IET Systems Biology

Special Section: Selected Papers from the International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2017)

IET Journals
The Institution of Engineering and Technology

Overlapping functional modules detection in PPI network with pair-wise constrained non-negative matrix tri-factorisation

ISSN 1751-8849
Received on 8th December 2017
Accepted on 12th December 2017
E-First on 7th February 2018
doi: 10.1049/iet-syb.2017.0084
www.ietdl.org

Guangming Liu[1], Bianfang Chai[2], Kuo Yang[1], Jian Yu[1], Xuezhong Zhou[1] ✉

[1]*Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, No. 3 Shangyuancun Haidian District, Beijing, People's Republic of China*

[2]*Department of Information Engineering, Hebei GEO University, Shijiazhuang, People's Republic of China*

✉ *E-mail: xzzhou@bjtu.edu.cn*

Abstract: A large amount of available protein–protein interaction (PPI) data has been generated by high-throughput experimental techniques. Uncovering functional modules from PPI networks will help us better understand the underlying mechanisms of cellular functions. Numerous computational algorithms have been designed to identify functional modules automatically in the past decades. However, most community detection methods (non-overlapping or overlapping types) are unsupervised models, which cannot incorporate the well-known protein complexes as a priori. The authors propose a novel semi-supervised model named pairwise constrains nonnegative matrix tri-factorisation (PCNMTF), which takes full advantage of the well-known protein complexes to find overlapping functional modules based on protein module indicator matrix and module correlation matrix simultaneously from PPI networks. PCNMTF determinately models and learns the mixed module memberships of each protein by considering the correlation among modules simultaneously based on the non-negative matrix tri-factorisation. The experiment results on both synthetic and real-world biological networks demonstrate that PCNMTF gains more precise functional modules than that of state-of-the-art methods.

1 Introduction

Protein seldom exerts its biological function as unitary independent entity but usually plays as an organised group or functional module [1]. With the development of high-throughput experiment technology, such as mass spectrometry [2, 3], two-hybrid systems [4, 5], large amounts of protein-protein interaction (PPI) data are available which makes it possible to reveal the fundamental regular patterns of the cellular systems. Generally, these PPI data sets are expressed as undirected networks in which proteins act as a collection of vertices, and interactions between pairs of proteins play as a set of links [6]. In addition, protein networks have different topological qualities, including: (i) small-word property [7], (ii) scale-free degree distribution [8], and (iii) functional modular organisation [9]. Therefore we need to detect functional modules in PPI networks to discover the underlying mechanisms of cellular functions.

Proteins interacted with each other usually are more likely to partake the same or similar biological functions than those not interacted with each other [10]. Hence, the closely connected regions in PPI networks can be regarded as functional modules. To address this problem, a plenty of advanced computing approaches have been proposed to identify densely linked sub-graphs automated as functional modules (or protein complexes) in recent biological researches [11, 12]. In terms of the detected modules, the functional module detection methods can be divided into two categories: non-overlapping and overlapping algorithms.

An entropy-based functional module detection method has been proposed by Kenley [13] in which a protein was selected randomly as a seed and then absorbs its neighbours to form an original module, then proteins that are adjacent to this module were added or removed according to the increase or decrease of entropy. UVCluster [14], proposed by Arnau *et al.*, is a hierarchical clustering method based on the shortest path between pairs of proteins.

In recent years, a plenty of overlapping module detection methods have been proposed [12, 15–17]. Xiang *et al.* [17] have proposed a weighted gene co-expression network analysis algorithm to identify overlapping modules related to glioblastoma multiforme prognosis. Bader and Hogue have proposed a functional module detection algorithm named MCODE [18] which identifies functional modules by fully employing the degree of proteins. Another well-known overlapping functional modules detection algorithm named CFinder has been developed by Adamcsek *et al.* [15] which uncovers k-cliques by utilising clique percolation [16] firstly and then merges the adjacent k-cliques into the functional modules. Nepusz *et al.* [12] have proposed an overlapping protein module detection method called ClusterONE, in which the proteins accompany with the highest degree were selected as seeds firstly and then their neighbour nodes are decided to append or remove from them measured by a cohesiveness score.

There are some algorithms have the ability of detecting both non-overlapping and overlapping modules, such as non-negative matrix factorisation (NMF)-based methods. NMF is a broadly used matrix decomposition approach which factorises an original non-negative matrix into two non-negative matrices with low rank and it has been successfully applied in text, image, natural language analysis [19] and functional module detection [20]. Nevertheless, the physical meaning of the two factorised matrices is ambiguous. Luckily, non-negative matrix tri-factorisation (NMTF) has been proposed which can assign a clear physical meaning to each factorised matrix and we will introduce it in Section 2.2. Wang *et al.* [21] have used NMTF to co-cluster multi-type relational data simultaneously, Zhu *et al.* [22] have used NMTF to analyse both user-level and tweet-level sentiments on social media and Pei *et al.* [23] utilised NMTF to detect community structure in social networks. All these three works are unsupervised methods and the performance of them depending on the selection of similarity function which was used as manifold regularisation terms.

Only topological information is considered by the above-mentioned methods; however, PPI data acquired from high-throughput biological experiments is incomplete [24], and a plenty of noise and error interactions exist in these sparse PPI networks. For instance, the percentage of false-positive interactions is occasionally up to 50% [25]. Therefore, protein module detection methods which are simply based on topological structure may not obtain accurate functional modules. Fortunately, some manually

curated protein complex databases, such as CORUM [26], are available and in high quality. Compared to PPI, the number of proteins in protein complexes is small but these complexes can be viewed as prior information to help address the limitations of PPI for functional module detection.

To address these limitations of PPI networks, we propose a novel semi-supervised model named pairwise constrained non-negative matrix tri-factorisation (PCNMTF) which uses known high-quality protein complexes as prior information to identify functional modules more precisely than unsupervised methods. We expect to uncover new functional modules from PPI networks using prior information. Some of the detected modules are contained and some are not contained in the complex database. We first extract must-link constraints from protein complexes, where a pair of proteins within a same complex indicates a must-link constraint. Then these limited constraints are used to guide the factorising iteration. The main contributions of this work including: (1) we present a novel semi-supervised functional module detection model PCNMTF which makes full use of known protein complexes as prior information to help detecting functional modules; (ii) a Frobenius constraint is imposed on community relationship matrix $\boldsymbol{G}$ to make the solution stable; (iii) different from existing NMF and NMTF methods, the module membership of a protein is decided not only based on the indicator matrix but also in terms of the module relationship matrix.

2 Related work

Models based on NMF [27] and NMTF [28] have been successfully used in community detection in recent years. There are roughly two kinds of algorithms: unsupervised and supervised (or semi-supervised) methods. Given a similarity matrix $\boldsymbol{S}$ of a network, the module memberships of nodes are derived from it. In this section, we first introduce several classic similarity matrix calculation methods and then introduce the unsupervised and semi-supervised NMF models for module detection.

2.1 Similarity matrix of a network

Extracting similarity matrix $\boldsymbol{S}$ of nodes from the topological information is a fundamental task. There are three methods to construct similarity matrix $\boldsymbol{S}$: (i) Adjacency matrix. Using the adjacent matrix $\boldsymbol{A}$ directly as similarity matrix $\boldsymbol{S}$ or construct the matrix $\boldsymbol{S}$ based on matrix $\boldsymbol{A}$, such as $\boldsymbol{S}(\boldsymbol{V}_i, \boldsymbol{V}_j) = [\boldsymbol{A} + \alpha \boldsymbol{A}^2]$, where α is a parameter to control the role of A^2 [29]. (ii) Shortest path. If p_{ij} is the shortest path from node i to node j then $\boldsymbol{S}(\boldsymbol{V}_i, \boldsymbol{V}_j) = 1/p_{ij}^k$ [30], where k is a constant. (iii) Diffusion kernel feature matrix. First, an opposite Laplacian matrix $\boldsymbol{L}$ is constructed according to a network as follows:

$$L_{ij} = \begin{cases} 1 & \text{if } i \text{ linked to } j \\ -d_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where d_i is the degree of vertex i. Then define the exponential of matrix $\boldsymbol{L}$ as $K = \exp(\beta L)$, where β is a positive parameter to control the extent of diffusion. Finally, the similarity matrix $\boldsymbol{S}$ is acquired by $\boldsymbol{S}(\boldsymbol{V}_i, \boldsymbol{V}_j) = K_{ij}/\sqrt{K_{ij}K_{ii}}$ [31].

