MIPCE: An MI-based protein complex extraction technique

PRIYAKSHI MAHANTA[1,*], DHRUBA KR BHATTACHARYYA[1] and ASHISH GHOSH[2]

[1]*Department of Computer Science and Engineering, Tezpur University, Napaam 784 028, India*

[2]*Machine Intelligent Unit, Indian Statistical Institute, Kolkata 700 108, India*

**Corresponding author (Email, priyakshi@tezu.ernet.in)*

Protein–protein interaction (PPI) networks are believed to be important sources of information related to biological processes and complex metabolic functions of the cell. Identifying protein complexes is of great importance for understanding cellular organization and functions of organisms. In this work, a method is proposed, referred to as MIPCE, to find protein complexes in a PPI network based on mutual information. MIPCE has been biologically validated by GO-based score and satisfactory results have been obtained. We have also compared our method with some well-known methods and obtained better results in terms of various parameters such as precession, recall and F-measure.

[Mahanta P, Bhattacharyya DK and Ghosh A 2015 MIPCE: An MI-based protein complex extraction technique. *J. Biosci.* **40** 701–708] **DOI** 10.1007/s12038-015-9553-1

1. Introduction

Protein–protein Interaction (PPI) networks are believed to be important sources of information related to biological processes and complex metabolic functions of the cell. Recent advances in biotechnology have resulted in a large amount of PPI data. The increasing amount of available PPI data necessitates a fast, accurate approach to biological complex identification. Because of its importance in the studies of protein interaction network, there are different models and algorithms for identifying functional modules in PPI networks. To analyse the complex networks of PPIs and to identify protein complexes or functional modules from them is one of the most important challenges in the post-genomic era. Identification of functional modules in protein interaction networks is a first step in understanding the organization and dynamics of cell functions. System-level understanding of biological organization is a key objective of the post-genomic era. Complex cellular processes are modular and are accomplished by the concerted action of functional modules. These modules encompass groups of genes or proteins involved in common elementary biological functions as found in Collins *et al.* (2007). Revealing modular structures in biological networks Mahanta *et al.* (2012) will help us understand how cells function. To cope with the ever-increasing volume and complexity of protein interaction data, many methods which are based on modelling the PPI data with graphs have been developed for analysing the structure of PPI networks.

PPI networks are represented as undirected graphs with nodes corresponding to proteins and edges representing the interactions between two proteins, where self-loops and parallel edges are not considered, as stated in Adamcsek *et al.* (n.d.). PPI networks are important sources of information related to biological process and complex metabolic functions of the cell. Cluster analysis is a choice of methodology for the extraction of functional modules Mahanta *et al.* (2014) from protein–protein interaction networks. Clustering can be defined as the grouping of objects based on their sharing of discrete and measurable properties. In PPI networks, clusters correspond to two types of modules:

- Protein complex: It is a physical aggregation of several proteins via molecular interaction with each other at the same location and time.
- Functional module: It consists of a number of proteins (and other molecules) that interact with each other to

Keywords. Mutual information; PPI network; protein complex; topological property

Published online: 28 September 2015

control or perform a particular cellular function. Unlike protein complexes, these proteins do not necessarily interact at the same time and location.

Mutual information Hoque *et al.* (2014) measures the correlation between two random variables. In the context of biological network inference, a higher mutual information between two genes or gene products indicate a higher dependency Mahanta *et al.* (2013), and therefore a possible interaction between them. For a given two variables A and B, the mutual information measures how much knowing one of these variables reduces our uncertainty about the other. The mutual information between two gene expression patterns by Cover and Thomas (2012) is given by the following equation:

$$M\ I(A;B) = H(A) + H(B) - H(A,B) \tag{1}$$

where $H(A, B)$ is the joint entropy of two gene expressions A and B.

