI networks, the predictions evaluated by the known annotated complexes indicate that BYY-BMF is more robust than MCL for most cases. Moreover, BYY-BMF obtains a better balanced prediction accuracies than MCL and a spectral analysis method, on real PPI networks from the MIPS and DIP databases.**

I. Introduction

Protein-protein interactions (PPI) play key roles in the biological processes including cell cycle control, differentiation, protein folding, signaling, transcription, translation and transport etc. Protein complexes are groups of proteins that densely interact with one another [1]. They are key molecular entities that perform cellular functions. Identifying these interacting functional modules is essential to understand the organization of biological systems. A large amount of protein interactions produced by high-throughput experimental techniques enables us to uncover the protein complexes. However, high-throughput methods are known to yield non-negligible rates of false-positives and false-negatives, due to the limitations of the experimental techniques and the dynamic nature of protein interactions. Thus, it is difficult to accurately predict protein complexes from a PPI network.

PPI networks are generally represented as undirected graphs with nodes being proteins and edges being interactions. Various algorithms have been used to detect subgraphs with high internal connectivity [2], [3], [4]. One reputed algorithm is Markov Cluster Algorithm (MCL) [5], which simulates flow in a graph, causes flow to spread out within natural clusters and evaporate inbetween different clusters. The value of a so-called inflation parameter strongly influences the clusters and the cluster number. MCL was used to detect protein families [6], and was shown to be remarkably robust against random edge additions and deletions in quantitative evaluations [3], [7]. Particularly, "MCL had the best performance on both simulated and real data sets" [7]. In addition, a spectral clustering (SC) method was introduced in [8] for finding functional modules from a PPI network. Clusters are constructed by selecting a proportion of top absolute values of elements of each eigenvector corresponding to large eigenvalues, and controlling the cluster internal connectivity and cluster-size through thresholds.

In this paper, we propose a binary matrix factorization (BMF) algorithm under Bayesian Ying-Yang (BYY) learning [9], [10], [11] to predict protein complexes from PPI networks. The BMF models the binary adjacent matrix X of the PPI interaction graph as a product of two low-rank matrices A and Y with binary entries, i.e., $X \approx AY$, where each column of Y represents the interaction pattern of the corresponding protein via weighting the columns of A. A cluster consists of proteins sharing similar interaction patterns. The roles of A and Y are exchangeable due to their symmetric positions in $X \approx AY$, and thus BMF gives a biclustering on both the rows and columns of X [12].

We propose a BMF learning algorithm, shortly denoted as BYY-BMF, under the BYY best harmony principle [9], [10]. It has the following merits: (1) It automatically determines the cluster number (or equivalently the low-rank) during the learning process, in contrast to most existing BMF algorithms [13], [14] which require a given cluster number; (2) Its clustering result does not depend on any thresholds or parameters, as opposed to MCL [5] which relies on the inflation parameter for the partition boundaries, as well as SC [8] which strongly depends on thresholds to construct clusters through eigen-decomposition. Moreover, BYY-BMF can be applied to biclustering on a rectangular dyadic matrix.

We adopt the strategy in [3] to test the performance of our algorithm. A test interaction graph is constructed from a set of annotated complexes from the MIPS database [15] by linking the proteins in the same complex, and then altered by random edge additions or deletions under various proportions to simulate the false positives and false negatives in PPI data. The predictions are evaluated with annotated complexes by Sensitivity, Positive-predictive

[*] Corresponding Author: lxu@cse.cuhk.edu.hk
[†] These authors contributed equally to this work.

978-1-4244-8302-0/10/$26.00 ©2010 IEEE

value (PPV), Accuracy and Separation [3]. Since MCL was evaluated in [3] to be more robust than other three popular complex-prediction algorithms on the above four criteria, and regarded in [7] as "the leading technique for direct and module-assisted function prediction", we focus on comparing BYY-BMF with MCL. The BYY-BMF may converge to a local optimum due to the current implementation technique. By selecting the output with the highest harmony measure under repeated random initializations, BYY-BMF's predictions are more robust against the false positives and false negatives than MCL's best predictions with the inflation parameter optimally tuned according to the test performance which is impractical because the test performance is evaluated with the true annotated complexes. Thus, BYY-BMF still has room for improvement with a possible more effective implementation guided by the harmony measure [10]. Moreover, for real PPI networks from MIPS [15] and DIP [16], the BYY-BMF by averaging all repeated evaluation results is better than MCL (with the most frequently used value for the inflation parameter) and SC, in balancing Sensitivity and PPV. In addition, we demonstrate BYY-BMF's biclustering performance on synthetic gene expression data given in [17].

