Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

Mathematical Biosciences and Engineering, 16(4): 1933–1948
DOI: 10.3934/mbe.2019094
Received: 17 December 2018
Accepted: 15 February 2019
Published: 07 March 2019

Research article

SNFM: A semi-supervised NMF algorithm for detecting biological functional modules

Yutong Man[1†], Guangming Liu[2†], Kuo Yang[1] and Xuezhong Zhou[1,3,*]

[1] Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[2] School of Computer Science & Engineering, Xi'an University of Technology, Xi'an 710048, China
[3] Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, China

* **Correspondence:** Email: xzzhou@bjtu.edu.cn; Tel: 86-10-51684931.

Abstract: Unraveling protein functional modules from protein-protein interaction networks is a crucial step to better understand cellular mechanisms. In the past decades, numerous algorithms have been proposed to identify potential protein functional modules or complexes from protein-protein interaction (PPI) networks. Unfortunately, the number of PPIs is rather limited, and some interactions are false positive. Therefore, the algorithms that only utilize PPI networks may not obtain the expected results related to functional modules. In this study, we propose a novel semi-supervised functional module detection method based on non-negative matrix factorization(NMF)(SNFM), which incorporate high-quality supervised PPI links from complexes as prior information.Our method outperforms all the other competitors with improvements on performance by around 15.4% in Precision, 28.9% in Recall, 27.1% in F-score (on DIP data set) by using PCDq as gold standards.

Keywords: PPI; NMF; semi-supervised; functional modules; DIP

1. Introduction

Proteins in cells seldom perform biological functions alone but form a larger molecular component with each other to exert specific functions [1, 2] . With the development of high-throughput technologies such as mass spectrometry [3, 4] and two-hybrid systems [5, 6], numerous interactions among proteins have been acquired, resulting in large-scale protein-protein interaction (PPI) networks. Identifying the underlying protein functional modules from PPI networks is important to explore the biological process [7] in cells and to elucidate the disease-causing mechanisms [8] . Recent studies suggest that proteins densely interacting with each other are more likely to perform similar or the same

biological functions [9, 10], and various functional modules or protein complex detection algorithms have been proposed [11–13].

Typically, the PPI networks are modeled as graphs in which the proteins are the nodes and the interactions between the proteins are the edges [14]. A functional module can be regarded as sub-graphs with the nodes connected to each other more densely than others in the rest of the graph [15, 16] with specific molecular functions. Thus, the goal of detecting functional modules from PPI networks can be determined using graph clustering method. For example, the cliques (fully linked sub-graphs) in a graph have been used for developing functional module detection algorithms, such as MC [17], LCMA [18], and CFinder [19]. Other proposed particular methods include core-attachment based methods, such as COACH [20]. However, these previous methods are only relied on the topological structure of the PPI networks. In addition, the current available protein interactions are rather limited and prone to be specious [21]. Therefore, to improve the performance, Zhang et al. [22] proposed a generative model that takes the topological structure and Gene Ontology (GO) information together to detect functional modules. Georgii et al. [23] used gene expression data or phenotype data to build a weighted PPI network and then analyzed the functional modules in the PPI networks. Zhang et al. [24] developed an NMF-based functional module detection method named CoNMF that employs gene expression data and topological information simultaneously. To some extent, these methods reduce the drawback of those solely relied on the PPI network data, however, incorporating multiple data points would also induce more noises.

The databases on high-quality protein complexes, such as CORUM and PCDq [25], can be used for detect functional modules. Qi et al. [26] propose a semi-supervised functional module detection method named SCI-BN that extracts topological and biological features from known complexes and then uses them to train a Bayesian network for each sub-graph. Yu et al. [2] utilized the known protein complexes and the PPI network together to extract features and then used clique based algorithms to detect the functional modules. However, these two supervised approaches extract features from known complexes, which make them difficult to discover unknown functional modules.

To tackle these challenges, in this study, we propose a novel semi-supervised functional module detection algorithm(SNFM). Instead of extracting features from known complexes, we treated the protein pairs in one common complex as must-link constraints that can be viewed as prior information for our method. SNFM includes two main steps. The first step is to construct a similarity matrix of the proteins according to the adjacency matrix, in which we use a self-similarity manner to construct the similarity matrix as the featured matrix for the proteins. The second step is to factorize the similarity matrix instead of the adjacency matrix to obtain the protein module indicator matrix. Meanwhile, the protein pairs with must-link constraints are formulated as a graph regularization term. We will depict main components of the algorithm in detail in the Methods section.

2. Methods

We first use the sparse subspace clustering method to construct the similarity matrix according to the protein adjacency matrix and next take the constructed similarity matrix as the protein feature matrix needed for the second step. The similarity matrix is constructed by taking into account the global information of the entire PPI network. Then, the similarity matrix is factorized into two non-negative matrices, and we propose an objective function based on NMF that takes protein pairs from complexes

as must-link constraints. The similarity matrix obtained in the first step is decomposed to obtain the protein module indicator matrix, and the protein pairs of must-link constraints are formulated as a graph regularization term. We aim to improve the accuracy of the algorithm detecting the functional modules of protein interaction networks. The main steps of the proposed model are described in Algorithm 1.