An alternative to this method based on kernel density estimation (KDE) was suggested by Moon *et al.* (1995). The MI between two genes X and Y with continuous expression values is given by equation 2:

$$M\ I(X,Y) = \int_x \int_y f(x,y) log\left(\frac{f(x,y)}{f(x)f(y)}\right) dxdy \tag{2}$$

where $f(x;y)$ is the joint probability density of the two random variables, and $f(x)$ and $f(y)$ are the marginal densities. For a given m data points or conditions in the dataset, the joint and marginal densities can be estimated by the Gaussian kernel estimator given by equations 3 and 4:

$$f(x,y) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2\pi h^2} \exp\left(-\frac{1}{2h^2}\left((x-x_i)^2 + (y-y_i)^2\right)\right) \tag{3}$$

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2\pi h^2} \exp\left(-\frac{1}{2h^2}\left((x-x_i)^2\right)\right) \tag{4}$$

where h is the width of the kernels. The MI equation can be approximated by the sample average as given in equation 5:

$$M\ I(X,Y) = \frac{1}{m}\sum_i log \frac{m\sum_j \exp\left(-\frac{1}{2h^2}\left((x_i-x_j)^2 + \left(y_i-y_j\right)^2\right)\right)}{\sum_j \exp\left(-\frac{1}{2h^2}\left(x_i-x_j\right)^2\right)\sum_j \exp\left(-\frac{1}{2h^2}\left(y_i-y_j\right)^2\right)}. \tag{5}$$

One of the advantages of kernel density estimation over simple frequency histograms is that only one parameter (i.e. kernel width, h) must be set, whereas the histogram requires two parameters, i.e. bin width and origin. But the traditional kernel estimation method takes high computation time due to repeated calculations of the kernel distances for each gene expression while computing the pair-wise MI for all pairs. To avoid this, we use the modified kernel estimation method described in Qiu *et al.* (2009) for MI estimation where the computation time is reduced by changing the order of the loops.

1.1 *Motivation*

The main idea was to go beyond pair-wise interactions and utilize additive features between any nodes. While this increases the network inference complexity over pair-wise interaction-based approaches, it achieves much higher accuracy. Moreover, we focus on reducing wrong edges while extracting protein complexes from PPI network. To achieve these two goals, we considered proteins with high mutual information while neglecting those with lower mutual information.

1.2 *Contributions*

The following contributions have been made in this study.

- An effective mutual-information-based PPI network construction technique referred to as MIPCE has been introduced which is able to extract protein complexes of high functional coherence.
- GO-based validation of the protein complexes given by MIPCE.
- Evaluation of MIPCE in terms of precision, recall and F-measure and comparison with its several counterparts.

The rest of the article is organized as follows. In section 2 related research is reported. The proposed method is explained in section 3. The detailed experimental results are presented in section 4. Finally, concluding remarks and research directions are given in section 5.

2. Related work

In order to solve the problem of detecting all possible protein complexes in a PPI network, several computational proposals have been introduced. In this section we discuss some of the popular methods for protein complex identification.

Cfinder: Adamcsek *et al.* (2006) provide a software called CFinder which can detect the K-clique percolation clusters as functional modules using a Clique Percolation Method. A K-clique is a clique with K nodes, and two K-cliques are adjacent if they share common nodes. The final cluster is constructed by linking all the adjacent K-cliques as a bigger subgraph.

DPClus: Altaf-Ul-Amin *et al.* (2006) propose a cluster periphery tracking algorithm called DPClus to detect protein complexes by keeping track of the periphery of a detected cluster. DPclusfirst weighs each edge based on the common neighbours between its two proteins and further weighs nodes by its weighted degree. To form a protein complex, DPClusfirst chose the

node with highest weight as the initial cluster and then iteratively augments this cluster by including vertices one by one.

MCL: Saturuli *et al.* (2010) introduces Markov Clustering (MCL) which can be applied to detect functional modules and protein complexes by simulation of random walks in PPI network. MCL manipulates the weighted or unweighted adjacency matrix with two operators called expansion and in inflation. Iterative expansion and inflation will separate the PPI network into many segments as protein complexes.

MCODE: It is one of the first computational methods proposed by Bader *et al*. to detect protein complex based on protein connectivity values in a PPI network. In MCODE, Bader and Hogue (2003) first weigh each node based on their local neighbourhood densities and then select the seed node with high weights as initial clusters and augments these clusters by outward traversing from the seed.

Coach: In the algorithm, Wu *et al.* (2009), the preliminary cores are first detected from the neighborhood graph of each vertex in the PPI network. This core-attachment-based method to detect protein complexes from PPI networks mines the protein-complex cores from the neighbourhood graphs and then forms protein complexes by including attachments into cores.

