HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

Original Article

Detecting protein complexes based on a combination of topological and biological properties in protein-protein interaction network

Pooja Sharma [a], D.K. Bhattacharyya [a,*], J.K. Kalita [b]

[a] *Department of Computer Science & Engineering, Tezpur University Napaam, Tezpur 784028, Assam, India*
[b] *Department of Computer Science, University of Colorado at Colorado, Springs, CO 80933-7150, USA*

ARTICLE INFO

Article history:
Received 1 August 2017
Received in revised form 1 November 2017
Accepted 17 November 2017
Available online 26 November 2017

Keywords:
Protein complex
Connectivity
Semantic similarity
Contribution

ABSTRACT

Protein complexes are known to play a major role in controlling cellular activity in a living being. Identifying complexes from raw protein protein interactions (PPIs) is an important area of research. Earlier work has been limited mostly to yeast. Such protein complex identification methods, when applied to large human PPIs often give poor performance. We introduce a novel method called CSC to detect protein complexes. The method is evaluated in terms of positive predictive value, sensitivity and accuracy using the datasets of the model organism, yeast and humans. CSC outperforms several other competing algorithms for both organisms. Further, we present a framework to establish the usefulness of CSC in analyzing the influence of a given disease gene in a complex topologically as well as biologically considering eight major association factors.

1. Introduction

The understanding of functional and physical interactions among molecules in the living body is of utmost importance in biology. Proteins are referred to as the most essential molecules in the living body and hence determination of the interactions among them can add to the existing domain knowledge in biology. Certain experimental methods such as mass spectrometry [1] and yeast-to-hybrid help identify interconnections among pairs of proteins and also build the network of interacting proteins. However, a major drawback of these methods is that they are unable to detect presence of interactions involving more than two protein partners [2]. From biological discoveries, we are aware that often a group of proteins act together at the same time and place to form a protein complex [3] and these complexes are responsible for carrying out various activities in the body. For example, the SWI/SNF complex is associated with remodelling of the DNA [4]. Thus, understanding interactions among groups of proteins (i.e., protein complexes) is more beneficial rather than emphasizing just on the interconnection between any two protein pairs. Protein complex detection using experimental techniques is challenging as the amount of PPI data is increasing at a very fast rate. Thus computational methods that can go beyond pairwise analysis may be able to complement the dominant study paradigm in PPI analysis. A PPI network is generally represented as a graph where vertices correspond to proteins and edges between them correspond to interactions between proteins. The task of identifying protein complexes from the PPI network (PPIN) can be modelled as a clustering problem, where the task is to identify similar types of proteins. The similarity between proteins can be in terms of distance, graph similarity or any other suitable metric. Use of different metrics results in different set of complexes. Analysis of these complexes has always been a research issue to the biologists.

Interactions among proteins decide the molecular and cellular functions in healthy and diseased states of organisms [5]. The molecular basis of diseases can be explored from the PPI network point of view, and relevant discoveries may aid in the prevention, diagnosis and treatment of diseases. Protein complexes in a PPI network can be used for gaining insight into genetic pathways and may also aid in analyzing the different progression stages in diseases [5]. Studies have also shown that disease genes tend to lie at the periphery of the networks and that they are highly connected among themselves. Careful analysis of disease networks can prove helpful in drug design as well. Association links between disease gene(s) and other genes in a complex can also help in prioritization [6,7] of complexes to narrow down the search space for bioinformaticists and enhance the analysis process.

Peer review under responsibility of Academy of Scientific Research & Technology.

* Corresponding author.
E-mail addresses: poojasharma0303@gmail.com (P. Sharma), dkb@tezu.ernet.in (D.K. Bhattacharyya), jkalita@uccs.edu (J.K. Kalita).

https://doi.org/10.1016/j.jgeb.2017.11.005
1687-157X/© 2017 Production and hosting by Elsevier B.V. on behalf of Academy of Scientific Research & Technology.