2.2 Unsupervised NMF

The unsupervised methods only focus on utilising topological structure of network to detect modules. Thus, the similarity matrix $\boldsymbol{S} \in \mathbb{R}_+^{n \times n}$ is viewed as an original input matrix, the NMF aims to factorise $\boldsymbol{S}$ into two non-negative low rank matrices $\boldsymbol{U} \in \mathbb{R}_+^{n \times k}$ and $\boldsymbol{V} \in \mathbb{R}_+^{n \times k}$, where $k \ll n$. We use the Euclidean distance to quantify the quality of the approximation gained by product $\boldsymbol{U}$ and $\boldsymbol{V}$. The objective function is defined as $\min_{U,V} J = \| \boldsymbol{S} - \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \|_F^2$. Meanwhile, a symmetric NMF (SNMF) has been proposed to identify community structure since $\boldsymbol{S}$ is a symmetric similarity matrix and its objective function is defined as $\min_{\boldsymbol{V}} J = \| \boldsymbol{S} - \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \|_F^2$.

Since both NMF and SNMF do not considered the relationships between modules, then NMTF is designed to uncover underlying modules from networks which is formulated as $\min_{V,G} J = \| \boldsymbol{S} - \boldsymbol{V}\boldsymbol{G}\boldsymbol{V}^{\mathrm{T}} \|_F^2$, where $\boldsymbol{V} \in \mathbb{R}_+^{n \times k}$ is the node membership indicator matrix and $\boldsymbol{G} \in \mathbb{R}_+^{k \times k}$ represents the relationship between modules. NMTF can give any connection between two nodes in one network by the term $\boldsymbol{V}\boldsymbol{G}\boldsymbol{V}^{\mathrm{T}}$ while NMF and SNMF cannot.

2.3 Semi-supervised NMF

In real-world applications, some prior information is easily obtained with pairwise form which can be used to improve the performance of community detection algorithms. In recent years, a plenty of algorithms have been proposed to incorporate these prior information to aid detecting modules. Zhang *et al.* [32] have designed a model which used the must-link constraints to enhance adjacent matrix $\boldsymbol{A}$ so that a novel adjacent matrix $\bar{\boldsymbol{A}}$ is defined as follows:

$$\bar{\boldsymbol{A}}_{ij} = \begin{cases} \alpha & \text{if } i \text{ and } j \text{ have the same label} \\ \boldsymbol{A}_{ij} + 1 & \text{if } i = j \\ 0 & \text{if } i \text{ and } j \text{ have different labels} \end{cases} \tag{2}$$

Based on the new adjacent matrix $\bar{\boldsymbol{A}}$, NMF, SNMF and NMTF are able to identify modules from PPI networks. Yang *et al.* [33] have proposed a semi-supervised module detection framework which combines NMF and SNMF with pair-wise constraints to uncover communities. The objective functions are $\min_{V,U} J = \| \boldsymbol{S} - \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \|_F^2 + \beta \mathrm{Tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{V})$ and $\min_{\boldsymbol{V}} J = \| \boldsymbol{S} - \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \|_F^2 + \beta \mathrm{Tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{V})$, where $\boldsymbol{L}$ is the Laplacian matrix of pair-wise constraints and β is a positive parameter to balance the tradeoff between topology structure and must-link information.

3 Functional module detection based on PCNMTF

Notations: A PPI network can be formed typically as an undirected graph $\boldsymbol{P} = (\boldsymbol{V}, \boldsymbol{E})$ in which $\boldsymbol{V} = \{V_i\}_{i=1}^n$ represents the proteins, and E denotes the edge set which represents the interactions between protein pairs. Let an $n \times n$ non-negative symmetric matrix $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}_+^{n \times n}$ denote the adjacency matrix of graph P, generally, the element a_{ij} denotes whether an interaction is existed between the ith protein and jth protein. For convenience, we set $a_{ij} = 1$ if and only if protein p_i interacts with protein p_j, and $a_{ij} = 0$ otherwise.

3.1 Problem statement

Given an adjacency matrix $\boldsymbol{A}$ of a PPI network and a known protein complex database, we extract pair-wise information, must-link constraint, from complex database. Thus, a must-link matrix $\boldsymbol{M}$ is built based on these must-link constraints. The goal of the proposed semi-supervised module detection model in this work is trying to find protein module membership matrix $\boldsymbol{F}$ and module relationship matrix $\boldsymbol{G}$ with the given information, adjacency matrix $\boldsymbol{A}$ and must-link matrix $\boldsymbol{M}$. We attempt to explore an objective function based on matrix factorisation which can identify underlying module structures from PPI networks and the objective function that needs to be minimised is defined as follows:

$$J(\boldsymbol{F}, \boldsymbol{G}) = Q(\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{A}) + P(\boldsymbol{F}, \boldsymbol{M}) + \rho(\boldsymbol{G}) \tag{3}$$

The first term $Q(\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{A})$ indicates the deviation between product of $\boldsymbol{F}$ and $\boldsymbol{G}$ and the adjacency matrix $\boldsymbol{A}$, the second term denotes the penalty term of must-link constraints and the last term is a regularisation term on $\boldsymbol{G}$.

3.2 Matrix tri-factorisation

The interactions between protein pairs are rare in human PPI networks [24, 34] at present. Thus, the corresponding graph P with respect to these incomplete interactions is considerable sparse. If a feature-vector of one protein is assigned directly by each row in adjacency matrix $\boldsymbol{A}$, the time consuming will be expensive due to the high dimensionality which is equal to the number of proteins in the whole PPI network. Furthermore, the performance of module detection in terms of this feature vector is unsatisfactory [35]. The NMF [30, 36] and SNMF [37] models have been proposed since they are able to explore a high-quality lower dimensional feature as the new representation for each protein in PPI network. What is more, previous studies have confirmed that NMF models offer obviously advantages in detecting modules within biological network [38]. However, the correlation between modules which denotes the interactions between modules, there will be more interactions between two overlapped modules than those non-overlapped, has not been considered when assigning module membership to a protein that may lead to an inaccurate module division result.

To overcome the drawback of NMF, then NMTF is employed in this paper and the objective function is defined as follows:

$$\min_{\boldsymbol{F}\geq 0,\boldsymbol{G}\geq 0} Q(\boldsymbol{F},\boldsymbol{G},\boldsymbol{A}) = \| \boldsymbol{A}-\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}} \|_F^2 \tag{4}$$

where $\boldsymbol{F}\in\mathbb{R}_+^{n\times k}$ is an $n\times k$ matrix representing the module membership of proteins (k is the maximum possible number of modules) and the element F_{ij} represents the probability that node i should be belonged to module j, $\boldsymbol{G}\in\mathbb{R}_+^{k\times k}$ is a $k\times k$ symmetric matrix denoting the correlations between any module pairs. The product of $\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}$ indicates the relationship between any two proteins in accordance with module structure. $\| \bullet \|_F$ denotes the Frobenius norm. Since the adjacency matrix $\boldsymbol{A}$ is positive, the non-negative constrains are also added to matrixes $\boldsymbol{F}$ and $\boldsymbol{G}$ simultaneously.

3.3 Pairwise constrained

Protein complex is a group of proteins that interact with each other densely and tend to share similar biological functions [39]. Intuitively, the proteins within a same complex should be considered to be clustered into a same module and then the must-link constraints are generated according to these proteins. Therefore, the must-link constrained matrix $\boldsymbol{M}=[m_{ij}]\in\mathbb{R}_+^{n\times n}$ is constructed in terms of extracted must-link constraints, where $m_{ij}=1$ if protein i and protein j co-occur in one common protein complex and $m_{ij}=0$ otherwise. The module membership of any protein pair, protein i and j, with must-link constraint should be similar as much as possible, which means the difference between the ith row f_i and jth row f_j in the module indicator matrix $\boldsymbol{F}$ should be as small as possible. In this paper, the square distance between two vectors is used to measure the similarity between them, which is denoted as $\mathrm{d}(f_i,f_j)=\| f_i-f_j \|_2^2$.

The must-link constraints which are used as prior information can be formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{F}\geq 0} P(\boldsymbol{F},\boldsymbol{M}) &= \frac{1}{2}\times\sum_{i,j} m_{i,j}\times \mathrm{d}(f_i,f_j)\\ &= \mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{F})-\mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{F})\\ &= \mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F})\end{aligned} \tag{5}$$

where $\boldsymbol{D}=[d_{i,j}]\in\mathbb{R}_+^{n\times n}$ is a diagonal matrix about matrix $\boldsymbol{M}$ ($d_{i,i}=\sum_{j=1}^{n} m_{i,j}$) and $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{M}$ is the Laplacian matrix of matrix $\boldsymbol{M}$, $\mathrm{Tr}(\bullet)$ indicates the trace of a matrix.