A summary of the different protein complex extraction algorithms is given in table 1. From the literature survey, we see that in the case of PPI network prediction, the most popular statistical method is clustering. Clustering methods are more suitable for the PPI network inference problem as the main emphasis is on the identification of protein complexes. It is found that certain important and popular modeling techniques may fail to model PPI networks. Also, clustering methods based on mutual information could be used as stated in Zhou *et al.* (2003) to extract protein complex.

3. MIPCE: The proposed MI-based PPI complex extraction method

Definitions given below and various symbolic representations given in table 2 describe the theoretical foundation of the proposed method. The algorithm for finding protein complex from the PPI data is given in Algorithm 1. A flowchart of the proposed method is given in figure 1.

Definition 1: A protein complex is a set of connected proteins P_i where mi (P_i) € any top k MI values of the PPI network; and the member proteins of the complex shows high functional coherence.

Definition 2: Connectivity of a protein p_i is defined as the fraction of connection of p_i with other proteins to the maximum connection value of a protein.

Definition 3: Score of a node assigns higher weights to nodes whose immediate neighbors are more interconnected. Score are assigned to each node with the following steps:

1. Get the immediate neighbors of a node to score.
2. Find the highest k core network.
3. Calculate core density=edges presented in the subgraph/possible edges.
4. Score of the node=k×core density.

To extract the mutual information (MI) between every pair of proteins, we construct an MI matrix with the following information as shown in figure 2 from PPI data:

(a) Each row represents a protein.
(b) The first column gives the connectivity of proteins.
(c) The second column gives the score of proteins.

Algorithm 1: Protein Complex Extraction

```
Input :P and k
Output :C
Construct the adjacency matrix;
Find mi for the connected pairs;
Sort the mi values in descending order to find highest mi distinct
protein pairs; Select k distinct pairs from the sorted pairs ;
Each k^th pair now represents C_k;
For each p_l € P do
    If p_l not € any of the k^th set, C_k then
        for each k^th set do
            find avgmi(l,k)
        End
        Find k for which avgmi(l,k) is maximum;
        Assign p_l to this k^th set, C_k;
        cnt(k)= cnt(k) + 1;
        Else
        l=l+1;
    End
End
```

This method first constructs the adjacency matrix from the PPI data. After that it finds the MI between all the protein pairs for the connected proteins from the MI matrix using kernel estimation method. Computing h for each pair of proteins is a time-consuming process. Therefore, we propose a way to compute h which will work for all pairs of proteins. The steps can be summarized as follows:

1. Find standard deviation of each protein p_i, std(p_i) considering all conditions for i=1,2,3,….
2. h=std(std(p_i)).

The method sorts the MI values in descending order to select the k distinct highest MI seed pairs. Once the seed

Table 1. Comparison of protein complex extraction techniques

Techniques	#Input parameters	Datasets used	Gold standards	Performance measures
COACH (Wu *et al.* (2009))	2	DIP, Krogan	Friedel	*p*-value, Coannotation score, Colocalization score, F measure, Coverage rate
DPClus Altaf-Ul-Amin *et al.* (2006)	1	DIP, Krogan	MIPS, SGD, Alloy *et al.*	*p*-value, F measure
MCODE (Bader & Hogue (2003))	1	Gavin, MIPS, YPD	Gavin, MIPS	No. of complexes detected, Sensitivity, Specificity
CFinder (Adamcsek *et al.* (2006))	1	N/A	N/A	N/A
MCL (Satuluri *et al.* (2010))	1	Uetz, Ito, Gavin, Ho, Krogan	MIPS	Sensitivity, PPV, Accuracy

pairs are found, we expand these seed pairs to k protein complexes by adding to them the remaining proteins based on the average mutual information content of the chosen protein with proteins in each of the k-th seed pairs. The protein will belong to that pair with which the average mutual information is maximum. This continues till all the proteins of the dataset are assigned to its respective complex.

The following theorem ensures proper formation of core and peripheral regions in a cluster.

Theorem 1: Seed expansion phase of MIPCE involves inclusion of those nodes which have high functional coherence.