II. Protein Complex Prediction Problem

The PPI network is usually represented as an undirected graph $G = (V, E)$ [3], [4], where a node v_i $(i = 1, \dots, n)$ in V represents a protein, and an edge $e = (v_i, v_j)$ in E represents an interaction between the proteins v_i and v_j. The symmetric adjacent matrix is defined as $X = [x_{ij}]$, where $x_{ij} = 1$ if there is an interaction between v_i and v_j, otherwise $x_{ij} = 0$. Mathematically, protein complexes are defined as sets of nodes with more edges amongst its members than between its members and the rest. Many methods (see e.g., [4]) were used to detect proteins complexes. A reputed one is called the Markov Cluster Algorithm (MCL) [5], which was shown to be very robust [3].

MCL [5] simulates flow using two algebraic operations on matrices. The first operation is expansion, which models the spreading out of flow. The second is inflation to model the contraction of flow, mathematically a Hadamard power followed by a diagonal scaling. The flow becomes thicker in regions of higher current and thinner in regions of lower current. MCL generates non-overlapping clusters by controlling the flow to spread out within natural clusters and to evaporate inbetween different clusters. The value of an inflation parameter strongly influences the cluster number.

A spectral clustering (SC) method was introduced in [8] to find quasi-cliques (and quasi-bipartites) in a PPI network. First, it calculates the eigen-decomposition $X = UDU^T$ for eigenvectors (the columns of U) and corresponding eigenvalues (diagonal elements of the diagonal matrix D); Then, it constructs clusters by selecting top $\alpha_{sc}\%$ absolute values of each eigenvector corresponding to large eigenvalues; Finally, it discards the nodes linked to less than $\beta_{sc}\%$ of nodes within a cluster. The obtained clusters depend on the proportion of selection $\alpha_{sc}\%$ and the internal connectivity by $\beta_{sc}\%$.

III. A Novel Binary Matrix Factorization Algorithm Under Bayesian Ying-Yang Learning

Binary Matrix Factorization (BMF) has been studied in various factorization forms [13]. In the following, we focus on $X \approx AY$, the same form as in [14], where $X = [x_{ij}]_{n \times N}$, $x_{ij} \in \{0, 1\}$, and $A = [a_{ij}]_{n \times m}$, $Y = [y_{jt}]_{m \times N}$, $a_{ij}, y_{jt} \in \{0, 1\}$. As interpreted in [12], $X \approx AY$ equivalently performs a biclustering on the rows (features) of X by A and on the columns (items) of X by Y, where each feature/item is assigned to one cluster or more. Most existing BMF algorithms are implemented for a given low-rank m (or equivalently the cluster number). For the protein-complex prediction problem, X is a symmetric binary adjacent matrix of the PPI network with $n = N$, and thus we can further constrain $A = Y^T$. Next, we propose a novel BMF algorithm under the Bayesian Ying-Yang (BYY) harmony learning [9], [10], [11].

We present a probabilistic model for the task of binary matrix factorization. The joint likelihood is $q(X, A, Y, \boldsymbol{\theta}) = q(X|A, Y, \boldsymbol{\theta}) q(A|\boldsymbol{\theta}) q(Y|\boldsymbol{\theta})$, where

$$\begin{aligned} & q(X|Y, A) = \textstyle\prod_{t=1}^{N} \prod_{i=1}^{n} (1 - u_{it})^{x_{it}} (u_{it})^{1 - x_{it}}, \\ & u_{it} = \exp\left\{-\eta \textstyle\sum_{j=1}^{m} a_{ij} y_{jt} - \nu\right\}, \eta > 0, \nu \geq 0, \end{aligned} \qquad (1)$$