3.4 PCNMTF

There is a plenty of ways to make use of both topological information and pairwise constraints simultaneously for protein module detection. The main idea in this work is to use pairwise constraints as a penalty term rather than simply to incorporate the prior information into the original PPI network, then the objective function will be subjected to a penalty if the must-link constraints are not satisfied. To address this issue, the objective function of the proposed model PCNMTF is defined as follows:

$$\min_{\boldsymbol{F}\geq 0,\boldsymbol{G}\geq 0} J(\boldsymbol{F},\boldsymbol{G}) = \| \boldsymbol{A}-\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}} \|_F^2 + \alpha \| \boldsymbol{G} \|_F^2 + \beta\mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F}) \tag{6}$$

where β is a parameter with the function of balancing the tradeoff between prior knowledge formulated as must-link constraints and topological structure of PPI network. Furthermore, the Frobenius norm is imposed on matrix $\boldsymbol{G}$ as a regularisation term that is used to generate stable solutions for (6) and prevent overfitting, α is a smoothing parameter.

Although the proposed model PCNMTF is similar with previous studies which are proposed by Wang *et al.* [28], Zhang *et al.* [32] and Yang *et al.* [33], it is quite different among them in several aspects. Wang's method only concerned on topological information without considering prior information to discover modules from networks, it is difficult to detect modules accurately from networks with no clear modular structures. Zhang *et al.* directly used must-link constraint to modify the adjacency matrix; however, the enhanced adjacency matrix did not guarantee that a node pair with must-link constraint can be clustered into a same module. Yang *et al.* proposed a semi-supervised framework based on NMF to uncover modules; however, the physical meaning of the two factorised matrices were not clear and the relationship between modules was not learned. Our proposed model PCNMTF utilised prior information to guide the learning process of protein membership matrix and module relationship matrix simultaneously. Furthermore, we proposed a novel overlapping module detection method by considering these two matrices at the same time.

Using the knowledge of trace as follows: $\mathrm{Tr}(\boldsymbol{X})=\mathrm{Tr}(\boldsymbol{X}^{\mathrm{T}})$, $\| \boldsymbol{X} \|_F^2=\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ and $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{Y})=\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{X})$, then (6) is rewritten as follows:

$$\begin{aligned}\min_{\boldsymbol{F}\geq 0,\boldsymbol{G}\geq 0} J(\boldsymbol{F},\boldsymbol{G}) = \mathrm{Tr}(&\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{F}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}-\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\\ &+\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})+\alpha\mathrm{Tr}(\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}})+\beta\mathrm{Tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F})\end{aligned} \tag{7}$$

In order to satisfy non-negative constraints $\boldsymbol{F}=[f_{ij}]\geq 0$ and $\boldsymbol{G}=[g_{ij}]\geq 0$, we brought in two Lagrange multipliers $\Psi=[\psi_{ij}]\in\mathbb{R}_+^{n\times k}$ and $\Phi=[\phi_{ij}]\in\mathbb{R}_+^{k\times k}$ separately, then the Lagrange function of (7) is rewritten as follows:

$$\begin{aligned}\min_{\boldsymbol{F}\geq 0,\boldsymbol{G}\geq 0} J(\boldsymbol{F},\boldsymbol{G}) = \mathrm{Tr}(&\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}-\boldsymbol{A}\boldsymbol{F}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}-\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\\ &+\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})+\alpha\mathrm{Tr}(\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}})+\beta\mathrm{Tr}(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^{\mathrm{T}})\\ &+\mathrm{Tr}(\Psi\boldsymbol{F}^{\mathrm{T}})+\mathrm{Tr}(\Phi\boldsymbol{G}^{\mathrm{T}})\end{aligned} \tag{8}$$

Since (8) is non-convex in terms of both matrixes $\boldsymbol{F}$ and $\boldsymbol{G}$ as variables simultaneously, in order to minimise function J, we first acquired the partial derivative against matrixes $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively, as follows:

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{F}} &= 2\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{G}-2\boldsymbol{A}\boldsymbol{F}\boldsymbol{G}+\beta\boldsymbol{L}\boldsymbol{F}+\Psi\\ \frac{\partial J}{\partial \boldsymbol{G}} &= \boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{G}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}-\boldsymbol{F}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{F}+\alpha\boldsymbol{G}+\Phi\end{aligned} \tag{9}$$

then let (9) equal to zero and used the KKT conditions $\psi_{ik}f_{ik}=0$ and $\phi_{jk}g_{jk}=0$, then the updating rules of protein indicator matrix $\boldsymbol{F}$ and module relationship matrix $\boldsymbol{G}$ were given as follows:

Input: α, β, adjacency matrix $A \in R^{n\times n}$, must-link set: MS, number of modules k

Build the must-link constraint matrix $M \in R^{n\times n}$ according to MS
Initialize: $F \geq 0; G \geq 0$
while not converge **do**
  Fix G, and update F according to Eq. (10)
  Fix F, and update G according to Eq. (11)
  Check the convergence
end while

Output: module indicator matrix F and module relationship matrix G

Fig. 1 *Algorithm 1: The proposed PCNMTF*

$$\boldsymbol{F} = \boldsymbol{F} \otimes \frac{2\boldsymbol{AFG} + \beta\boldsymbol{MF}}{2\boldsymbol{FGF}^{\mathrm{T}}\boldsymbol{FG} + \beta\boldsymbol{DF}} \tag{10}$$

$$\boldsymbol{G} = \boldsymbol{G} \otimes \frac{\boldsymbol{F}^{\mathrm{T}}\boldsymbol{AF}}{\boldsymbol{F}^{\mathrm{T}}\boldsymbol{FGF}^{\mathrm{T}}\boldsymbol{F} + \alpha\boldsymbol{G}} \tag{11}$$

where $\otimes$ means the element-wise multiplication between two matrices. In order to minimise (6), the updating strategy employed is to update one matrix while keeping another unchanged iteratively. The iterative process will be terminated when the objective function is converged or the number of iteration bigger than a given threshold. We lay out the proposed PCNMTF model in Algorithm 1 (see Fig. 1.

3.5 Overlapping module detection

We developed a novel overlapping module detection method with using both protein module membership matrix $\boldsymbol{F}$ and module relationship matrix $\boldsymbol{G}$. The dimensions of each row in matrix $\boldsymbol{F}$ is k, which is equal to the number of all possible modules, the element f_{ij} denotes the membership strength how protein i serves to module j [37]. Intuitively, if protein i belongs to multiple modules, there must exist some relationship among them to some extent. Then, for protein i, we first assign it to module c to which it most likely belongs when module c meets $c = \underset{c}{rmarg\ \max}\ f_{ic}$. Furthermore, in addition to module c, we also consider clustering protein i into another module j if the following conditions are satisfied in the mean time: $f_{ij} \geq \omega$ and $g_{cj} > 0$ where g_{cj} is the (c, j)th element in matrix $\boldsymbol{G}$ which denotes the relationship between module j and module c. As a consequence, each protein in PPI network can be clustered into one or more modules effectively and efficiently. In this manuscript, the value of threshold ω is set equal to 0.2 by experience as a similar way of Zhang's work [40].

4 Experimental results

4.1 Data sets

We introduce two common synthetic networks to verify the effectiveness of the proposed model PCNMTF. Girvan and Newman [41] design a synthetic network benchmark generator, each network (denoted as GN network) contains 128 nodes which are belonged to four modules. The average degree of each node is 16. For each node, let Z_{in} indicate the number of edges randomly linked to it in its own module and Z_{out} denote the amount of links randomly connected to it in other modules, obviously, $Z_{\mathrm{in}} + Z_{\mathrm{out}} = 16$. As the value of Z_{out} increases the modular structure becomes less clear. Previous studies have proved that when $Z_{\mathrm{out}} > 6$ the modular structure of the generated networks becomes vague and most state-of-the-art methods are difficult to identify modules from these networks accurately. In this work, we set $Z_{\mathrm{out}} = 8$ and then generated 100 networks with benchmark randomly. The average benchmark modularity of these GN8 networks is 0.27. Lancichinetti *et al.* [42] developed another well-used artificial network benchmark generator (denoted as LFR network), it provides several parameters to control the properties of generated networks, such as the number of nodes (n), the average degree of each node (ad), the maximum degree of each node (md), the minimum module size (M_{min}), the maximum module size (M_{max}) and a mixing parameter (mp) which represents the fraction edges between modules. Similar to Z_{out} in GN networks, a larger mp leads to a more unclear modular structure network. In our experiment, we set $n = 1000$, ad = 15, md = 50, $M_{\mathrm{min}} = 20$, $M_{\mathrm{max}} = 50$ and mp = 0.7, and then we generated 100 networks with benchmark (denoted as LFR) randomly and the average benchmark modularity of these LFR networks is 0.26.

Two human related PPI networks are used in our work, one is derived from database of interacting proteins (DIP) [43] human subset and the other is human protein reference database (HPRD) [44]. Two protein complex databases are used in this work. The first one is CORUM [26] which concerns the protein complexes in mammalian, thus, the protein complexes which are not existed in human organism are filtered out in this study. The second one is PCDq [45] which concerns the human related protein complexes. The protein complexes which have less than three proteins are filtered out in our experiments. The complexes and proteins coverage of the two human related PPI networks by these two complex databases and the properties of PPI networks are listed in Table 1, where #*p* and #*e* denote the number of proteins and edges in PPI network, respectively, #*cc* and #*cp* denote the number of coverage complexes and proteins of PPI network by complex database, respectively, #*as*, #*ai* and #ad denote average size, average number of interactions and average degree of complexes, respectively.