Proof: Mutual Information (MI) is reasonably immune against missing data and outliers and also potentially more robust for differentiating erroneous clustering solutions (Priness *et al.* 2007). In Butte and Kohane (2000), the authors hypothesize that an association between two genes indicated by large amount of mutual information between them would also signify biological relationship. So, high mutual information between protein pairs indicates high functional coherence.

Seed pairs are formed from the highest MI values of protein pairs. MI values of proteins are extracted based on protein connectivity and score based on highest k core network. Higher connectivity and score ensures higher density, and so higher MI values have higher functional coherence. While expanding the seed, a protein gets included in a protein complex if its average MI value is maximum with respect to that particular protein complex. Hence, a node must have higher MI value to get included into the protein complex along the phase of seed expansion which ensures high functional coherence.

Table 2. Symbols used

Symbol	Meaning
P	PPI data
A	Adjacency matrix
Mi	Matrix containing mutual information of all pairs of proteins in A
(s_{i1},s_{i2})	i-th seed pairs
k	Number of protein complex
cnt(k)	Number of proteins in the k-th complexes
avgmi(l; k)	Average mi values of the l-th protein with all cnt(k) proteins of k-th complex
C	Set of protein complexes
C_k	k-th protein complex

3.1 *Effect of input parameter*

To evaluate the effect of the parameter k on the results, we changed the values of parameter from 3 to 10 with increment 1. The value of input parameter k was chosen heuristically from the graph shown in figure 3, where the p-value is minimum. A low p-value (highest biological significance) indicates that the proteins belonging to the enriched functional categories are biologically significant in the corresponding complex.

3.2 *Complexity analysis*

The time complexity analysis of MIPCE is $O(n^2)$, where n is the number of proteins in the dataset, k is the number of protein complex. Finding the Mutual Information (MI) matrix is of $O(n^2)$. Next the average MI of each protein with each of the k

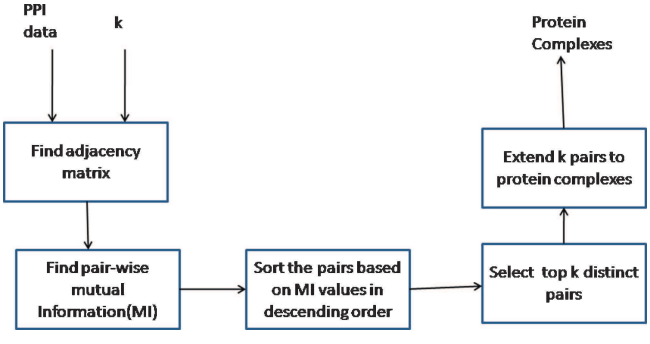

Figure 1. Conceptual framework of our proposed method.

Proteins	Connectivity	Score
P_1	$Conn(P_1)$	$Score(P_1)$
P_2	$Conn(P_2)$	$Score(P_2)$
P_3	$Conn(P_3)$	$Score(P_3)$
.	.	.
.	.	.
P_n	$Conn(P_n)$	$Score(P_n)$

Figure 2. Input for the MI computation.

complex requires k computations, and further k computations (worst case) are required for finding the maximum out of these k average values, thus making it of the order of $O(n^2) + O(nk^2)$. So, the final complexity of MIPCE is approximately that of $O(n^2)$.

4. Experimental results

We implemented the algorithm in MATLAB and tested it on three benchmark microarray datasets. The test platform was a Sun workstation with Intel(R) Xenon(R) 3.33 GHz processor and 6 GB memory running Windows XP operating system.

4.1 *Dataset used*

In this study, we used four well-known datasets, viz. Gavins dataset (dataset 1) Gavin et al. (2006), DIP dataset (dataset 2) Xenarios *et al.* (2002) and Krogan et al. 2006 core (dataset 3)

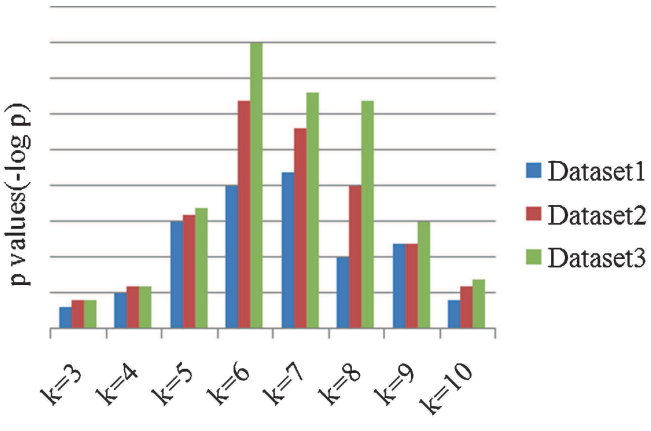

Figure 3. Tuning of input parameter k.