2.1. Subspace clustering

Given the PPI network G, G can be expressed as an adjacency matrix A = $(x_1, ..., x_N) \in \mathbb{R}^{n\times n}$. The adjacency matrix is too sparse due to the limited protein-protein interactions. More importantly, the adjacency matrix only considers the relationship between the proteins and their neighbors and ignores the other topological information such as proteins with common neighbors that do not interact with each other. Each protein in the PPI network can be expressed as a linear combination of other proteins. Additionally, these n samples are from k subspaces S_i , i = 1,...,k. The basic idea of sparse subspace clustering [27] is that a protein x_i, x_i is represented as a linear combination of other proteins in the same subspace. Its corresponding objective function is as follows:

$$\begin{aligned} \min_{Z} J_1(Z) &= \frac{1}{2}\|A - AZ\|_F^2 \\ s.t. \quad & Z_{ii} = 0 \end{aligned} \tag{2.1}$$

where Z= $[z_1, z_2, ..., z_n]$, Z$\in \mathbb{R}^{n\times n}$ is the coefficient matrix. $Z_{ii} = 0$ is to avoid a protein expressed by itself. The larger the coefficient Z_{ij} means the sample x_i is more similarity to sample x_j; therefore, the obtained similarity matrix has superior subspace structure and robustness. However, in the specific application, there are noise data, or the sample is not clean, where E $\in \mathbb{R}^{N\times N}$ is the error matrix.

2.2. Supervised non-negative matrix factorization

Non-negative matrix factorization (NMF) [28] is a widely used matrix decomposition algorithm that can be used to detect overlapping protein modules in unsupervised manner. NMF decomposes the original matrix into two non-negative low-rank matrices, the product of which is the approximation of the original matrix.

However, most unsupervised module detection algorithms only consider the topological information of the PPI network. Moreover, due to the limited amount of PPI data obtained from high-throughput biological experiments [29] and the false positive links in PPI data, it is difficult to develop accurate functional modules by solely using PPI data. Currently, we can obtain information on high-quality human protein complexes from databases, such as CORUM [30]. Although in small scale, human protein complexes data is a high-quality data source of interaction network, which can be used as prior information. Therefore, we can design a semi-supervised functional module detection algorithm that combines the PPI data and the protein complexes data to compensate for the shortcomings of the PPI data.

The network similarity matrix A$\in \mathbb{R}_+^{n\times n}$ is used as the original input matrix. The main purpose of the NMF algorithm is to decompose the similarity matrix into the product of two non-negative low-rank matrices W $\in \mathbb{R}_+^{n\times k}$ and H$\in \mathbb{R}_+^{k\times n}$, where k$\ll n$. In this study, we use the Euclidean distance to measure the distance between Z and WH. Yang et al. [31] proposed a supervised community detection framework based on NMF and symmetric NMF. The objective function of the supervised model based on the NMF is as follows:

$$\min_{W\geq 0,H\geq 0} J_2(W,H) = \|A - WH\|_F^2 + \beta Tr(HLH^T) \tag{2.2}$$

The matrix L is the Laplacian matrix, L = D - M. M is the must-link constraint matrix, and D is its corresponding diagonal matrix. β is a positive parameter used to adjust the intensity of the prior information.

The column vector of the original matrix A is the weighted sum of all the column vectors in the left matrix W, and the weighted matrix is the primary color of the column vector corresponding to the right matrix H; therefore, W is called the base matrix, and H is the coefficient matrix. The original matrix can be replaced by a coefficient matrix, and the features of the original matrix can be characterized and dimensionality reduced.

2.3. Semi-supervised NMF Functional Module

In the original semi-supervised NMF model, the adjacency matrix only considers the information from the proteins and their neighboring proteins. Here, for SNFM, we use the similarity matrix constructed by sparse subspace clustering that takes into account the higher order neighbor information of the proteins, such as the protein neighbor's neighbor. We first use the subspace clustering method to construct the similarity matrix based on the adjacency matrix. Then, we performed the semi-supervised NMF algorithm with protein pairs from high-quality protein complexes as must links to improve the quality of the detected functional modules from the PPI network. Next we describe the algorithm of SNFM as follows:

$$\begin{gathered}\min_{W\geq 0,H\geq 0} J_2(W,H) = \|Z - WH\|_F^2 + \beta Tr(HLH^T) \\ s.t. \quad Z_{ii} = 0\end{gathered} \tag{2.3}$$

where $Z = [z_1, z_2, ..., z_n]$, $Z \in \mathbb{R}^{n\times n}$ is the coefficient matrix.

2.4. Optimization

2.4.1. Similarity matrix of PPI network

Subspace clustering is used to build the similarity of the PPI network that can be formulated as follows:

$$\begin{aligned}\min_Z J_1(Z) &= \frac{1}{2}\|A - AZ\|_F^2 \\ &= \frac{1}{2}Tr(AA^T - AZ^T - AZA^T + AZZ^TA^T)\end{aligned} \tag{2.4}$$

where A is the adjacency matrix of a PPI network and is a constant, and Z is the only variable. We can obtain the derivative of Z as follows:

$$\nabla_Z J_1(Z) = A^TAZ - A^TA \tag{2.5}$$

Then using the KKT condition, we get the multiplicative iterative formula for Z as follows:

$$Z = Z \odot \frac{A^T}{A^TAZ} \tag{2.6}$$

$\odot$ represents the product of the two matrix corresponding elements.