4.2 Evaluation metrics

Since each node in the two artificial networks mentioned above has specific community membership, then the normalized mutual information (NMI) [46] and accuracy are employed to measure the quality of detected modules. The accuracy metric is used to evaluate the percentage of nodes with correct module membership identified by the community detection method. Let g_i and d_i denote the ground-truth label and detected label for node i, the accuracy can be defined as

$$\text{accuracy} = \frac{\sum_{i=1}^{n} \delta(g_i, l_{\mathrm{map}}(d_i))}{n} \tag{12}$$

where $\delta(x, y) = 1$ if $x = y$, or $\delta(x, y) = 0$ if $x \neq y$, l_{map} is a function that maps each detected label d_i to the equivalent ground-truth label g_i which is implemented by Kuhn–Munkres algorithm [47]. The NMI metric is used to measure the similarity between ground-truth module sets G_s and detected module sets D_s and is defined as follows:

$$\mathrm{NMI}(G_s, D_s) = \frac{-2\sum_{g,d=1}^{k} n_{gd}\log\left(n_{dg}n/n_g n_d\right)}{\left(\sum_{g=1}^{k} n_g \log\left(n_g/n\right)\right) + \left(\sum_{d=1}^{k} n_d \log\left(n_d/n\right)\right)} \tag{13}$$

where n_g denotes the number of proteins in the ground-truth gth module G_g and n_d is the number of proteins in the detected dth

Table 1 Properties of human networks and complexes

Network	#*p*	#*e*	CORUM					PCDq				
			#*cc*	#*cp*	#*as*	#*ai*	#ad	#*cc*	#*cp*	#*as*	#*ai*	#ad
DIP	2943	4673	746	1018	5.51	3.43	6.85	340	1090	4.58	2.22	5.83
HPRD	9453	36,888	1069	1823	5.76	5.47	30.49	874	2892	4.39	3.96	23.72

Table 2 Accuracy of compared methods with different percentage of must-link constraints on GN8

Method	0.05	0.1	0.15	0.2	0.25	0.3
MCODE	0.794 ± 0.01	0.794 ± 0.01	0.794 ± 0.01	0.794 ± 0.01	0.794 ± 0.01	0.794 ± 0.01
K-rank-D	0.739 ± 0.05	0.739 ± 0.05	0.739 ± 0.05	0.739 ± 0.05	0.739 ± 0.05	0.739 ± 0.05
NMF	0.859 ± 0.03	0.859 ± 0.03	0.859 ± 0.03	0.859 ± 0.03	0.859 ± 0.03	0.859 ± 0.03
PCNMF	0.867 ± 0.03	0.961 ± 0.02	1.000 ± 0.01	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.01
SNMF	0.867 ± 0.01	0.867 ± 0.01	0.867 ± 0.01	0.867 ± 0.01	0.867 ± 0.01	0.867 ± 0.01
PCSNMF	0.937 ± 0.01	0.984 ± 0.00	0.993 ± 0.00	1.000 ± 0.01	1.000 ± 0.00	1.000 ± 0.00
NMTF	0.862 ± 0.02	0.862 ± 0.02	0.862 ± 0.02	0.862 ± 0.02	0.862 ± 0.02	0.862 ± 0.02
NMTFADJ	0.859 ± 0.01	0.859 ± 0.01	0.859 ± 0.01	0.859 ± 0.01	0.859 ± 0.01	0.859 ± 0.01
NMTFJAC	0.846 ± 0.03	0.846 ± 0.03	0.846 ± 0.03	0.846 ± 0.03	0.846 ± 0.03	0.846 ± 0.03
PCNMTF	0.997 ± 0.01	1.000 ± 0.01	1.000 ± 0.00	1.000 ± 0.02	1.000 ± 0.00	1.000 ± 0.00

Table 3 Accuracy of compared methods with different percentage of must-link constraints on LFR

Method	0.05	0.1	0.15	0.2	0.25	0.3
MCODE	0.46 ± 0.03	0.46 ± 0.03	0.46 ± 0.03	0.46 ± 0.03	0.46 ± 0.03	0.46 ± 0.03
K-rank-D	0.53 ± 0.06	0.53 ± 0.06	0.53 ± 0.06	0.53 ± 0.06	0.53 ± 0.06	0.53 ± 0.06
NMF	0.393 ± 0.04	0.393 ± 0.04	0.393 ± 0.04	0.393 ± 0.04	0.393 ± 0.04	0.393 ± 0.04
PCNMF	0.431 ± 0.02	0.359 ± 0.02	0.776 ± 0.01	0.679 ± 0.04	0.671 ± 0.02	0.635 ± 0.06
SNMF	0.567 ± 0.02	0.567 ± 0.02	0.567 ± 0.02	0.567 ± 0.02	0.567 ± 0.02	0.567 ± 0.02
PCSNMF	0.591 ± 0.02	0.856 ± 0.01	0.977 ± 0.05	0.994 ± 0.02	1.000 ± 0.01	1.000 ± 0.03
NMTF	0.571 ± 0.04	0.571 ± 0.04	0.571 ± 0.04	0.571 ± 0.04	0.571 ± 0.04	0.571 ± 0.04
NMTFADJ	0.542 ± 0.03	0.542 ± 0.03	0.542 ± 0.03	0.542 ± 0.03	0.542 ± 0.03	0.542 ± 0.03
NMTFJAC	0.475 ± 0.05	0.475 ± 0.05	0.475 ± 0.05	0.475 ± 0.05	0.475 ± 0.05	0.475 ± 0.05
PCNMTF	0.639 ± 0.01	0.922 ± 0.02	0.999 ± 0.01	1.000 ± 0.02	1.000 ± 0.03	1.000 ± 0.01

module D_d, n is the total number of proteins in PPI network, n_{gd} is the number of proteins overlapped between module G_g and D_d.

As for the human related PPI network, the precision, recall and F-measure metrics are utilised to assess the quality of detected modules. The extent of overlapping between gold-standard complexes G_c and detected module sets D_s is presented as follows:

$$\mathrm{OL}(G_c, D_s) = \frac{|G_c \cap D_s|^2}{|G_c| \times |D_s|} \tag{14}$$

where $|G_c|$ indicates the size of one known protein complex, $|D_s|$ denotes the size of one detected protein module, and $|G_c \cap D_s|$ is the quantity of overlapped proteins between them. If $\mathrm{OL}(g, d) \geq \gamma$, the two sets p and d are considered to be matched each other. In this paper, we assign $\gamma = 0.2$ with the same manner of previous studies [12, 37]. Then the precision, recall and F-measure are defined as follows:

$$\begin{aligned} N_{cb} &= |\{b | b \in G_c, \exists p \in D_s, \mathrm{OL}(p, b) \geq 0.2\}| \\ N_{cp} &= |\{p | p \in D_s, \exists b \in G_c, \mathrm{OL}(b, p) \geq 0.2\}| \\ \text{precision} &= \frac{|N_{cp}|}{|D_s|}; \quad \text{recall} = \frac{|N_{cb}|}{|G_c|} \\ F\text{-score} &= \frac{2 \times \text{precision} \times \text{recall}}{|\text{precision} + \text{recall}|} \end{aligned} \tag{15}$$

where F-measure is the harmonic mean of recall and precision.

4.3 Performance on synthetic networks

To evaluate the module identification capability of our proposed algorithm PCNMTF, seven well-known state-of-the-art NMF-based community detection methods and two non-NMF-based methods are employed to compare with our method. The compared seven NMF-based methods include NMF [48], pair-wise constrained NMF (PCNMF) [33], symmetric NMF (SNMF) [28], pair-wise constrained SNMF (PCSNMF) [33], NMTF [49], NMTF with Jacarrd similarity matrix (MNTFJAC) and NMFADJ [21–23]. The graph regularisation term used in NMTFJAC is based on Jaccard similarity between two proteins. The proteins linked to each other in PPI network are thought to have similar functions then the adjacency matrix is viewed as a similarity matrix which is served to NMTFADJ. The two non-NMF-based methods are K-rank-D [50] and MCODE [18].

The NMI and accuracy metrics are used to evaluate the performance of module detection methods, and the parameters of PCNMF and PCSNFM are chosen to obtain the best results. The parameter β which is used to balance the tradeoff between topology information and prior information of PCNMF and PCSNMF set equal to 10 and 100 separately. Note that, when $\beta = 0$, the PCNMF is equivalent to NMF and PCSNMF is equivalent to SNMF. The sensitivity analysis of the two parameters α and β are conducted in Section 4.5. Then we set the smoothing parameter $\alpha = 0.05$ and $\beta = 10$ for the proposed method PCNMTF. $\beta = 10$ indicates that must-link constraints play an important role in detecting modules from complicated networks which is consistent with previous studies [33, 51, 52].