Table 3. Accuracy comparison using DIP dataset

Methods	Precision	Recall	F-measure
MCODE (Bader and Hogue 2003)	0.43	0.1	0.17
MCL (Satuluri *et al.* 2010)	0.18	0.59	0.27
Coach (Wu *et al.* 2009)	0.38	0.56	0.47
DPClus (Altaf-Ul-Amin *et al.* 2006)	0.34	0.27	0.29
Cfinder (Adamcsek *et al.* 2006)	0.17	0.71	0.27
MIPCE	0.39	0.5	0.44

Krogan *et al.* (2006), to validate the proposed algorithm. Among 4,087 different proteins identified with high confidence by mass spectrometry from 2,357 successful purifications, the Krogan data set (median precision of 0.69) comprises 7,123 protein–protein interactions involving 2,708 proteins. BioGRID does not contain confidence scores for all interactions, and confidence scores of interactions from different data sources may be incompatible with each other, and so we used all interactions without any confidence filters and did not add any weight information to this dataset. In the DIP dataset, a total of 17,201 interactions among 4,934 proteins were considered.

4.2 *Cluster accuracy validation*

To measure the accuracies of prediction, we calculated precision, recall and F-measure for different algorithms. Precision is the ratio of predicted complexes which are true to the total number of predicted complexes, and Recall is the ratio of true complexes predicted to the total number of true complexes. Suppose A is the number of predicted complexes which are true and B is the number of predicted complexes which are not true. Mathematically, Precision is defined as

$$\text{Precision} = \frac{A}{A + B} \tag{6}$$

Again, A is the number of true complexes which are predicted and C is the number of total complexes which are not predicted. Mathematically, Recall is defined as

Table 4. Accuracy comparison using Krogan core dataset

Methods	Precision	Recall	F-measure
MCODE	0.43	0.1	0.17
MCL	0.18	0.59	0.27
COACH	0.38	0.56	0.47
Cfinder	0.34	0.27	0.29
DPClus	0.17	0.71	0.27
MIPCE	0.39	0.5	0.28

Table 5. Accuracy comparison using Gavin dataset

Methods	Precision	Recall	F-measure
MCODE	0.73	0.29	0.41
COACH	0.54	0.27	0.36
Cfinder	0.66	0.19	0.29
MCL	0.52	0.33	0.4
MIPCE	0.53	0.41	0.38

$$\text{Recall} = \frac{A}{A + C} \tag{7}$$

F-measure is the harmonic mean of Precision and Recall.

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \textit{Recall}}{\text{Precision} + \textit{Recall}} \tag{8}$$

The neighbourhood affinity (NA) score Wu *et al.* (2009) between a predicted complex p and a real complex b in the benchmark, N A(p; b), was used to determine whether they match with each other. If N A(p; b) w, they are considered to be matching (w is set as 0.20 in most approaches, which was also used in this study).

To analyse the effectiveness of the algorithm for protein complex identification purpose, we compared the results of MIPCE with existing methods in terms of various evaluation metrics, including Precession, Recall and F-measure for datasets. The comparison result is presented in tables 3, 4 and 5 and in figures 4, 6 and 7. As we can see, the proposed method gives satisfactory result in all measures. However, COACH shows better F-measure than MIPCE for dataset 2 and 3 because the complex size is much bigger in case of MIPCE which includes some additional proteins that belong to some other functional category. The roc curve of MIPCE for dataset 2 is given in figure 5.