2.4.2. Semi-supervised NMF

To minimize Eq. (2.3), we need to use the following matrix trace related knowledge: Tr (WH)=Tr (HW), $\|A\|_F^2 = Tr(AA^T)$, so Eq. (2.3) can be rewritten as follows:

$$\min_{W\geq 0, H\geq 0} J_2(W,H) = Tr[(Z-WH)(Z-WH)^T] + \beta Tr(HLH^T) \tag{2.7}$$

Since W and H are both variables, Eq. (2.3) is not convex; therefore, to find the optimal W and H that can make Eq. (2.3) reach the minimum, we update one while keeping the other fixed as a constant. If we fix H as a constant, we can obtain a partial derivative of the objective function about W. The derivation process is reported as follows:

$$\nabla_W J_2(W,H) = 2(WHH^T - ZH^T) \tag{2.8}$$

Similarly, if we fix W and treat H as a variable, the partial derivative of H can be calculated as follows:

$$\nabla_H J_2(W,H) = 2(W^T WH - W^T Z + \beta HL) \tag{2.9}$$

Next, using the KKT condition, we obtain the updated rules of W and H as follows:

$$W = W \odot \frac{ZH^T}{WHH^T} \tag{2.10}$$

$$H = H \odot \frac{W^T Z + \beta(HM)}{W^T WH + \beta(HD)} \tag{2.11}$$

where D $\in \mathbb{R}_+^{n\times n}$ is a diagonal matrix of M, L = D - M is the Laplacian matrix of the matrix M, and $Tr(\cdot)$ is the trace of a matrix.

Matrices W and H are updated by the iterative iteration until the Eq. (2.3) converges or reaches the maximum number of iterations sets in advance. The time complexity is shown in Table 1.

Table 1. Calculate the number of iterations in the SNFM model.

	F-norm formulation			
	fladd	flmlt	fldiv	overall
Eq. (2.6)	$3n^3$	$3n^3+n^2$	n^2	n^3
Eq. (2.10)	$2k^2n+n^2k$	$2k^2n+n^2k$+nk	nk	n^2k
Eq. (2.11)	$3n^2k+2k^2n+$ 2nk	$3n^2k+2k^2n$+nk	nk	n^2k

3. Results

To validate the performance of SNFM, we conducted experiments on three human-related PPI networks.

Algorithm 1 The proposed SNFM model

Input: adjacency matrix A, must-link set L, number of modules K and β , the maximum number of iterations T_1, error ε_1, error ε_2, the maximum number of iterations T_2

Build the must-link constraint matrix $M \in \mathbb{R}_+^{n\times n}$ according to L
Initialize: W $\geq$ 0, H$\geq$ 0, Z
while $t < T_1$ **do**
    Update Z according to Eq. (2.6)
    Calculate the objective function value according to Eq. (2.1), denoted as J_1^t
    If the objective function value obtained by two iterations satisfies$\|J_1^t - J_1^{t-1}\| < \varepsilon_1$ or $t \geq T_1$, the algorithm ends. Otherwise, $t = t + 1$, and repeat steps 4 and 5
end while
while $t < T_2$ **do**
    Fix H, and update W according to Eq. (2.10)
    Fix W, and update H according to Eq. (2.11)
    Calculate the objective function value according to Eq. (6), denoted as J_2^t
    If the objective function value obtained by two iterations satisfies$\|J_2^t - J_2^{t-1}\| < \varepsilon_2$ or $t \geq T_2$, the algorithm ends. Otherwise, $t = t + 1$ and repeat steps 9, 10 and 11
end while

Output: Protein module membership matrix H

3.1. Data and experimental setup

3.1.1. Data sets

Three human-related PPI networks, namely the DIP, HPRD and STRING10, were used in our experiments. The DIP (Database of Interacting Proteins) [32] is a high-quality PPI database that integrates protein interactions from various data sources. It also contains interaction data calculated by algorithms from other reliable protein interaction databases with manually reviews. We extracted human-related PPI data from DIP, which includes a total of 4673 protein interactions and 2943 proteins. The HPRD (Human Protein Reference Database) [33] is a human-related database in which all of the information was manually extracted by biologists from the literatures. We extracted a total of 9453 proteins and 36888 protein interactions. The version of the STRING database [34] used in this study is STRING10, which is derived from several approaches, such as biological experiments, computational methods and literature curation. To obtain a high-quality PPI network, according to the introduction in [35], we only extracted the PPI data with a confidence score greater than 700 from the STRING10 database. Finally, we established a PPI network containing 14,380 proteins and 218,163 interactions.

In this experiment, we also used two human-related protein complex databases, CORUM and PCDq, as the validation data set. The CORUM (Comprehensive Resources for Mammalian Protein Complexes) [36] database provides data on protein complexes in human-reviewed mammalian organisms, all from the literature, excluding proteins obtained from high-throughput experimental data. We obtained the human related protein complexes from CORUM and removed the proteins that were not included in the human protein interaction data, so the number of protein complexes corresponding to different PPI data was different. PCDq is a human protein complex database. The protein complex data in this database is extracted from the integrated PPI network and then manually labeled for review and matched to complex data in the existing literature.

The corresponding CORUM protein complex in the DIP network contains 746 proteins, and PCDq contains 340 proteins. The corresponding CORUM protein complex in the HPRD network contains 1069 proteins, and PCDq contains 874 proteins. For STRING network, the corresponding CORUM protein complex contains 1134 proteins and PCDq contains 845 proteins.

3.1.2. Evaluation metrics

We used Precision, Recall, and F1 to measure the quality of the detected protein functional modules. First, we defined the degree of coincidence between a protein complex G_a and a detected protein functional module D_b as follows:

$$OL(G_a, D_b) = \frac{|G_a \cap D_b|^2}{|G_a| \times |D_b|} \tag{3.1}$$

where $|G_a|$ refers to the number of proteins in the protein complex G_a, $|D_b|$ refers to the number of proteins in the detected functional module D_b, and $|G_a \cap D_b|$ refers to the number of proteins appearing in both sets of G_a and D_b. The larger the OL value, the more similar the two contrast sets are. In this study, if $OL(m, n) \geq \gamma$, we believe that the two protein sets m and n match successfully. According to the introduction in [37,38], we set $\gamma = 0.25$ in this study. Next we defined the following two variables:

$$N_{ag} = |\{g|g \in G_a, \exists d \in D_b, OL(d, g) \geq 0.25\}| \tag{3.2}$$

$$N_{ad} = |\{d|d \in D_b, \exists g \in G_a, OL(g, d) \geq 0.25\}| \tag{3.3}$$

Subsequently, we defined the Precision, Recall, and F1 as follows:

$$Precision = \left|\frac{N_{ag}}{D_b}\right|, Recall = \left|\frac{N_{ad}}{G_a}\right| \tag{3.4}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.5}$$

where F1 is the harmonic mean of the precision and recall. The larger the indicator, the better the result of the detected functional module is.