$$\begin{aligned} & q(Y|\boldsymbol{\alpha}) = \textstyle\prod_{t=1}^{N} \prod_{j=1}^{m} \alpha_j^{y_{jt}}, \\ & \quad \textstyle\sum_{j=1}^{m} \alpha_j = 1, \;\; \alpha_j \geq 0, \;\; \boldsymbol{\alpha} = \{\alpha_j\}, \\ & q(A|\boldsymbol{\beta}) = \textstyle\prod_{i=1}^{n} \prod_{j=1}^{m} \beta_j^{a_{ij}}, \\ & \quad \textstyle\sum_{j=1}^{m} \beta_j = 1, \;\; \beta_j \geq 0, \;\; \boldsymbol{\beta} = \{\beta_j\}. \end{aligned}$$

where both each coloumn of Y and each row A are contrained to have one and only one "1". Furthermore, we adopt Dirichlet priors $\mathcal{D}(\boldsymbol{\alpha}|\boldsymbol{\lambda}^{\alpha}, \xi^{\alpha})$ and $\mathcal{D}(\boldsymbol{\beta}|\boldsymbol{\lambda}^{\beta}, \xi^{\beta})$ respectively for the parameter $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ with hyper-parameters $\Xi = \{\xi^{\alpha}, \boldsymbol{\lambda}^{\alpha}, \xi^{\beta}, \boldsymbol{\lambda}^{\beta}\}$, where $\mathcal{D}(\boldsymbol{z}|\boldsymbol{a}, b) = \frac{\Gamma(b)}{\prod_{j=1}^{m} \Gamma(b a_j)} \prod_{j=1}^{m} z_j^{b a_j - 1}$.

Systematically developed over a decade [11], [9], Bayesian Ying-Yang (BYY) harmony learning is a general statistical learning framework for parameter learning and model selection under a best harmony principle. For the above BMF model, the harmony measure is as follows:

$$\begin{aligned} H(p\|q) = & \sum_{\boldsymbol{A}, Y, X} \int p(\boldsymbol{\alpha}, \boldsymbol{\beta}|X) p(\boldsymbol{A}, Y|X, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(X) \\ & \cdot \ln[q(X|Y, \boldsymbol{A}) q(Y|\boldsymbol{\alpha}) q(\boldsymbol{A}|\boldsymbol{\beta}) q(\boldsymbol{\alpha}|\Xi) q(\boldsymbol{\beta}|\Xi)] d\boldsymbol{\alpha} d\boldsymbol{\beta}, \end{aligned} \qquad (2)$$

where $q(\cdot)$ gives the Ying representation, and $p(\cdot)$ gives the Yang representation. All components in Ying representation follow from the above specifications. In Yang representation, the empirical density $p(X) = \delta(X - X_N)$ is adopted with $X_N = \{\boldsymbol{x}_t\}_{t=1}^{N}$, and the other components are free, i.e., no constraints on their probability density functions.

Algorithm 1 The Sketched BYY-BMF algorithm

```
Input: data $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]$
Initialize $\boldsymbol{A}$, $m = m_{init}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\xi^{\alpha} = \xi^{\beta} = m/2$,
        $\boldsymbol{\lambda}^{\alpha} = \boldsymbol{\lambda}^{\beta} = [1, \dots, 1]/m$, $\eta = 0.98$, $\nu = 0.01$.
repeat
  Yang-Step:
    $Y^{(\tau)} = \arg\max_Y \ln[q(X|Y, \boldsymbol{A}^{(\tau-1)})q(Y|\boldsymbol{\alpha}^{(\tau-1)})]$;
    $\boldsymbol{A}^{(\tau)} = \arg\max_{\boldsymbol{A}} \ln[q(X|Y^{(\tau-1)}, \boldsymbol{A})q(\boldsymbol{A}|\boldsymbol{\beta}^{(\tau-1)})]$;
  Ying-Step:
    $\boldsymbol{\alpha}^{(\tau)} = \arg\max_{\boldsymbol{\alpha}} \ln[q(Y^{(\tau)}|\boldsymbol{\alpha})q(\boldsymbol{\alpha}|\Xi)]$;
    $\boldsymbol{\beta}^{(\tau)} = \arg\max_{\boldsymbol{\beta}} \ln[q(\boldsymbol{A}^{(\tau)}|\boldsymbol{\beta})q(\boldsymbol{\beta}|\Xi)]$;
  Model-Selection-Step:
  for $j = 1$ to $m$ do
    if $\alpha_j < \eta_0$ or $\beta_j < \eta_0$ then
      Discard the $j$-th dimension; $m \leftarrow m - 1$;
    end if
  end for
until $|H^{(\tau)}(p\|q) - H^{(\tau-1)}(p\|q)| < 10^{-5}|H^{(\tau)}(p\|q)|$
Output: $\boldsymbol{A}$, $Y = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_N]$, $m$
Notations: $m_{init}$ is an initial integer for $m$; $\tau$ is the
iteration number; $\eta_0$ is a very small positive value.
```