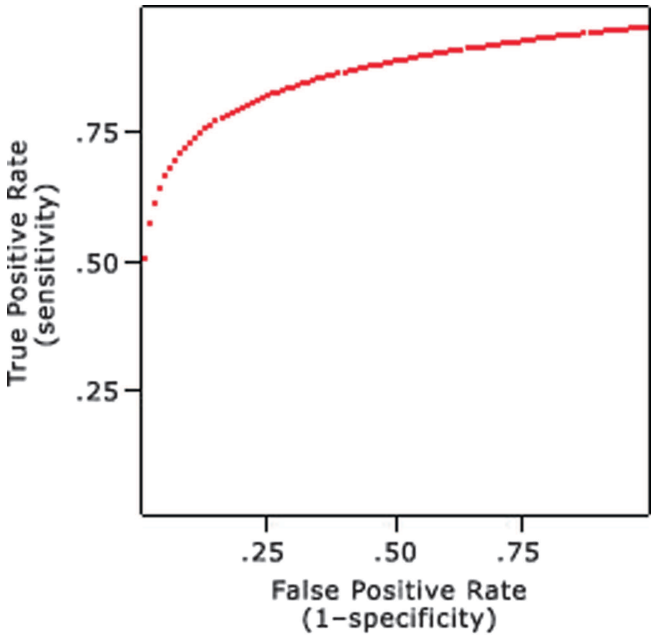

Figure 5. ROC curve of MIPCE for dataset 2.

4.3 *GO-based validation*

Gene ontology (GO) provides description of gene and gene product attributes in a structured and controlled manner across all species. To evaluate the functional enrichment of predicted protein complexes, the *p*-value of a protein

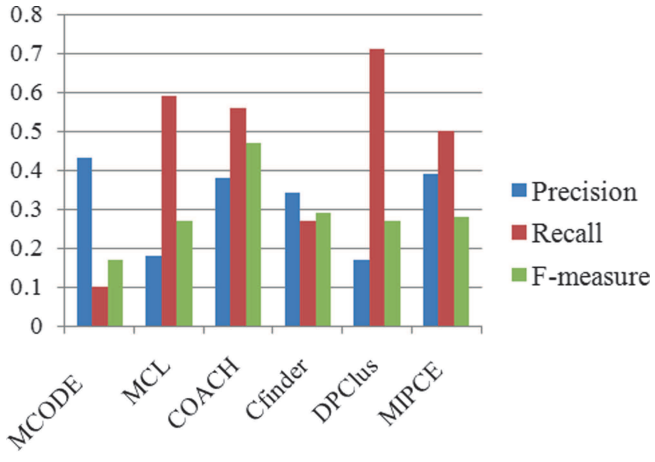

Figure 4. Performance comparison of MIPCE with existing methods for dataset 1.

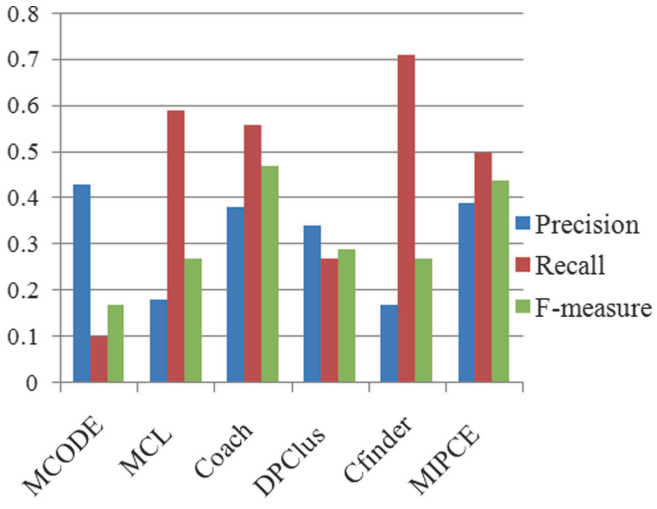

Figure 6. Performance comparison of MIPCE with existing methods for dataset 3.