For the human protein functional module, we first defined the neighborhood affinity score between the two protein sets as follows:

$$NA(p, g) = \frac{\left|N_p \cap N_g\right|^2}{\left|N_p\right| \left|N_g\right|} \tag{3.6}$$

This value is used to measure the degree of overlap of the module p detected from the PPI network with a known protein complex g. If the neighborhood affinity values of p and g satisfy $NA(p, g) \geq \omega$, then we believe that p and g match. In this study, we set ω= 0.25. Suppose $m = |P|, n = |G|$, t_{ij} represents the number of proteins that existed in the i_{th} detected protein functional module and the j_{th} protein complex simultaneously, and N_j represents the number of proteins in the j_{th} protein complex. Then the calculation formulas for Sn (Sensitivity) , PPV (Positive Predictive) , ACC (Accuracy) and MMR (Maximum Matching Ratio) are defined as follows:

$$Sn = \frac{\sum_{j=1}^{n} max_i \left\{t_{ij}\right\}}{\sum_{j=1}^{n} N_j}, PPV = \frac{\sum_{i=1}^{m} max_j \left\{t_{ij}\right\}}{\sum_{i=1}^{m} \sum_{j=1}^{n} t_{ij}} \tag{3.7}$$

$$ACC = \sqrt{Sn \times PPV}, MMR = \frac{\sum_{i=1}^{n} max_j NA(i,j)}{n} \tag{3.8}$$

The greater the ACC and MMR values, the greater the ability of the corresponding algorithm is to detect accurate functional modules from the PPI networks.

3.1.3. Parameter analysis

The SNFM has only one parameter: β, which is used to control the smoothness between the topological information and the prior information. To investigate the influence of various parameters β on performance, we set the value range of β to 1, 2, 5, 10, 20, 50, 100 and 1000. Taking the DIP database as an example, we used PCDq as the benchmark data set, and reported the changes of the corresponding ACC and MMR. As shown in Figure 1, we found that when the value of β is approximately 1000, our proposed SNFM algorithm achieved better results. Thus, in all of the experiments in this study, we set β =1000.

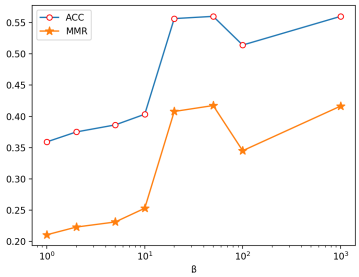


Figure 1. The ACC and MMR values corresponding to various β.

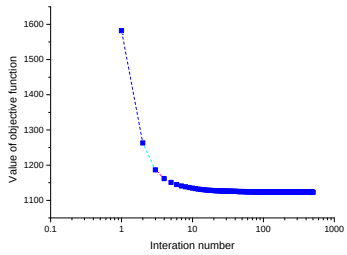


Figure 2. The objective function increases with the number of iterations.

3.1.4. Baseline algorithms

We chose the two state-of-the-art algorithms for functional module detection: DCU [39] and OH-PIN [40], and four classical clustering algorithms, namely Markov clustering algorithm (MCL) [41], By finding the maximal clique, [42] has proposed an algorithm named CMC that can remove or merge clusters according to the interconnectivity. The NMF algorithm [28] is also used to detect functional modules. The k-means [43] algorithm performs clustering by initializing the cluster center as baseline methods.

3.2. Experimental results

We obtained the related results for the six baseline methods and SNFM algorithm on three PPI networks. The corresponding Sn, PPV, ACC and MMR values of all comparison algorithms from three PPI networks (DIP,HPRD and STRING10) are shown in Table 2. For all comparative community detection algorithms, the algorithm proposed in this study SNFM includes a greater number of proteins in the functional modules and only slightly fewer than the DCU and k-means algorithm in the DIP