Despite the availability of numerous protein complex finding methods, not much attention has been paid to their performance in case of human datasets. Empirical analysis of a few of the existing methods has revealed that these methods perform well mostly for yeast datasets. Thus, it is necessary to find a method which can perform equally well for both yeast as well as human datasets. We have introduced an effective method to predict complexes of high biological significance in this paper. Further, we have established the method in terms of well-known performance measures such as Sensitivity, Positive Predictive Value and Accuracy for two model organisms, i.e., yeast and human. We have also given a framework to analyze the influence of a disease gene in a complex both biologically as well as topologically in terms of eight association parameters.

2. Background

The problem of protein complex finding from PPI networks can be thought of as a machine learning problem. This can be addressed either in a supervised or unsupervised manner. Since a protein complex is a natural grouping of similar proteins, it has been handled mostly using unsupervised approaches. A number of methods have been proposed in the literature using only topological features. For example, MCODE [8] uses the concept of vertex weighting to find dense subgraphs. These dense subgraphs are referred to as protein complexes. RNSC [9] uses a cost based search technique to find protein complexes from PPI networks. CFinder [10] and CMC [11] use a clique merging procedure to identify complexes. ClusterONE [12] uses a cohesion measure to find protein complexes. Recently, researchers have started integrating GO-based knowledge for better complex detection. Some examples are GMFTP [13] which uses a propensity score to better predict the complexes. This score estimates the affinity of a protein to belong to a complex based on GO annotations. Another method called WCOACH [14], which is an improved version of COACH integrates semantic similarity between proteins to find complexes. A recent method called TINCD [15] works using the ensemble framework. It uses information from various clustering methods and raw data sources to predict quality complexes. Few researchers have used supervised techniques too to find complexes from the PPI networks. They train a model using the topological and biological properties of the available benchmark complex sets and then use it to predict complexes. Examples include a bayesian network based complex finding method [16]. Each of these methods predicts complexes to the best of their ability. But in order to analyze their performance, we need to take the help of certain standards. The commonly used indices in machine learning are precision, recall and f-measure but we do not use them here due to conflicting uses by different researchers. It is not possible that a whole set of clusters would exactly match a set of benchmark complexes, and so researchers have come up with the idea of a overlapping threshold. This value was earlier set to 0.2 [8], which then changed to 0.6 or 0.75. However, in order to maintain uniformity while considering performance using both yeast and human datasets, we use Sensitivity (Sn), Positive Predictive Value (PPV) and Accuracy (Acc) [17] as performance measures. These indices consider the number of common elements between the predicted clusters and the benchmark complex to find the overall measure and do not need any kind of overlapping threshold. Suppose there are s number of benchmark complexes and t number of predicted clusters. If C_{ij} is the number of common proteins between the i^{th} benchmark complex and j^{th} predicted cluster, Sensitivity (Sn) and Positive Predictive Value (PPV) are given as

$$Sn = \frac{\sum_{i=1}^{s} max_j\{C_{ij}\}}{\sum_{i=1}^{s} P_i} \quad PPV = \frac{\sum_{j=1}^{t} max_i\{C_{ij}\}}{\sum_{j=1}^{t} C_j} \quad Acc = \sqrt{Sn \times PPV}$$

A high sensitivity value indicates that a large fraction of proteins found in real complexes is covered by those found in the predicted clusters and a large positive predictive value indicates that a large fraction of the predicted clusters corresponds to real complexes. These two criteria should be used by any complex finding method. The values of these two can be summarized by another measure known as the Accuracy (Acc) which is the geometric mean of the two.