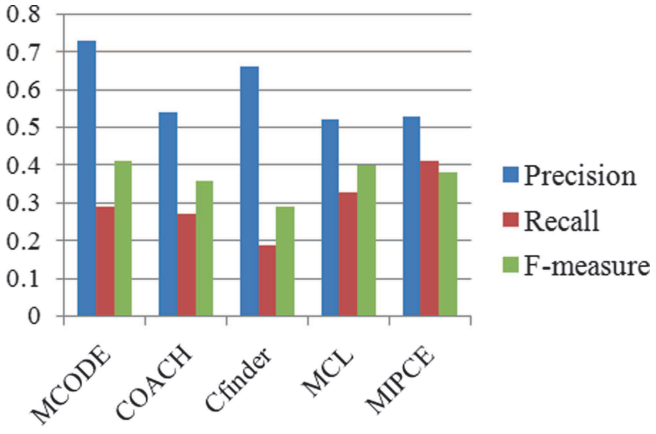

Figure 7. Performance comparison of MIPCE with existing methods for dataset 2.

complex with a given GO term was used to estimate whether the proteins in the complex are enriched for the GO term with a statistically significant probability compared to what one would expect by chance. The smaller *p*-value indicates the predicted protein complexes is not accumulated at random and is more biologically significant than the one with a larger *p*-value (Mete *et al.* 2008).

We analysed the protein complexes obtained by MCODE, COACH, PROCO-MOSS Mukhopadhyay *et al.* (2012) and the corresponding MIPCE complexes using Gene Ontology to compare the effectiveness of these algorithm in terms of identifying protein complexes. We are considering the *p*-value of *k* number of complexes for which biological significance is maximum. The most significant annotations which belong to known complexes are 1.05E−41 (GO:010499), 7.88E−40 (GO:042273), 3.72E−35 (GO:043634), 3.72E−35 (GO:043633) and 1.18E−34 (GO:000459), which are greater than MCODE. In case of PROCOMOSS, the significant annotation are 1.68E−23 (GO:0044238), 4.96E−22 (GO:0006364), 1.61E−26 (GO:0051603), 4.34E−13 (GO:0043413) and 3.46E−28 (GO:0006511). While comparing with COACH, the annotation for the real complex are 2.28E−34 (RNA polymerase II mediator complex), 1.38E−39 (DNA-directed RNA polymerase III complex), 3.46E−28 (SAGA complex) and 5.26E−36 (anaphase-promoting complex). Tables 6, 7, 8 and 9 present some GO values given by proposed method which are better or equally good as compared to existing methods. We can conclude from the above-mentioned tables that MIPCE shows highly satisfactory results, especially in terms of recall values and equally good in terms of F-measure.

Table 6. Performance comparison of the proposed method with MCODE in terms of *p*-value for dataset 3

Gene Ontology ID	MIPCE	MCode
GO:019941	6.78E−32	2.29E−18
GO:051603	2.23E−30	8.44E−18
GO:000459	1.18E−34	4.64E−22
GO:010499	1.05E−41	5.35E−27
GO:006508	2.46E−28	6.12E−14
GO:006353	6.68E−30	5.78E−17
GO:043634	3.72E−35	1.48E−22
GO:043633	3.72E−35	1.48E−22
GO:031123	3.93E−28	2.12E−15
GO:043241	8.13E−25	1.22E−11

Value of *k* is chosen 6 for dataset 3. The most significant term of each of the 6 complex is reported

Table 7. Performance comparison of the proposed method with MCODE in terms of *p*-value for dataset 1

GO annotation	MIPCE	MCode
GO:042273	7.88E−40	3.28E−09
GO:031327	2.87E−19	1.90E−04
GO:045934	4.96E−12	1.31E−04
GO:016481	1.78E−14	6.62E−05
GO:051172	9.46E−18	1.31E−04
GO:042257	4.26E−16	2.61E−06
GO:048519	7.63E−12	2.21E−03
GO:022618	3.53E−17	1.65E−06
GO:000462	4.77E−34	5.56E−05
GO:070925	1.84E−19	2.21E−03
GO:000447	3.53E−18	1.98E−04
GO:030490	2.07E−33	7.90E−05

Value of *k* is chosen 7 for dataset 1. The most significant term of each of the 7 complex is reported.

Table 8. Performance comparison of extracted protein complex with COACH in terms of *p*-value for dataset 2

GO annotation	MIPCE	COACH
RNA polymerase II mediator complex	2.28E−34	6.61 E−23
DNA-directed RNA polymerase III complex	1.38E−39	5.85 E−30
HOPS complex	4.96E−12	1.66 E−13
COMPASS complex	1.78E−14	1.57E−20
SAGA complex	3.46E−28	5.00 E−24
anaphase-promoting complex	5.26E−36	9.85 E−33
OST complex	5.9E−11	0.43 E−15

Value of *k* is chosen 6 for dataset 2. The most significant term of each of the 6 complex is reported.