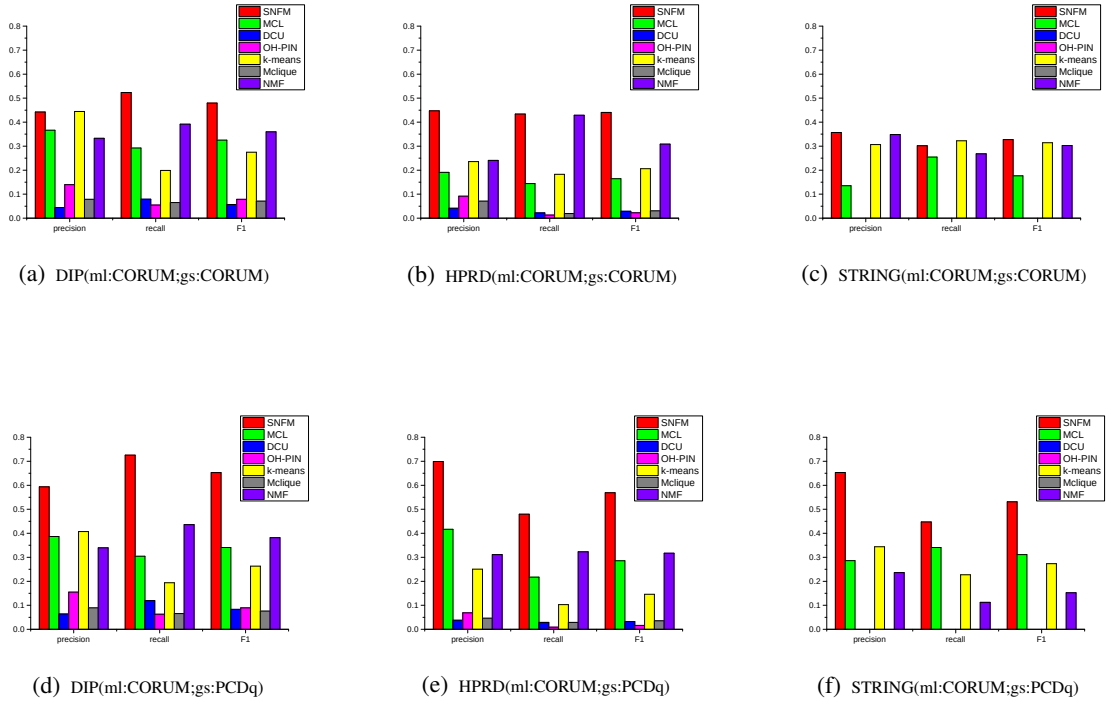
network compared with other algorithms. The optimal ACC and MMR values were obtained by SNFM in all three protein interaction networks, and the number of detected modules that are matched protein complexes achieve the best results in all of the algorithms. The DCU, OH-PIN, MCL, and Myclique algorithms do not use priori information, and NMF and k-means use priori information. SNFM adopts the must-link method, and the effect is strong. According to [31], the reliability of the prior information is affirmed, indicating that the prior information is well utilized.

Additionally, we also analyzed the convergence speed of the SNFM objective function. In Figure 2, we show the trend of the value of the objective function as the number of iterations increases. We found that after a certain number of iterations (approximately 100 times), the value of the objective function tends to be stable; that is, we obtain the local optimal solution of the objective function.

Table 2. Detailed results of compared algorithms on three human PPI networks using CORUM and PCDq as gold standards.The maximum value of each metric is shown in bold, and the next largest value is underlined. We used "—" to indicate the algorithms with no results after 24 hours. Among them, coverage represents the number of proteins contained in all of the modules detected by each algorithm. The table shows the number of modules detected by #m, the average size of the detected modules is marked by #as, and the number of modules detected that matched the known protein complex is marked by #mm.

Network	Method	Coverage	#as	#m	CORUM					PCDq				
					#mm	Sn	PPV	ACC	MMR	#mm	Sn	PPV	ACC	MMR
DIP	SNFM	2715	7.459	364	161	0.458	0.238	**0.330**	**0.293**	216	**0.620**	0.506	**0.560**	**0.417**
	MCL	2374	9.458	251	92	**0.603**	0.145	0.296	0.180	97	0.566	0.282	0.400	0.204
	DCU	2937	4.298	1143	51	0.049	0.207	0.100	0.044	73	0.054	0.397	0.147	0.052
	OH-PIN	984	9.109	129	18	0.031	0.307	0.098	0.029	20	0.032	0.485	0.124	0.027
	k-means	2740	20.296	135	60	0.542	0.123	0.258	0.137	55	0.560	0.188	0.325	0.131
	Mclique	985	3.280	751	142	0.031	**0.339**	0.102	0.033	145	0.032	0.447	0.120	0.031
	NMF	2114	4.950	427	59	0.286	0.297	0.291	0.223	67	0.325	**0.523**	0.412	0.233
HPRD	SNFM	9437	19.181	492	220	0.482	0.237	**0.338**	**0.256**	344	0.658	0.375	**0.497**	**0.295**
	MCL	8564	18.781	456	87	**0.802**	0.088	0.266	0.091	190	**0.789**	0.143	0.336	0.133
	DCU	9450	5.875	3832	158	0.009	0.172	0.040	0.010	144	0.014	0.273	0.061	0.013
	OH-PIN	3743	39.615	218	20	0.009	0.122	0.033	0.008	15	0.013	0.158	0.045	0.007
	k-means	8681	25.913	335	79	0.355	0.130	0.215	0.134	84	0.388	0.175	0.261	0.094
	Mclique	4288	3.563	11613	211	0.009	**0.312**	0.053	0.009	273	0.013	**0.414**	0.072	0.013
	NMF	9260	10.559	877	826	0.310	0.240	0.273	0.233	539	0.354	0.355	0.354	0.201
String	SNFM	15513	32.386	479	171	0.458	**0.243**	**0.334**	**0.211**	313	**0.680**	**0.377**	**0.506**	**0.290**
	MCL	13903	12.945	1074	145	**0.792**	0.105	0.289	0.143	307	0.646	0.202	0.362	0.172
	DCU	—	—	—	—	—	—	—	—	—	—	—	—	—
	OH-PIN	—	—	—	—	—	—	—	—	—	—	—	—	—
	k-means	14981	28.002	535	164	0.472	0.226	0.327	0.198	184	0.462	0.251	0.340	0.160
	Mclique	—	—	—	—	—	—	—	—	—	—	—	—	—
	NMF	15216	42.742	356	124	0.541	0.206	0.333	0.180	84	0.463	0.219	0.318	0.116

Finally, using DIP, HPRD and STRING data sets, we obtained both the results using CORUM data as benchmark data(see Figure 3 (a), (b) and (c)) and those using the PCDq data as benchmark data (see Figures 3 (d), (e) and (f)). The results demonstrate that the SNFM algorithm proposed in this study has obtained the best results for precision(SNFM vs MCL for DIP database to improve 8%, recall(SNFM vs NMF for DIP database to improve 14%) and F1(SNFM vs NMF for DIP database to improve 12%).