Proteins in a complex can be linked via different associations. These associations can be found using certain tools such as GeneMANIA [18] and STRING [19]. The associations can be physical interactions, which predict links between two proteins only if they are found to interact in some protein-protein interaction study. A link among complex members can also be due to the presence of co-expressions among their gene products, i.e,. they tend to show similar expression values across conditions in a gene expression study. Proteins in a complex may also be linked via predicted functional relationship, i.e., these proteins may be mapped to known interactions in some other organisms via orthology. There are some gene coding proteins linked via pathway information which suggests their involvement in the same reaction in the pathway. The member genes may also show co-localization among themselves suggesting their co-ocurrence in the same tissue or cell. Members within a complex may also be linked as given by some databases such as the metabolic pathway database or the protein complex database, or they may be linked w.r.t. text documents where they are assumed to be related. Thus, there may be various forms of association among members in a protein complex. Higher the number of such links in a complex, the better is its quality.

3. Method

The problem of protein complex finding is an unsupervised learning problem which involves partitioning a graph into similar natural groups. Given a graph $G = (V, E)$ corresponding to a PPI network, where V represents the set of proteins and E represents the set of edges, the task is to find a set of subgraphs such that these subgraphs closely correspond to the set of benchmark complexes. In order to describe our method, we use the following concepts.

Definition 1 (*HConfidence measure*). HConfidence measure between a pair of vertices (v_i, v_j) is given as the ratio of the common neighbors between them to its minimum connectivity. Mathematically,

$$H(v_i, v_j) = \frac{N_{v_i} \cap N_{v_j}}{min(deg_{v_i}, deg_{v_j})}$$

where N_{v_i}, N_{v_j} are the set of neighbors of v_i and v_j and deg_{v_i}, deg_{v_j} are the degrees of nodes, v_i and v_j respectively.

Definition 2 (*Seed pair*). Any two nodes or proteins, (v_i, v_j) which has the highest HConfidence measure, i.e., $H(v_i, v_j) > H(v_k, v_l)$, $\forall v_k, v_l \in \{V - \{v_i, v_j\}\}$ is a possible candidate for seed pair selection.

Definition 3 (*Connectivity*). The connectivity of a node $v_i \in V$ to a subgraph pC is defined as the ratio of the number of links, l_{v_i}, that exists between the node and the elements of the subgraph pC to the total degree of the node, i.e., deg_{v_i}.

$$Connectivity(v_i, pC) = \frac{l_{v_i}}{deg_{v_i}}$$

Definition 4 (*Semantic similarity*). The semantic similarity between a protein pair (v_i, v_j) is given by the similarity of concepts

(GO terms) with which they are associated in the corpus (GO database):

$$semsim(v_i, v_j) = sim(GOterms_i, GOterms_j)$$

where proteins v_i and v_j are associated with GO terms $GOterms_i$ and $GOterms_j$, respectively.

Definition 5 (*Reachability Index*). The reachability index of a node v_i in a cluster C is given as the ratio of the total number of links its direct neighbors have within C to the total number of edges in the cluster:

$$RI_{v_i} = \sum_{dN} \frac{l_{wC}}{tedges_C}$$

where dN is the set of direct neighbors of node v_i within cluster C, l_{wC} is the number of links each node $v_x \in dN$ has within the cluster, and $tedges_C$ is the total number of links in the cluster, C.

Definition 6 (*Contribution*). The contribution of a subgraph, say G' is the sum total of the reachability indices of all nodes $v_1, v_2 \ldots v_k$ in the subgraph.

$$Contribution(G') = \sum_{i=1}^{k} RI_{v_i}$$

Definition 7 (*Non-reachable proteins*). A pair of protein nodes (v_i, v_j) is considered as no-reachable if there is no protein v_k such that $Connectivity(v_k, (v_i, v_j)) \geqslant \alpha$ ($\alpha = 40\%$ based on emprical analysis(Fig. 1)).

Definition 8 (*Protein complex*). A subgraph $G' = (V', E')$ of G is said to be a protein complex if each $v_i \in G'$ is at least α connected to all v_j such that $v_j = \{V' - v_i\}$ and $Contribution(G'') \geqslant Contribution(G')$ where $G'' = G' \cup \{v_m\}, v_m \in v_j$ is a new candidate node to be added to G'.