The best harmony, i.e, maximizing $H(p\|q)$, leads the unknown Yang components to be Dirac delta functions. To achieve the best harmony, a Ying-Yang alternative procedure is implemented and sketched in Algorithm 1. In this algorithm, the cluster number starts from a large enough m_{init}, and reduces accordingly in the "Model-Selection-Step". This automatic reduction results from a least complexity nature in maximizing $H(p\|q)$. One interpretation [10] is as follows: The maximization forces Ying representation to match Yang representation, but they may not be perfectly equal due to a finite sample size and other constraints. At the equality, $H(p\|q)$ becomes the negative entropy, further maximizing which will minimize system complexity.

Our BYY-BMF algorithm considers an effective factorization and an automatic determination of the cluster number simultaneously, while most existing BMF algorithms need a given cluster number. In the "Yang-Step", $Y^{(\tau)}$ can be computed by individual maximizations over each column of Y. It is similar for computing $A^{(\tau)}$. With Y's columns (and A's rows) having one and only one "1", the above BYY-BMF outputs non-overlapping clusters. Due to the non-convexity of eq.(2), different initializations may lead to different local optima by BYY-BMF. To tackle this problem, we can repeat random initializations and select the output with the highest harmony measure. More effective implementations are possible.

IV. Experiments

A. Data Sets

As in [4], the reference protein complexes contain, in total, 428 complexes by combining manually curated 216 complexes from MIPS [15], 92 complexes from Aloy et al. [18], and 295 complexes from the SGD database [19]. The PPI network data sets are: (1) constructed from the MIPS complexes by instantiating a node for each protein and linking by an edge any two proteins within the same complex; (2) collected from MIPS database [15], with $12,317$ interactions among 4543 proteins, or from DIP database [16] with 4405 interactions among 2144 proteins [1].

B. Evaluation Criteria

To evaluate the accuracy of the predictions, we adopt the following four criteria used in [3], [4].

Sensitivity (Sn) is defined as follows:

$$Sn = \{\textstyle\sum_{i=1}^{n} \max_j\{T_{ij}\}\} \,/\, \sum_{i=1}^{n} N_i, \tag{3}$$

where n and m is the number of reference and predicted complexes respectively, and T_{ij} is the number of common proteins in the i-th reference complex and the j-th predicted complex, and N_i is the number of proteins in the i-th reference complex. A high S_n value implies a good coverage of proteins in the reference complexes.

Positive predictive value (PPV) is defined as

$$PPV = \left\{\textstyle\sum_{j=1}^{m} \max_i\{T_{ij}\}\right\} \,/\, \sum_{j=1}^{m} T_{\cdot j}, \tag{4}$$

where $T_{\cdot j} = \sum_{i=1}^{n} T_{ij}$. A high PPV value indicates the predicted complexes are likely to be true positive.

Accuracy(Acc) is the geometric average of Sn and PPV,

$$Acc = \sqrt{S_n \times PPV}, \tag{5}$$

which balances the complementary information provided by Sn and PPV: Sn increases to 1 for the big cluster of all proteins, while PPV reaches 1 for single-protein clusters.