(a) DIP(ml:CORUM;gs:CORUM) (b) HPRD(ml:CORUM;gs:CORUM) (c) STRING(ml:CORUM;gs:CORUM)

(d) DIP(ml:CORUM;gs:PCDq) (e) HPRD(ml:CORUM;gs:PCDq) (f) STRING(ml:CORUM;gs:PCDq)

Figure 3. The precision, recall and F1 of compared methods on DIP, HPRD and STRING. (a), (b) and (c) take CORUM as ground-truth; (d), (e) and (f) take PCDq as ground-truth (”ml” means must-link and ”gs” means gold standard database).

3.3. Functional module cases analyses

To investigate the biological insights of detected modules, we performed functional enrichment analysis using GO terms and listed the top 5 functional modules from DIP with significant biological processes (see Table 3). We found that there exist high ratio of proteins derived from same or similar protein complexes in the top 5 function modules. For example, in the M253, the proteins: PSMC2-6, PSMD3-4, PSMD6-7,PSMD10,PSMD12 and PSMD13, are from the multiprotein complex involved in the ATP-dependent degradation of ubiquitinated proteins. In addition, for the M263, which has 17 member proteins, we also found high ratio of proteins(e.g. MED6,12-17,19-20,23-26) in a same protein complex. We further showed the corresponding subgraph mapped from DIP network, which contains the 17 protein members of M263 and their first-order neighbors (see Figure 4). It showed that although the number of links between those protein members is rather high, there exist two obvious bipartite graph structures(Lee et al. [44]) in the subgraph. We see that for the one part of the bipartite network, MED20, MED17, and CCDC belong to M263, but GATA1 and MED10 do not belong to M263. In the other part of bipartite graph, MED15 and MED6 belong to M263, but SREBF2 and SREBF1 do not belong to M263. This indicates that SNFM could detect the functional modules even with bipartite structures in a sparse network, such as DIP. This has been further validated by the additional enrichment analyses of pathway for the typical protein groups(e.g. MED17, MED20 and CCNC; MED6 and MED15, see Table 4) and GO terms for the whole M263(see Table 5).

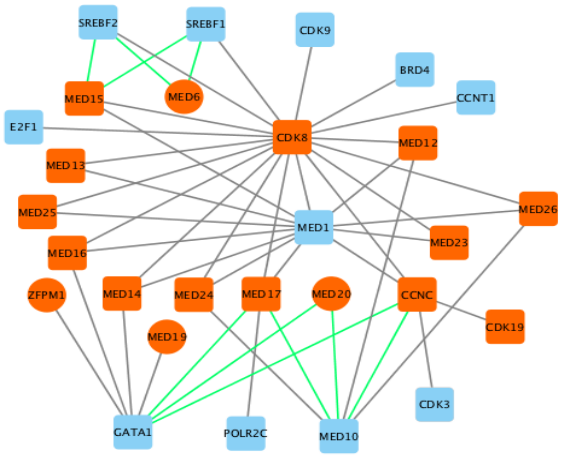


Figure 4. The protein members and their first-order neighbors of M263.The orange nodes indicate the protein members in M263, and the blue nodes are their first-order neighbors. The orange rectangular nodes satisfy the condition that they exist in the module, and they interact with proteins in the same module. The orange round node indicates that all of the proteins neighbors are not in M263.

Table 3. The top 5 functional modules from DIP with significant biological processes.

module ID	size	members	GO ID	p-value	GO term
M253	28	KDM5A PSMD10 PSMD12 DDX5 SPI1 PSMD13 PRDM2 PSMD6 PSMB5 PSMD7 PSMD4 PSMD3 TIFA PSMD1 SFN MAP3K5 CREBBP LRRC45 BAD C9orf16 ELF1 PSMC5 PSMA4 PSMC6 PSMC3 PSMC4 PSMC2 PFDN6	GO:0006521	3.63E-29	Regulation of cellular amino acid metabolic process
M331	16	UPF2 DDX6 UPF1 EPAS1 PARN ADAP1 EDC3 EDC4 EXOSC7 EXOSC5 XRN1 XRN2 EXOSC3 EIF4E2 DCP2 DCP1B	GO:0043928	9.93E-21	Exonucleolytic nuclear-transcribed mRNA catabolic process involved in deadenylation-dependent decay
M181	20	FBXW5 USP9X GPS1 CUL3 CUL2 CUL1 COPS7B COPS7A COPS4 LMAN1 COPS3 COPS6 COPS5 CDC34 COPS2 CASP4 ORMDL3 COPS8 CUL4B UBE2M	GO:0010388	1.46E-20	Cullin deneddylation
M263	17	CDK19 MED19 CCNC MED16 MED15 MED26 MED6 MED17 MED12 MED23 CDK8 MED25 MED14 MED24 MED13 MED20 ZFPM1	GO:0006367	2.94E-19	Transcription initiation from RNA polymerase II promoter
M186	18	SMARCE1 SMARCD1 PBRM1 SMARCD2 SMARCC1 SMARCC2 SMARCB1 SMARCD3 ARID1A SMARCA2 ACKR4 PVRL4 ARID2 PVRL3 PVRL2 RCOR1 PVRL1 SEC31A	GO:0006337	1.49E-18	Nucleosome disassembly

Table 4. The SupperPathway analysis of members in the bipartite graph.