Table 9. Performance comparison of extracted protein complex with PROCOMOSS in terms of *p*-value for dataset 2

GO annotation	MIPCE	PROCOMOSS
GO:0044238	1.68E−23	4.09E−16
GO:0006364	4.96E−22	3.64E−19
GO:0051603	1.61E−26	3.46E−19
GO:0043413	4.34E−13	4.89E−08
GO:0006511	3.46E−28	8.20E−25
GO:0044238	3.67E−15	4.09E−16

5. Conclusion and future work

In this article, a PPI network construction and protein complex extraction technique based on a Mutual Information was presented. The technique has been established over for three publicly available benchmark real-life datasets. The method was found to produce satisfactory results when compared with its counterparts. This work can be extended to detect functional modules using an integrated approach with different sources of data. In future, we aim to explore the possibility of MIPCE in domain–domain interaction network to extract interesting information (tables 8 and 9).

References

Adamcsek B, Palla G, Farkas IJ, Dernyi I and Vicsek T n.d. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 1021–1023

Adamcsek B, Palla G, Farkas IJ, Dernyi I and Vicsek T 2006 CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22** 1021–1023

Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K and Kanaya S 2006 Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinf.* **7** 207

Bader G and Hogue C 2003 An automated method for nding molecular complexes in large protein interaction networks. *BMC Bioinf.* **4** 2

Butte AJ and Kohane IS 2000 Mutual information relevance networks: functional genomic clustering using pair wise entropy measurements. *Pac. Symp. Biocomput.* **5** 415–426

Collins SR, Kemmeren KP, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS and Krogan NJ 2007 Toward a comprehensive atlas of the physical inter-actome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **6** 439–450

Cover TM and Thomas JA 2012 *Elements of information theory* (John Wiley & Sons)

Gavin A *et al*. 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nature* **631** 631–636

Hoque N, Bhattacharyya DK and Kalita JK 2014 MIFS-ND: a mutual information-based feature selection method. *Expert Syst. Appl.* **41** 6371–6385

Krogan NJ, Cagney G, Yu H, Zhong G and Guo X 2006 Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440** 637–643

Mahanta P, Ahmed HA, Bhattacharyya DK and Kalita JK 2012 An e ective method for network module extraction from microarray data. *BMC Bioinf.* **13** S4

Mahanta P, Bhattacharyya DK and Ghosh A 2013 A subspace module extraction technique for gene expression data. In *Pattern Recognition and Machine Intelligence. Proceedings of the 5th International Conference, PReMI 2013, Kolkata, India, December 10-14, 2013* pp 635–640

Mahanta P, Ahmed HA, Bhattacharyya DK and Ghosh A 2014 FUMET: a fuzzy network module extraction technique for gene expression data. *J. Biosci.* **39** 351–364

Mete M, Tang F, Xu X, and Yuruk N 2008 A structural approach for nding functional modules from large biological networks. *BMC Bioinf.* **9** S-9

Moon Y-I, Rajagopalan B and Lall U 1995 Estimation of mutual information using kernel density estimators. *Phys. Rev. E.* **52** 2318–2321

Mukhopadhyay A, Ray S and De M 2012 Detecting protein complexes in a PPI network: a gene ontology based multi-objective evolutionary approach. *Mol. BioSyst.* **8** 3036–3048

Priness I, Maimon O and Ben-Gal IE 2007 Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinf.* **8** 11

Qiu P, Gentles AJ and Plevritis SK 2009 Fast calculation of pairwise mutual information for gene regulatory network reconstruction. *Comput. Methods Prog. Biomed.* **94** 177–180

Satuluri V, Parthasarathy S and Ucar D 2010 Markov clustering of protein interaction networks with improved balance and scalability; in *Proceeding of the First ACM International Conference on Bioinformatics and Computational Biology (ACM)* pp 247–256

Wu M, Li X, Kwoh CK and Ng S-K 2009 A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinf.* **10** 169

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M and Eisenberg D 2002 DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30** 303–305

Zhou X, Wang X and Dougherty ER 2003 Construction of genomic networks using mutual-information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design. *Signal Process.* **83** 745–761