Separation(Sep) value is given by

$$Sep = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij} \cdot \tfrac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} s_{ij}}, \tag{6}$$

where $s_{ij} = T_{ij}^2/(T_{\cdot j}T_{i\cdot})$, and $T_{i\cdot} = \sum_{j=1}^{m} T_{ij}$. A high Sep indicates a better general correspondence between predicted and reference complexes.

[1] The file "Scere20100614CR.txt" from DIP is used, and the proteins in this file without systematic names according to the yeast protein list in "http://www.uniprot.org/downloads" are discarded.

Table I
EVALUATIONS OF DIFFERENT CLUSTERING ALGORITHMS ON THE TEST GRAPH $X_{0,0}$ (#C: NUMBER OF PREDICTED COMPLEXES)

algorithm	Sn	PPV	Acc	Sep	#C
true	**1.0000**	0.7219	0.8497	0.7826	216
BMF(opt)	0.9844	**0.8459**	**0.9125**	**0.8652**	179
BMF(avg)	0.9764	0.7805	0.8730	0.7861	147
MCL(1.8)	0.9920	0.7689	0.8734	0.8474	157
MCL(opt)	0.9818	0.7936	0.8827	0.8560	164
SC(10%, 1%)	0.6788	0.2661	0.4250	0.0238	622

C. Results

1) On Altered Graphs by Randomly Adding and Deleting Edges: As in [3], we build a *test graph* X from the MIPS complexes [15] by linking the protein nodes in the same complex. For a systematic evaluation, we alter the test graph X to be $X_{a,d}$, where a and d denote the percentages of randomly added or deleted edges with respect to the number of original edges in X. The set of percentage pairs (a, d) is $P_{AD} = \{(a, d) \,|\, a \in \{0, 0.05, 0.1, 0.2, 0.4, 0.8, 1.0\};\ d \in \{0, 0.05, 0.1, 0.2, 0.4, 0.8\}\}$.

Table I evaluates the predicted complexes by various algorithms with respect to the MIPS complexes. The "algorithm **true**" uses the MIPS complexes as the predicted complexes. The BYY-BMF algorithm is implemented with random initialization ($m_{init} = 300, \kappa = 1$) by 10^3 independent trials. The **BMF(avg)** averages the results of all trials, while the **BMF(opt)** denotes the trial with the highest value of the harmony measure by eq.(2). The **MCL(1.8)** means the MCL process with the inflation parameter being 1.8, while **MCL(opt)** denotes the MCL implementation of possible best accuracy on the test graph, with the optimal inflation parameter value 3.4 (see Table (2) in [3], where 1.8 is the most frequent value). **SC**(10%,1%) means SC is implemented with $\alpha_{sc}\% = 10\%$ and $\beta_{sc}\% = 1\%$.

The observations from Table I are as follows. (1) The **BMF(opt)** outperforms the **BMF(avg)** by relieving the local optimum problem with a better initialization guided by the harmony measure at the cost of more computation; (2) The values of the inflation parameter influences MCL's prediction accuracies; (3) The **BMF(opt)** is better than **MCL(opt)**, and they both outperform the rest algorithms.

Figure 1 presents the results of 9 out of 42 percentage pairs (a, d) in P_{AD}, due to space limit. According to Table I, we further compare the robustness of **BMF(opt)** (using the same initialization as in Table I) and **MCL(opt)** (using the optimal inflation parameter values given by the Table (2) in [3]). For each of 10 runs for each (a, d), a graph is generated via random edge additions and deletions on the test graph. The evaluation results are averaged in Figure 1.

It can be observed from Figure 1 that **BMF(opt)** becomes more robust than **MCL(opt)** as the percentages becomes large, except for the case $(a, d) = (0\%, 80\%)$. For more details about this case, we also include the results of **BMF(avg)** and **MCL(1.8)**. The results show that **MCL(opt)** (with inflation parameter being $\mathbf{1.3}$) balances the Sensitivity and PPV much better than **MCL(1.8)**, while **BMF(opt)** and **BMF(avg)** take the $2nd$ and $3rd$ places respectively.

2) On Real PPI Data Sets: Two real PPI data sets are collected from the MIPS [15] and DIP [16]. For a fair, practical comparison, we average the results of 10 runs of BYY-BMF with $m_{init} = 600$, and chose the most often used inflation parameter value 1.8 for MCL.