	MED20	MED17	CCNC	MED15	MED6	CDK8	MED24
Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	✓	✓	✓	✓	✓	✓	✓
Gene Expression	✓	✓		✓	✓	✓	✓
Metabolism	✓	✓		✓	✓		✓
Developmental Biology	✓	✓		✓	✓		
RNA Polymerase II Transcription Initiation And Promoter Clearance	✓				✓		
Thyroid hormone signaling pathway		✓					✓
Signaling by NOTCH1			✓			✓	
Transcriptional activity of SMAD2/SMAD3-SMAD4 heterotrimer			✓			✓	
GPCR Pathway			✓				
HIV Life Cycle			✓			✓	
DNA Damage/Telomere Stress Induced Senescence				✓			
PEDF Induced Signaling							✓

SuperPaths is the unified GeneCards pathways that links to other pathways according to information extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and other databases. MED20, MED17 and CCDC are the first bipartite structures, MED15 and MED6 are the second bipartite structures, and CDK8 and MED24 are other members randomly selected in the module.

The more pathways shared by genes within the two map structures, the stronger the correlation is between them, which means the proteins should be clustered into the same module. Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha) and Metabolism are more special gene functions. This finding also illustrates the reliability of our proposed SNFM model for functional module detection.

We performed GO analysis on all proteins in M263. Table 5 lists the GO-terms including the following: biological process (BP), cellular component (CC), and molecular Function (MF). For example, the MF is GO:0001104(the RNA polymerase II transcription cofactor activity), and the P-value is 2.16E-16, and the genes contained are MED19, MED15, MED26, MED17, MED24, MED14, MED13 and MED20, which account for 8/17 of all of the genes. Table 5 shows that MED17 and MED20 are highly correlated. Unfortunately, MED17 and MED20 have not interacted with each other in the DIP network, and most of the compared algorithms have not clustered them into the same module, except for k-means. However, the detected module by k-means, which contains MED17 and MED20, is too large (∼1700 proteins). [45] has shown that larger modules have more diverse biological meanings.

Table 5. Module Gene Ontology analysis.

GOTERM_BP_DIRECT	
[redacted]	MED16, MED26, MED17, MED24, MED14, MED13
GO:0030521 androgen receptor signaling pathway (P-value = 4.67E-10)	MED16, MED17, MED24, MED14, MED13
GOTERM_CC_DIRECT	
[redacted]	MED19, MED6, MED15, MED16, MED26, MED17, CDK8, MED24, MED14, CCNC, MED13, MED20
GO:0070847 core mediator complex (P-value = 0.06)	MED6, MED14
GOTERM_MF_DIRECT	
[redacted]	MED19, MED15, MED26, MED17, MED24, MED14, MED13, MED20
GO:0003713 transcription coactivator activity (P-value = 7.43E-08)	MED16, MED26, MED17, MED14, MED13, MED20

4. Conclusion

In this study, we propose a novel semi-supervised functional module detection model named SNFM, which uses the similarity matrix constructed by subspace clustering to not only consider the first-order neighbor information of proteins but also integrates global information from the PPI networks. The obtained similarity matrix is the feature matrix of the PPI network, and the use of limited prior information extracted from known reliable protein complexes allows for discovering the protein functional modules with significant biological meaning in the PPI networks. Through experiments on real human PPI networks (DIP, HPRD, and STRING10), we found that our proposed SNFM algorithm performs better in detecting protein functional modules. Because we currently have a high-quality protein complex with a small amount of data, our future work will consider designing a protein functional module detection algorithm that combines multiple biological datasets.

Acknowledgments

The work is partially supported by the National Key Research and Development Program (2017YFC1703506), the Fundamental Research Funds for the Central Universities(2017JBM020), the Special Programs of Traditional Chinese Medicine (201407001, JDZX2015170 and JDZX2015171), and the National Key Technology *R&D* Program (2013BAI02B01 and 2013BAI13B04).

Conflict of interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

References

1. AC. Gavin, M. Bsche and R. Krause, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415** (2002), 141.
2. FY. Yu, ZH. Yang and N. Tang, et al., Predicting protein complex in protein interaction network-a supervised learning based method, *BMC. Syst. Biol.*, **8** (2014), S4.
3. Y. Ho, A. Gruhler and A. Heilbut, et al., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature*, **415** (2002), 180.
4. R. Aebersold and M. Mann, Mass spectrometry-based proteomics, *Nature*, **422** (2003), 198.
5. T. Ito, T. Chiba and R. Ozawa, et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *P. Natl. A. Sci. India. B.*, **98** (2001), 4569–4574.
6. P. Uetz, L. Giot and G. Cagney, et al., A comprehensive analysis of proteinprotein interactions in Saccharomyces cerevisiae, *Nature*, **403** (2000), 623.
7. AJ. Enright, S. Van Dongen and CA. Ouzounis, An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res.*, **30** (2002), 1575–1584.
8. T. Pawson and R. Linding, Network medicine, *FEBS. Lett.*, **582** (2008), 1266–1270.
9. LH. Hartwell, JJ. Hopfield and S. Leibler, et al., From molecular to modular cell biology, *Nature*, **402** (1999), C47.
10. J. Ji, A. Zhang and C. Liu, et al., Survey: Functional module detection from protein-protein interaction networks, *IEEE. T. Knowl. Data. En.*, **26** (2014), 261–277.
11. B. Cao, J. Luo and C. Liang, et al., PCE-FR: A Novel Method for Identifying Overlapping Protein Complexes in Weighted Protein-Protein Interaction Networks Using Pseudo-Clique Extension Based on Fuzzy Relation, *IEEE. T. Nanobiosci.*, **15**, 728–738.
12. P. Sah, LO. Singh and A. Clauset,. et al., Exploring community structure in biological networks with random graphs, *BMC. Bioinform.*, **15** (2014), 220.
13. H. Rahmani, H. Blockeel and A. Bender, Predicting the functions of proteins in protein-protein interaction networks from global information, *Machine Learn. Systems Biol.*, (2009), 82–97.
14. HN. Chua, W. K. Sung and L. Wong, Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions, *Bioinformatics*, **22** (2006), 1623–1630.
15. Y. Wang and X. Qian, Functional module identification in protein interaction networks by interaction patterns, *Bioinformatics*, **30** (2013), 81–93.
16. G. Liu, B. Chai and K. Yang, et al., Overlapping functional modules detection in PPI network with pair-wise constrained non-negative matrix trifactorisation, *IET. Syst. Biol.*, **12** (2017), 45–54.
17. V. Spirin and LA. Mirny, Protein complexes and functional modules in molecular networks, *P. Natl. A. Sci. India. B.*, **100** (2003), 12123–12128.
18. X. L. Li , C. S. Foo and S. H. Tan, et al., Interaction graph mining for protein complexes using local clique merging, *Genome Inform.*, **16** (2005), 260–269.
19. B. Adamcsek, G. Palla and IJ. Farkas, et al., CFinder: locating cliques and overlapping modules in biological networks, *Bioinformatics*, **22** (2006), 1021–1023.