The must-link constraints are extracted from benchmark modules with the same way of Yang's work [33]. Suppose that there are N nodes in one module, the possible number of node pairs with must-link constraint is $N_{\mathrm{ml}} = N(N-1)/2$. The percentage of node pairs with must-link constraints are based on N_{ml} in this section. Tables 2 and 3 illustrate the accuracy of modules detected by different methods in term of various percentage prior information. Figs. 2*a* and *b* display the NMI of different algorithms with various percentage of prior information. Both the accuracy and NMI of all supervised algorithms have been improved consistently with the increase of must-link information. The proposed model PCNMTF has the best performance which has the rapidly growth trend. The NMI and accuracy of PCNMTF approach to 1 rapidly when the percentage of must-link information exceeds 10% on GN8 networks and 15% on LFR networks, which means PCNMTF can identify modules effectively and efficiently from the network with unclear modular structure. The most significant improvement of PCNMTF is due to making full use of must-link information and module correlation simultaneously.

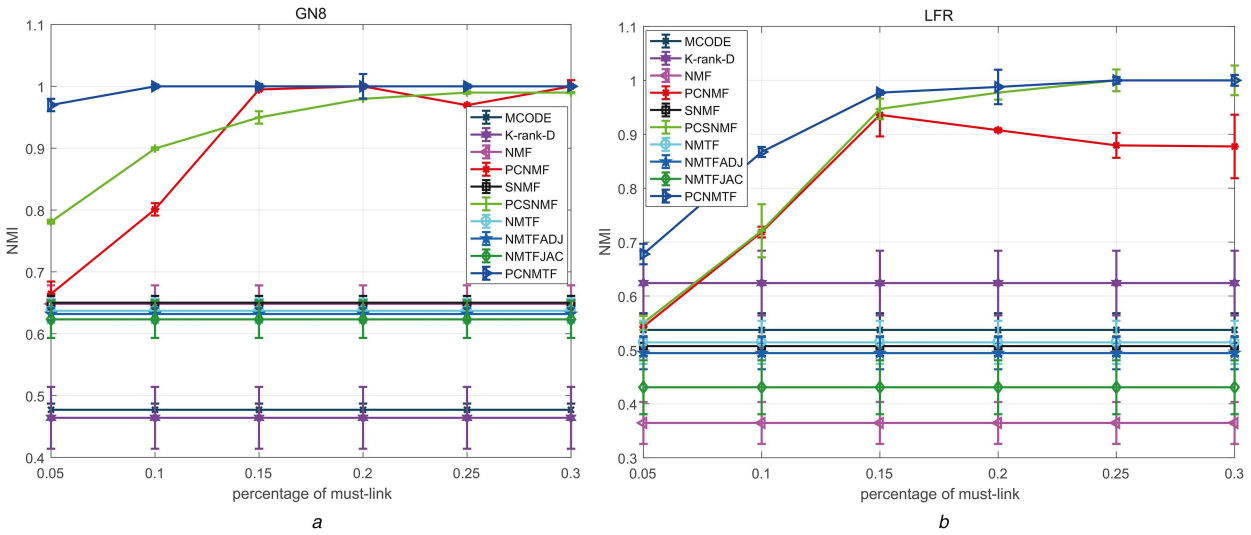


Fig. 2 *NMI of different methods with different percentage of must-link constraints derived from ground-truth*
(a) GN8 network, ***(b)*** LFR network

4.4 Performance on human PPI networks

The must-link constraints can improve the performance of detecting modules from networks, then the proposed model PCNMTF was used to detect protein functional modules on two human-related PPI networks, DIP and HPRD, with the same parameter settings as discussed in Section 4.3.

4.4.1 Must-link constraints: The must-link prior information is extracted from two known protein complex databases, CORUM and PCDq. Since the protein complexes are overlapped, the proteins included in more than one complex are not considered when we extracted must-link constraints from protein complexes. For each protein complex, the proteins only contained in one complex are used to extract must-link constraints and the number of corresponding proteins are denoted as N_p. Then we extracted $N_p(N_p-1)/2$ protein pairs with must-link constraint. However, the must-link constraint only provides the information about that the two corresponding proteins should belong to one module rather than clarify to which module they should belong. In this work, the must-link constraints are extracted from CORUM. Thus, 803 must-link constrains with 470 proteins and 2876 must-link constraints with 997 proteins are extracted for DIP and HPRD, respectively.

4.4.2 Detected modules: One challenge is how to determine the amount of modules, k, because of there is no prior knowledge about the number of modules in real PPI network. The NMF-based methods usually assign community membership according to the real value of the row of matrix $\boldsymbol{F}$ for each node, if there is no value bigger than a given threshold ω for a specific column of matrix $\boldsymbol{F}$ and then the corresponding module of this column will be omitted. Therefore, we can fit the proposed model PCNMTF with a larger value of k as it is able to identify the amount of modules adaptively. We set the value of k equal to 500 and 1000 for DIP and HPRD, respectively, in this paper. In this work, the detected modules with size smaller than 2 are filtered out. The compared results of all methods used in this paper are reported in Table 4 where coverage is the number of detected proteins, *#as*, #ad and *#ai* indicate the average size, average degree and average interactions in the detected modules, *#m* is the number of detected modules, *#mm* is the number of modules matched with known complexes, *#ai_ma* is the average interactions of matched modules and *#ai_ml* denotes the average interactions in matched modules but not in must-link constraints. To evaluate the performance of PCNMTF on detecting functional modules, we first compared the detected modules with known complexes and then we conducted enrichment analysis to evaluate the functional homogeneity of detected modules.

The convergence of our proposed model PCNMTF was investigated, the values of objective function (6) with respect to the number of iterations is plotted in Fig. 3. Then we can see that our proposed model NMTF can get a local optimal value after some iterations.

4.4.3 Protein complexes: Although the priori information is extracted from complex database CORUM, the amount of proteins contained in priori information is less than the number of proteins in CORUM. Since only part of proteins in CORUM is used as priori information, we need to compare the detected modules with complexes in CORUM. Another well-known human-related protein complex database named PCDq is used to as gold standard also. The precision, recall and F-measure of all compared algorithms on both two PPI networks are showed in Figs. 4*a* and *b* which using CORUM as ground truth and Figs. 4*c* and *d* which using PCDq as ground truth, then we can find that the proposed algorithm PCNMTF outperforms other compared methods by means of all these three metrics except for MCODE and K-rank-d on precision. That is because they have detected fewer modules and proteins compared with PCNMTF (Table 4). Incorporating prior must-link information into models can significantly improve the ability of detecting functional modules efficiently. The results on real human-related PPI networks indicate that the proposed PCNMTF model offers a more effective way to discover considerable protein functional modules in PPI networks.

4.4.4 Enrichment analysis of detected modules: In order to explore the biological significance of the protein modules which are not considered in known protein complex databases, we conducted the enrichment analysis for all detected modules in terms of gene ontology (GO) annotations which contain three categories: Biological process (BP), cellular component (CC) and molecular function (MF). The extend of enrichment for each module is measured by p-value that can be obtained by hypergeometric test [53]. The functional homogeneity can be evaluated by p-value. For a specific GO function, a smaller p-value always indicates that the module has a more significance biological meaning to this function. Then, the proportion of modules with p-values less than a given threshold was calculated for all computational methods. The threshold was set from 10^{-10} to 0.01, and then, for a specific threshold, the higher percentage of modules in the interval means the more effective of detecting functional modules from PPI networks for an algorithm. Fig. 5 presents the distribution of proportion of modules in different intervals of p-value on DIP and HPRD in terms of BP, CC and MF and we can see that the PCNMTF performs better than the compared methods on both DIP and HPRD networks. Thus, the proposed model PCNMTF can be used to detect more homogeneous functional modules from PPI networks. In order to show what modules of