Figure 2 evaluates the predictions with the 428 reference complexes. BYY-BMF has a better prediction Accuracy, which balances the Sensitivity and the PPV, than MCL, followed by SC, while MCL obtains the best separation value. That the separation value of BYY-BMF is lower than MCL in Figure 2 is likely due to the initialization problem as indicated in Table I. The used reference complexes probably cannot cover all true complexes underlying the PPI networks from MIPS and DIP, and thus, as indicated in [3], PPV and Separation only indicate factional actual complexes annotated already, whereas Sensitivity is likely to provide more relevant information of the coverage of the reference complexes recovered in the predictions.

D. On Gene Expression Data for Biclustering

In addition, we demonstrate to use our BYY-BMF as a biclustering algorithm on synthetic gene expression data in [17]. The original data, which consist of non-overlapping biclusters, are added with random Gaussian noise under increasing noise levels (i.e., the standard deviation). Figure 3 indicates that the performance of BYY-BMF is very robust against noise.

V. CONCLUSION

We have proposed a Binary Matrix Factorization (BMF) algorithm under Bayesian Ying-Yang (BYY) harmony learning, to tackle the problem of predicting protein complexes from a protein-protein interaction (PPI) network. The algorithm has the following merits: (1) The input of the known cluster number required by most existing BMF algorithms is not necessary; (2) As opposed to MCL and SC, BYY-BMF has no dependence on any parameters or thresholds.

Experimental results show that our BYY-BMF algorithm, if implemented by searching the output with the highest BYY harmony measure under repeated random initializations, is more robust against PPI false positives and false negatives than MCL using optimal inflation parameters tuned by the testing accuracies. The prediction results on large real world PPI networks indicate that the average results of repeated independent trials by BYY-BMF obtains a better balanced prediction accuracy, while MCL has a relative advantage in separation value. In addition, we have demonstrated the effectiveness and robustness of BYY-BMF in biclustering on synthetic gene expression data.