20. M. Wu, X. Li and C. K. Kwoh,. et al., A core-attachment based method to detect protein complexes in PPI networks, *BMC Bioinform.*, **10** (2009), 169.
21. J. Menche, A. Sharma and M. Kitsak, et al., Uncovering disease-disease relationships through the incomplete interactome, *Science*, **347** (2015), 1257601.
22. X. F. Zhang, D. Q. Dai and L. Ou-Yang, et al., Detecting overlapping protein complexes based on a generative model with functional and topological properties, *BMC Bioinform.*, **15** (2014), 186.
23. E. Georgii, S. Dietmann and T. Uno, et al., Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, **25** (2009), 933–940.
24. Y. Zhang, N. Du and L. Ge, et al., A collective nmf method for detecting protein functional module from multiple data sources, *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM*, (2012), pp. 655–660.
25. S. Kikugawa, K. Nishikata and K. Murakami, 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.*, **6**, S7.
26. Y. Qi, F. Balem and C. Faloutsos, et al., Protein complex identification by supervised graph local clustering. *Bioinformatics*, **24** (2008), i250–i268.
27. E. Elhamifar and R. Vidal, Sparse subspace clustering, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), pp 2790–2797.
28. F. Wang, T. Li and X. Wang, et al., Community discovery using nonnegative matrix factorization, *DATA. Min. Knowl. Disc.*, **22** (2011), 493–521.
29. J. Menche, A. Sharma and M. Kitsak, et al., Uncovering disease-disease relationships through the incomplete interactome, *Science*, **347** (2015), 1257601.
30. A. Ruepp, B. Waegele and M. Lechner, et al., CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic acids Rec.*, **38** (2009), D497–D501.
31. L. Yang, X. Cao and D. Jin, et al., A unified semi-supervised community detection framework using latent space graph regularization, *Data. Min. Knowl. Disc.*, **45** (2015), 2585–2598.
32. I. Xenarios, L. Salwinski and X. J. Duan, et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.*, **30** (2002), 303–305.
33. TS. Keshava Prasad, R. Goel and K. Kandasamy, et al., Human protein reference database 2009 update, *Nucleic Acids Res.*, **37** (2008), D767–D772.
34. C. Von Mering, LJ. Jensen and B. Snel, et al., STRING: known and predicted proteinprotein associations, integrated and transferred across organisms, *Nucleic Acids Res.*, **33** (2005), D433–D437.
35. C. Peng and A. Li, A heterogeneous network based method for identifying GBM-related genes by integrating multi-dimensional data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB).*, **14** (2017), 713–720.
36. A. Ruepp, B. Brauner and I. Dunger-Kaltenbach, et al., CORUM: the comprehensive resource of mammalian protein complexes, *Nucleic Acids Res.*, **36** (2007), D646–D650.

37. B. Chen, W. Fan and J. Liu, et al., Identifying protein complexes and functional modules from static PPI networks to dynamic PPI networks, *Brief Bionfrom.*, **15** (2013), 177–194.
38. Y. Yu, J. Liu and N. Feng, et al., Combining sequence and Gene Ontology for protein module detection in the Weighted Network, *J. Theor. Biol.*, **412** (2017), 107–112.
39. B. Zhao, J. Wang and M. Li, et al., Detecting protein complexes based on uncertain graph model, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **11** (2014), 486–497.
40. J. Wang, J. Ren and M. Li, et al., Identification of hierarchical and overlapping functional modules in PPI networks, *IEEE. T. Nanobiosci.*, **11** (2012), 386–393.
41. S. van Dongen, Graph clustering by flow simulation, PhD thesis, University of Utrecht,(2000).
42. G. Liu, L. Wong and H. N. Chua, Complex discovery from weighted PPI networks, *Bioinformatics*, **25** (2009), 1891–1897.
43. J. MacQueen, Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, (1967). p. 281–297.
44. J. Lee and J. Lee, Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction, *PloS one*, **8** (2013), e60372.
45. G. Liu, H. Wang and H. Chu, et al., Functional diversity of topological modules in human protein-protein interaction networks, *Sci. Rep-UK.*, **7** (2017), 16199.

Supplementary

† These authors contributed equally to this work.

AIMS Press