Table 4 Information of modules detected by all compared methods on DIP and HRPD

Network	Method	Coverage	*#as*	#ad	*#ai*	*#m*	CORUM			PCDq		
							#mm	*#ai_ma*	*#ai_ml*	*#mm*	*#ai_ma*	*#ai_ml*
DIP	MCODE	421	5.19	5.82	7.71	81	47	11.93	10.80	49	8.68	8.11
	K-rank-D	1666	12.34	2.73	11.25	135	60	10.01	9.16	65	9.22	8.65
	NMF	2679	11.26	3.69	3.73	255	102	2.80	1.61	109	2.21	1.81
	PCNMF	2748	7.33	3.46	2.81	375	178	1.37	2.11	164	2.07	1.93
	SNMF	1873	7.28	3.80	2.09	294	148	3.70	2.30	111	2.25	1.95
	PCSNMF	2876	6.21	3.48	5.93	463	229	6.20	5.63	215	6.05	5.72
	NMTF	2766	8.26	3.59	2.64	335	155	1.63	1.40	138	1.42	1.26
	NMTFADJ	2701	9.72	3.63	3.26	278	138	2.50	1.65	95	1.61	1.36
	NMTFJAC	2874	9.97	3.12	4.38	223	96	1.94	1.38	83	1.60	1.13
	PCNMTF	2920	10.50	3.52	9.53	278	137	11.20	9.50	133	8.96	8.44
HPRD	MCODE	1161	11.38	16.58	18.77	102	37	27.25	25.68	48	12.49	11.45
	K-rank-D	5316	33.22	2.90	55.12	160	33	9.00	7.91	54	3.85	3.37
	NMF	9178	11.125	8.13	2.94	825	241	3.97	3.80	300	1.52	1.44
	PCNMF	9055	12.63	8.77	3.88	888	277	6.32	5.36	259	4.04	3.57
	SNMF	9392	9.77	7.93	10.03	961	315	12.37	11.58	464	10.14	9.74
	PCSNMF	9159	22.85	9.63	6.19	954	257	11.00	9.71	216	10.32	9.45
	NMTF	9266	10.54	7.00	7.34	879	284	4.58	3.89	286	2.64	2.24
	NMTFADJ	9243	10.66	9.09	1.55	867	247	10.85	8.98	234	1.60	1.27
	NMTFJAC	9239	10.64	7.32	5.73	868	214	13.30	10.11	326	2.78	2.12
	PCNMTF	9337	9.52	8.78	10.27	991	391	12.11	11.14	525	11.18	10.77

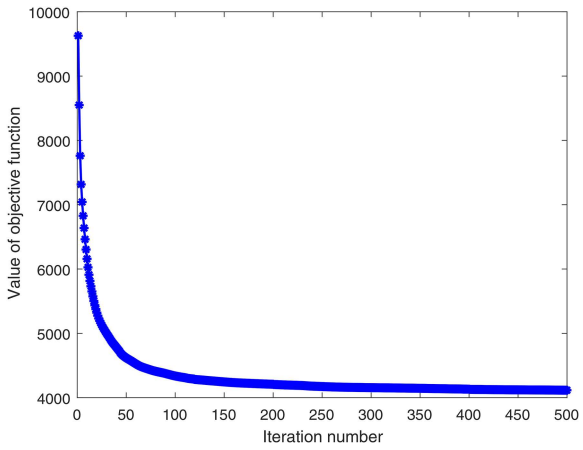


Fig. 3 *Values of (6) with respect to various iteration numbers on DIP network*

human-related PPI networks were detected, we list the top 5 significant modules in terms of BP in Tables 5 and 6 separately.

4.5 Parameter analysis

There are two parameters α and β which can affect the performance of our proposed model PCNMTF. In order to make it clear how these two parameters work, we apply PCNMTF on GN8 networks with changing the value of α and β at the same time and we illustrate the distribution of NMI in terms of different parameter values. We present the influence of these two parameters in Fig. 6*a*. We vary the value of α from 10^{-8} to 10^{2} and β from 0.1–1000. With the same setting, we evaluate the influence in terms of f-measure for DIP and the distribution is showed in Fig. 6*b*. Then, we observe that proposed PCNMTF performs better when α in the vicinity of 0.05 and β bigger than 10. The two parameters have the same influence on LFR and HPRD. The presented results are averaged over 50 repeated experiments.

5 Conclusion

In this manuscript, we propose a novel semi-supervised model PCNMTF to detect overlapping protein functional modules from human PPI networks. The proposed model, PCNMTF, makes better use of topological property of PPI networks and human-curated protein complex sufficiently. The experiments are executed on both synthetic networks and real-world human-related PPI networks, DIP and HPRD, and PCNMTF shows superior performance on finding functional modules although we incorporate very limited must-links which are extracted from CORUM. Our future work would consider how to incorporate other biological function of proteins, such as gene expressions and GO functional annotations, to obtain high-quality functional modules from human PPI networks.

6 Acknowledgments

This work was supported by the National Science Foundation of China (grant nos. 61105055, 81230086), National Basic Research Program of China (grant no. 2014CB542903), National Key Technology R&D Program (grant nos. 2013BAI02B01, 2013BAI13B04), Special Programs of Traditional Chinese Medicine (grant nos. 201407001, JDZX2015171, JDZX2015170), the Fundamental Research Funds for the Central Universities (grant no. 2017JBM020), National Keypoint Research and Invention Program (grant no. SQ2017YFC170370).

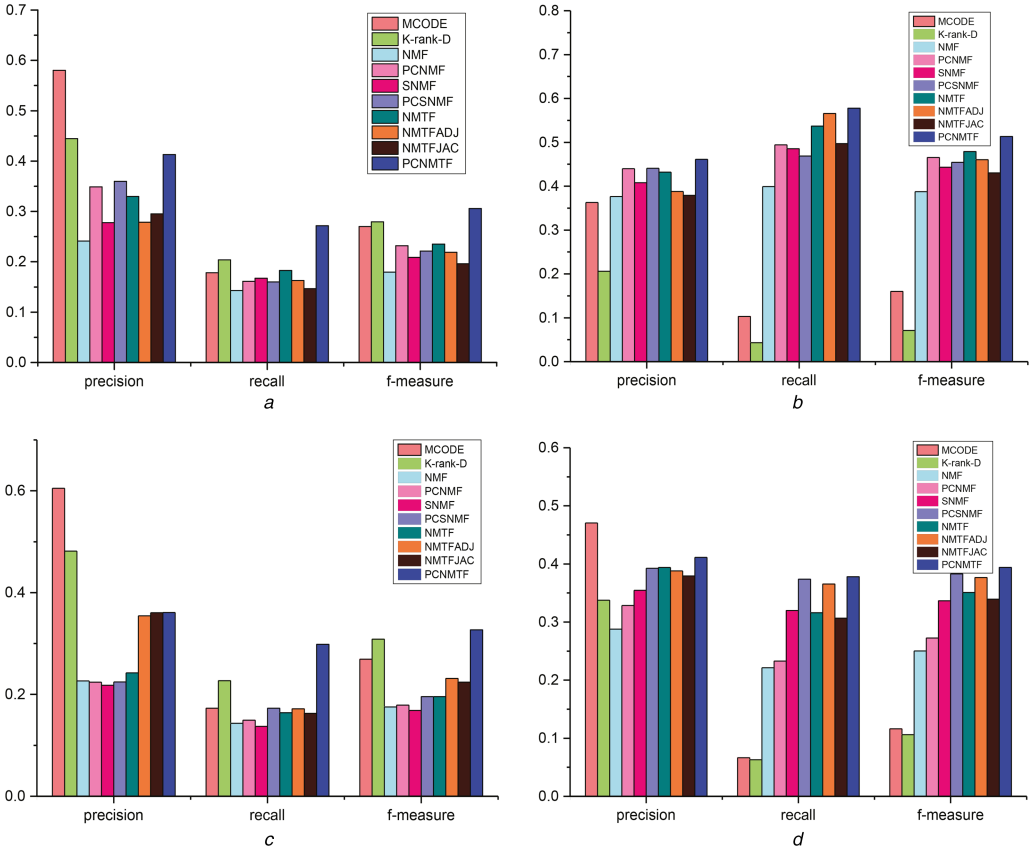


Fig. 4 *Precision, recall and F-measure of compared methods on DIP and HPRD*

(a), ***(b)*** Take CORUM as ground-truth, ***(c)***, ***(d)*** Take PCDq as ground-truth. ('ml' means must-link and 'gs' means gold standard database)