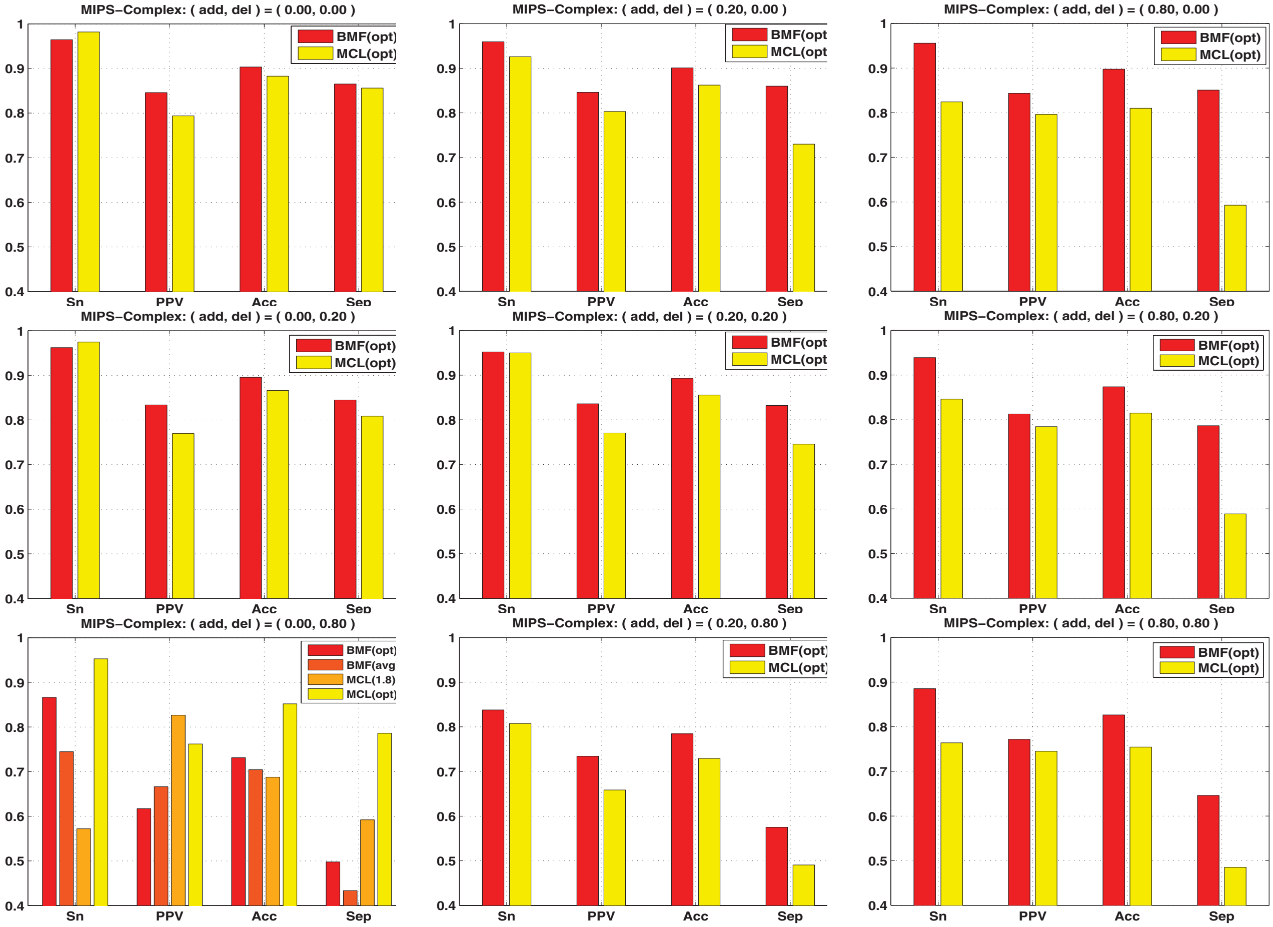

Figure 1. Prediction evaluations of BMF and MCL on altered graphes constructed from a test graph, with *add*% edges randomly added and/or *del*% edges randomly deleted with respect to the original number of edges

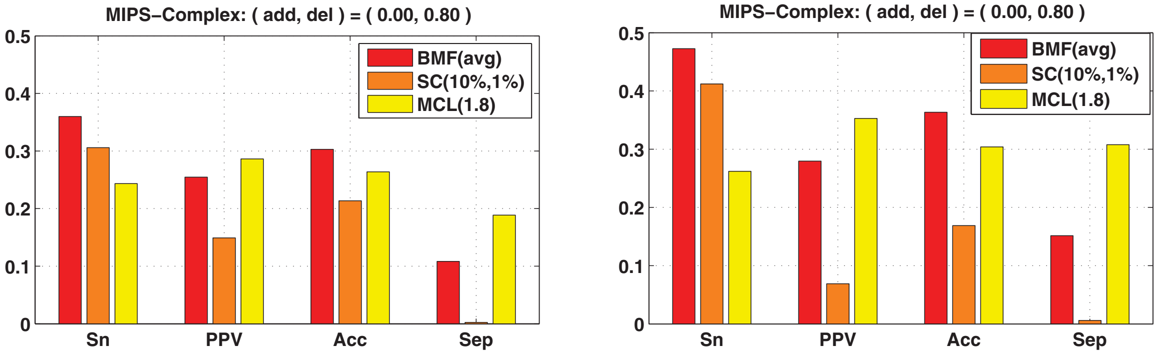

Figure 2. Prediction accuracies of BMF, MCL and SC on real world PPI networks collected from MIPS (*left*) and DIP database (*right*)

Furthermore, although BYY-BMF has a local optimum problem resulted from the current implementation procedure, the improvement by repeating the random initializations for a higher harmony measure indicates BYY-BMF still has room for improvement via more effective implementations [10]. Also, BYY-BMF can be extended and used on those data with non-overlapping clusters.

Acknowledgment

The work described in this paper was fully supported by a grant of the General Research Fund (GRF) from the Research Grant Council of the Hong Kong SAR (Project No:CUHK4180/10E), and by National Key Basic Research & Development Program 973 under Grant No. 2009CB825404.

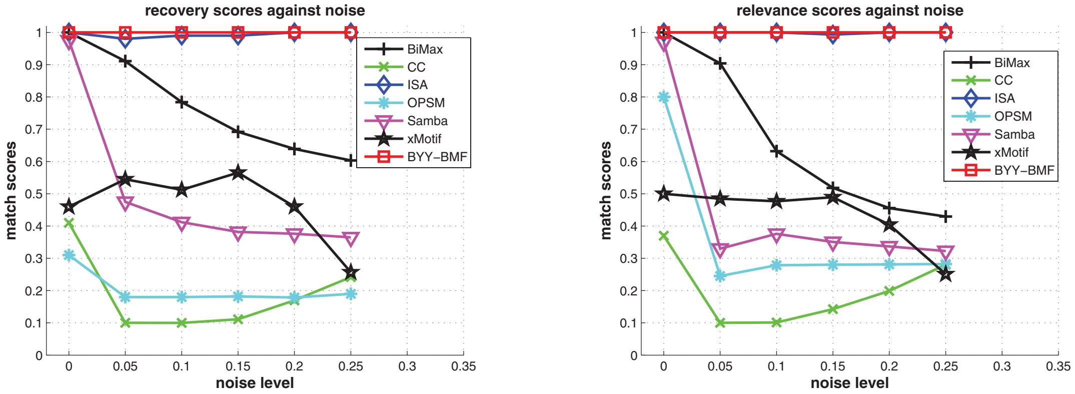

Figure 3. The matching score is calculated by DEFINITION 2 in [17]. The bicluster relevance reflects to what extent the generated biclusters represent true biclusters, while the module recovery quantifies how well each true bicluster is recovered. The details of other algorithms are referred to [17].

REFERENCES

[1] J. Poyatos and L. Hurst, "How biologically relevant are interaction-based modules in protein networks?" *Genome Biology*, vol. 5, no. 11, p. R93, 2004.

[2] X. H. Daniel Wu, "Topological analysis and sub-network mining of protein-protein interactions," in *Advances in Data Warehousing and Mining*, D. Taniar, Ed. Idea Group Publisher, 2006, pp. 209–240.

[3] S. Brohee and J. van Helden, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 488, 2006.

[4] M. Wu, X.-L. Li, and C.-K. Kwoh, "Algorithms for detecting protein complexes in PPI networks: An evaluation study," in *Proceedings of Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008)*, Australia, Oct, 15-17, 2008.

[5] S. van Dongen, "Graph clustering by flow simulation," Ph.D. dissertation, Univ. of Utrecht, Utrecht, The Netherlands, 2000.

[6] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucl. Acids Res.*, vol. 30, no. 7, pp. 1575–1584, 2002.

[7] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular System Biology*, vol. 3, no. 88, 2007.

[8] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[9] L. Xu, "Bayesian Ying-Yang System, Best Harmony Learning, and Five Action Circling," *A special issue on Emerging Themes on Information Theory and Bayesian Approach, Journal of Frontiers of Electrical and Electronic Engineering in China*, vol. 5, no. 3, pp. 281–328, 2010.

[10] ——, "Machine learning problems from optimization perspective," *A special issue for CDGO 07, Journal of Global Optimization*, vol. 47, pp. 369–401, 2008.

[11] ——, "A unified learning scheme: Bayesian-Kullback Ying-Yang machines," in *NIPS*, 1995, pp. 444–450.

[12] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. SIAM Data Mining Conference*, 2005, pp. 606–610.

[13] T. Li, "A unified view on clustering binary data," *Machine Learning*, vol. 62, no. 3, pp. 199–215, 2006.

[14] Z. Zhang, T. Li, C. Ding, and X. Zhang, "Binary matrix factorization with applications," in *Proc. of the 2007 Seventh IEEE International Conference on Data Mining*. Washington, USA: IEEE Computer Society, 2007, pp. 391–400.

[15] ftp://ftpmips.gsf.de/yeast/PPI/PPI 18052006.tab.

[16] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[17] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[18] P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A.-C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. B. Russell, "Structure-Based Assembly of Protein Complexes in Yeast," *Science*, vol. 303, no. 5666, pp. 2026–2029, 2004.

[19] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry, "Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 69–72, 2002.