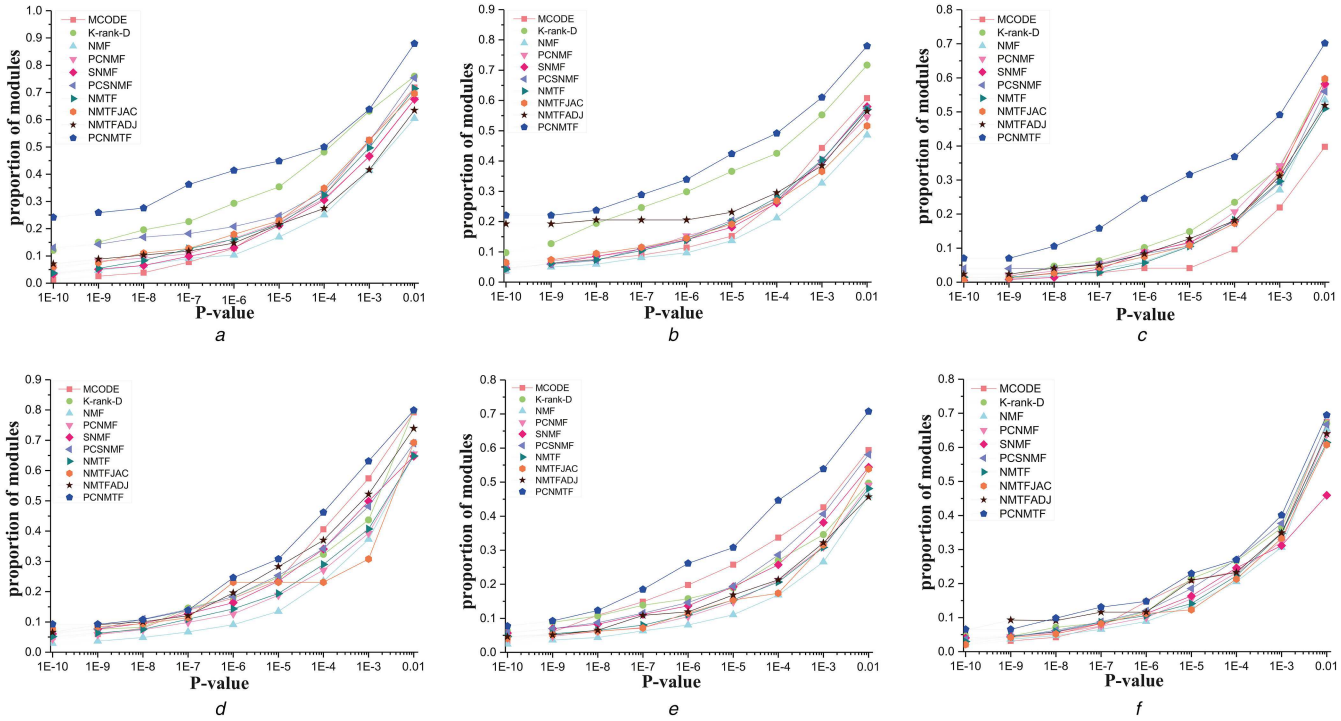


Fig. 5 *Enrichment analysis of all methods*

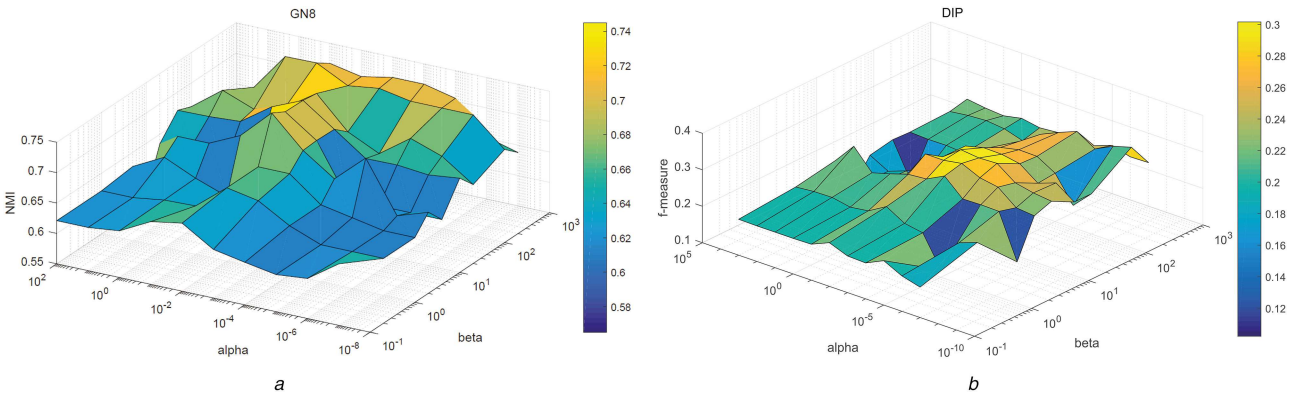
(a), ***(b)***, ***(c)*** are the proportion of enriched modules from DIP with different *p*-value in terms of BP, CC and MF respectively, ***(d)***, ***(e)***, ***(f)*** are the proportion of enriched modules from HPRD with different *p*-value in terms of BP, CC and MF respectively

Table 5 Top 5 modules enrich on BP terms from DIP

Module ID	Size	Members	*p*-Value	GO ID	GO term
256	12	ANAPC2 FZR1 MAD2L1 RAE1 CDC23 FBXO5 BUB1B ANAPC10 CDC20 ANAPC7 PTTG1 CDC16	4.29E−17	GO:0031145	transcription initiation from RNA polymerase II promoter
102	14	EIF3C EIF3D EIF3A EIF3B EIF3F EIF3G EIF3H EIF1AX EIF3E EIF3M EIF3K EIF3L EIF1 EIF3I EIF3J	3.08E−15	GO:0006413	translational initiation
458	13	TIFA TAB2 PSMB5 UBE2N PSMD13 PSMD12 PSMC3 PSMC2 PSMD1 IL1RAP PSMD3 PSMD6 PSMD7	4.75E−15	GO:0006521	regulation of cellular amino acid metabolic process
335	11	FGF6 FGF5 FGF8 FGF7 FGF9 FGF10 MMP14 FGF1 FGF2 FGF3 FGF4	1.41E−14	GO:0051781	positive regulation of cell division
218	11	NFE4 PPM1G WRAP53 SNUPN SNRPD3 SNRPD2 DDX20 SNRPF COIL SNRPE SMN1	7.54E−13	GO:0034660	ncRNA metabolic process

Table 6 Top 5 modules enrich on BP terms from HPRD

Module ID	Size	Members	*p*-Value	GO ID	GO term
8	13	ABCF1 PDK1 PDK2 WFS1 PDK3 PDK4 DLAT PDHB ACAP3 PDHA2 C4orf27 PDHA1 PDHX	7.05E−21	GO:0010510	regulation of acetyl-CoA biosynthetic process from pyruvate
97	19	GJA8 CLDN16 CLDN8 CLDN7 CLDN3 CLDN6 CLDN5 GJB3 GJA3 ARVCF KIRREL CGN TJP3 JAM2 JAM3 TJP2 CLDN4 GJC1 CLDN2	2.99E−19	GO:0016338	calcium-independent cell–cell adhesion
460	16	CCL2 CXCL9 CCL19 CCL8 CCL28 CCL7 CCL27 CXCL10 CCL5 CCL24 CCL25 CCL13 CXCL11 CCL11 XCL2 CCL21	9.65E−19	GO:0006955	immune response
45	14	PRKAG3 PFKFB2 PRKAG1 PRKAG2 PRKAB2 PRKAB1 NHLRC1 EEF2K PRKAA1 CAB39 AGL GCKR NDUFA7 ACACB	1.62E-18	GO:0046320	regulation of fatty acid oxidation
395	18	FLRT3 FGF19 FGF6 CCDC17 FGF8 FGF7 FGF17 SMG7 IL17RD HBZ C6orf47 FGF1 FGF3 FGF18 FGF5 FGF10 FGF23 FGF4	1.66E−18	GO:0008286	insulin receptor signalling pathway



a *b*

Fig. 6 *Influence of α and β*
(a) On GN8 network, ***(b)*** On DIP network

7 References

[1] Davis, D., Yaveroğlu, Ö.N., Malod-Dognin, N., *et al.*: 'Topology-function conservation in protein–protein interaction networks', *Bioinformatics*, 2015, **31**, (10), pp. 1632–1639

[2] Aebersold, R., Mann, M.: 'Mass spectrometry-based proteomics', *Nature*, 2003, **422**, (6928), pp. 198–207

[3] Ho, Y., Gruhler, A., Heilbut, A., *et al.*: 'Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry', *Nature*, 2002, **415**, (6868), pp. 180–183

[4] Ito, T., Chiba, T., Ozawa, R., *et al.*: 'A comprehensive two-hybrid analysis to explore the yeast protein interactome', *Proc. Natl. Acad. Sci.*, 2001, **98**, (8), pp. 4569–4574

[5] Uetz, P., Giot, L., Cagney, G., *et al.*: 'A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*', *Nature*, 2000, **403**, (6770), pp. 623–627

[6] Wu, H., Gao, L., Dong, J., *et al.*: 'Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks', *PLoS ONE*, 2014, **9**, (3), p. e91856

[7] Pereira-Leal, J.B., Levy, E.D., Teichmann, S.A.: 'The origins and evolution of functional modules: lessons from protein complexes', *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 2006, **361**, (1467), pp. 507–517

[8] Albert, R., Jeong, H., Barabási, A.L.: 'Error and attack tolerance of complex networks', *Nature*, 2000, **406**, (6794), pp. 378–382

[9] Wagner, G.P., Pavlicev, M., Cheverud, J.M.: 'The road to modularity', *Nat. Rev. Genetics*, 2007, **8**, (12), pp. 921–931

[10] Shih, Y.K., Parthasarathy, S.: 'Identifying functional modules in interaction networks through overlapping Markov clustering', *Bioinformatics*, 2012, **28**, (18), pp. i473–i479

[11] Lei, X., Wu, S., Ge, L., *et al.*: 'Clustering and overlapping modules detection in PPI network based on IBFO', *Proteomics*, 2013, **13**, (2), pp. 278–290

[12] Nepusz, T., Yu, H., Paccanaro, A.: 'Detecting overlapping protein complexes in protein-protein interaction networks', *Nat. Meth.*, 2012, **9**, (5), pp. 471–472

[13] Kenley, E.C., Cho, Y.R.: 'Detecting protein complexes and functional modules from protein interaction networks: a graph entropy approach', *Proteomics*, 2011, **11**, (19), pp. 3835–3844

[14] Arnau, V., Mars, S., Marín, I.: 'Iterative cluster analysis of protein interaction data', *Bioinformatics*, 2004, **21**, (3), pp. 364–378

[15] Adamcsek, B., Palla, G., Farkas, I.J., *et al.*: 'CFinder: locating cliques and overlapping modules in biological networks', *Bioinformatics*, 2006, **22**, (8), pp. 1021–1023

[16] Palla, G., Derényi, I., Farkas, I., *et al.*: 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, 2005, **435**, (7043), pp. 814–818

[17] Xiang, Y., Zhang, C.Q., Huang, K.: 'Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data', *BMC Bioinform.*. BioMed Central, 2012, **13**, (S2), p. S12

[18] Bader, G.D., Hogue, C.W.V.: 'An automated method for finding molecular complexes in large protein interaction networks', *BMC Bioinform.*, 2003, **4**, (1), p. 2

[19] Li, T., Zhang, Y., Sindhwani, V.: 'A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge'. Proc. Joint Conf. 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. Natural Language Processing of the AFNLP, 2009, vol. **1**, pp. 244–252

[20] Devarajan, K.: 'Nonnegative matrix factorization: an analytical and interpretive tool in computational biology', *PLoS Comput. Biol.*, 2008, **4**, (7), p. e1000029

[21] Wang, H., Nie, F., Huang, H., *et al.*: 'Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation'. 2011 IEEE 11th Int. Conf. Data Mining (ICDM), 2011, pp. 774–783

[22] Zhu, L., Galstyan, A., Cheng, J., *et al.*: 'Tripartite graph clustering for dynamic sentiment analysis on social media'. Proc. 2014 ACM SIGMOD international Conf. Management of data, 2014, pp. 1531–1542

[23] Pei, Y., Chakraborty, N., Sycara, K.: 'Nonnegative matrix tri-factorization with graph regularization for community detection in social networks'. Twenty-Fourth Int. Joint Conf. Artificial Intelligence. 2015

[24] Menche, J., Sharma, A., Kitsak, M., *et al.*: 'Uncovering disease-disease relationships through the incomplete interactome', *Science*, 2015, **347**, (6224), p. 1257601

[25] Von Mering, C., Krause, R., Snel, B., *et al.*: 'Comparative assessment of large-scale data sets of protein–protein interactions', *Nature*, 2002, **417**, (6887), pp. 399–403

[26] Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., *et al.*: 'CORUM: the comprehensive resource of mammalian protein complexes', *Nucleic Acids Res.*, 2007, **36**, (Suppl. 1), pp. D646–D650

[27] Psorakis, I., Roberts, S., Ebden, M., *et al.*: 'Overlapping community detection using Bayesian non-negative matrix factorization', *Phys. Rev. E*, 2011, **83**, (6), p. 066114

[28] Wang, F., Li, T., Wang, X., *et al.*: 'Community discovery using nonnegative matrix factorization', *Data Min. Knowl. Discov.*, 2011, **22**, (3), pp. 493–521

[29] Lu, H., Zhu, X., Liu, H., *et al.*: 'The interactome as a tree—an attempt to visualize the protein–protein interaction network in yeast', *Nucleic Acids Res.*, 2004, **32**, (16), pp. 4804–4811

[30] Wang, R.S., Zhang, S., Wang, Y., *et al.*: 'Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures', *Neurocomputing*, 2008, **72**, (1), pp. 134–141

[31] Kondor, R.I., Lafferty, J.: 'Diffusion kernels on graphs and other discrete input spaces', *ICML*, 2002, **2**, pp. 315–322

[32] Zhang, Z.-Y., Sun, K.-D., Wang, S.-Q.: 'Enhanced community structure detection in complex networks with partial background information', *Sci. Rep.*, 2013, **3**, pp. 3241

[33] Yang, L., Cao, X., Jin, D., *et al.*: 'A unified semi-supervised community detection framework using latent space graph regularization', *IEEE Trans. Cybern.*, 2015, **45**, (11), pp. 2585–2598

[34] Wass, M.N., David, A., Sternberg, M.J.E.: 'Challenges for the prediction of macromolecular interactions', *Curr. Opin. Struct. Biol.*, 2011, **21**, (3), pp. 382–390

[35] Zhu, S., Yu, K., Chi, Y., *et al.*: 'Combining content and link for classification using matrix factorization'. Proc. 30th annual Int. ACM SIGIR conf. Research and development in Information Retrieval, 2007, pp. 487–494

[36] Wu, Q., Wang, Z., Li, C., *et al.*: 'Protein functional properties prediction in sparsely-label PPI networks through regularized non-negative matrix factorization', *BMC Syst. Biol.*, BioMed Central Ltd, 2015, **9**, (Suppl. 1), p. S9

[37] Zhang, Y., Du, N., Ge, L., *et al.*: 'A collective NMF method for detecting protein functional module from multiple data sources'. Proc. ACM Conf. Bioinformatics, Computational Biology and Biomedicine, 2012, pp. 655–660

[38] Brunet, J.P., Tamayo, P., Golub, T.R., *et al.*: 'Metagenes and molecular pattern discovery using matrix factorization', *Proc. Natl. Acad. Sci.*, 2004, **101**, (12), pp. 4164–4169

[39] Hartwell, L.H., Hopfield, J.J., Leibler, S., *et al.*: 'From molecular to modular cell biology', *Nature*, 1999, **402**, pp. C47–C52

[40] Zhang, X.F., Dai, D.Q., Le, O.-Y., *et al.*: 'Detecting overlapping protein complexes based on a generative model with functional and topological properties', *BMC Bioinform.*, 2014, **15**, (1), p. 186

[41] Girvan, M., Newman, M.E.J.: 'Community structure in social and biological networks', *Proc. Natl. Acad. Sci.*, 2002, **99**, (12), pp. 7821–7826

[42] Lancichinetti, A., Fortunato, S., Radicchi, F.: 'Benchmark graphs for testing community detection algorithms', *Phys. Rev. E*, 2008, **78**, (4), p. 046110

[43] Xenarios, I., Salwinski, L., Duan, X.J., *et al.*: 'DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions', *Nucleic Acids Res.*, 2002, **30**, (1), pp. 303–305

[44] Peri, S., Navarro, J.D., Kristiansen, T.Z., *et al.*: 'Human protein reference database as a discovery resource for proteomics', *Nucleic Acids Res.*, 2004, **32**, (Suppl. 1), pp. D497–D501

[45] Kikugawa, S., Nishikata, K., Murakami, K., *et al.*: 'PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset', *BMC Syst. Biol.*, 2012, **6**, (Suppl. 2), p. S7

[46] Radicchi, F., Castellano, C., Cecconi, F., *et al.*: 'Defining and identifying communities in networks', *Proc. Natl. Acad. Sci. USA*, 2004, **101**, (9), pp. 2658–2663

[47] Lovász, L., Plummer, M.D.: '*Matching theory*' (American Mathematical Society, Providence, 2009)

[48] Ou-Yang, L., Dai, D.Q., Zhang, X.F.: 'Protein complex detection via weighted ensemble clustering based on Bayesian nonnegative matrix factorization', *PLoS ONE*, 2013, **8**, (5), p. e62158

[49] Zhang, Y., Yeung, D.Y.: 'Overlapping community detection via bounded nonnegative matrix tri-factorization'. Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2012, pp. 606–614

[50] Li, Y., Jia, C., Yu, J.: 'A parameter-free community detection method based on centrality and dispersion of nodes in complex networks', *Phys. A Stat. Mech. Appl.*, 2015, **438**, pp. 321–334

[51] Shi, X., Lu, H., He, Y., *et al.*: 'Community detection in social network with pairwisely constrained symmetric non-negative matrix factorization'. Proc. 2015 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining 2015, 2015, pp. 541–546

[52] Wang, D., Gao, X., Wang, X.: 'Semi-supervised nonnegative matrix factorization via constraint propagation', *IEEE Trans. Cybern.*, 2016, **46**, (1), pp. 233–244

[53] Bu, D., Zhao, Y., Cai, L., *et al.*: '"Topological structure analysis of the protein–protein interaction network in budding yeast', *Nucleic Acids Res.*, 2003, **31**, (9), pp. 2443–2450