Definition 9 (*Overlapped complex*). Two protein complexes $C1$ and $C2$ are said to overlap if intersection among the member elements of both the sets is non-empty i.e,.

$$PC1 \cap PC2 \neq \phi$$

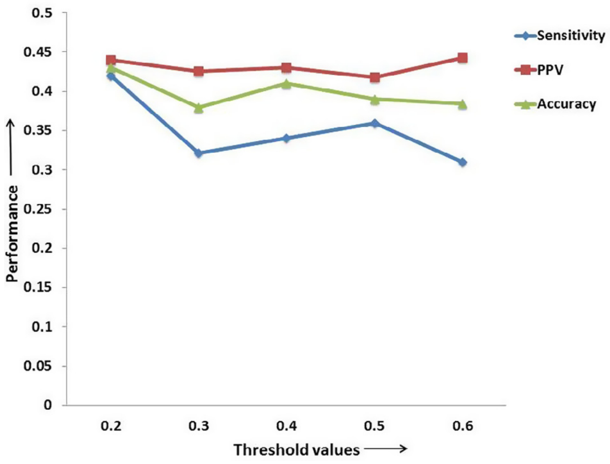


Fig. 1. Performance indices obtained at varying thresholds using HPRD dataset.

The proposed complex finding method follows the seed selection and expansion approach to extract the complexes from PPI network data. The method is called CSC as it uses the concepts of connectivity, semantic similarity and contribution during complex extraction. CSC works in four steps. The first step involves finding pairs of seed nodes from the PPI network to help form high quality clusters. Seed pair selection is done using the HConfidence score for each pair of node. At every iteration, a pair of nodes having the highest HConfidence score is chosen as the seed pair for cluster expansion. Once the seed nodes, say, (v_a, v_b) are selected, the pair is inserted into the *partialCluster*. Then the process of cluster expansion is performed in an unsupervised manner. A node v_c with the highest connectivity (among all nodes) with the *partialCluster* is chosen as the first candidate for cluster expansion. During cluster expansion, we try to make the topological and functional contribution (α and β respectively) during cluster formation to be 1, i.e.,

$$\alpha + \beta = 1 \quad (1)$$

From our experimental analysis part, we found the most suitable connectivity threshold (α) to be 0.4. This is explained by a performance graph shown in Fig. 1, which shows stable performance at around 40% connectivity in terms of all the three parameters, i.e,. Sensitivity, PPV and Accuracy. The membership of node v_c is further strengthened by the semantic similarity value existing between the nodes in the *partialCluster* and v_c. The threshold for semantic similarity (β) is accordingly adjusted to 0.6 for Eq. (1) to hold. Once, these two criteria are satisfied, it is confirmed that node v_c is a good choice both topologically and functionally to form a complex with nodes v_a and v_b present in the *partialCluster*. However, the decisive role is played by the contribution function calculated for the *partialCluster* before and after adding node v_c to it. If the value of the contribution function after new node addition is greater than the old value, only then the new node v_c is added to the *partialCluster*, else the elements in the *partialCluster* are returned as outlier proteins. This process is repeated until no further node is left satisfying all three criteria. The next complex extraction begins by choosing another pair of candidate seed nodes and the process is repeated to extract a set of complexes.

To establish the effectiveness of CSC method over other existing methods, we have given a proposition here.

Proposition 1. *The CSC method is capable of finding high quality complexes.*

Explanation: *Initially, CSC selects candidate seed pairs with the help of HConfidence measure. This measure involves choosing the best possible candidate for cluster expansion depending on their topological position in the network. Next, we use the connectivity criterion, semantic similarity value and contribution factor to determine if a new node can be inserted into the existing partialCluster. Two of these criteria, viz, connectivity and contribution are topological while semantic similarity uses corpus knowledge. This process is repeated with new seed pairs at every iteration to generate a set of clusters (complexes). These three criteria ensure the selection of an appropriate protein during expansion. Hence, the proposed CSC ensures extraction of quality complexes.* □

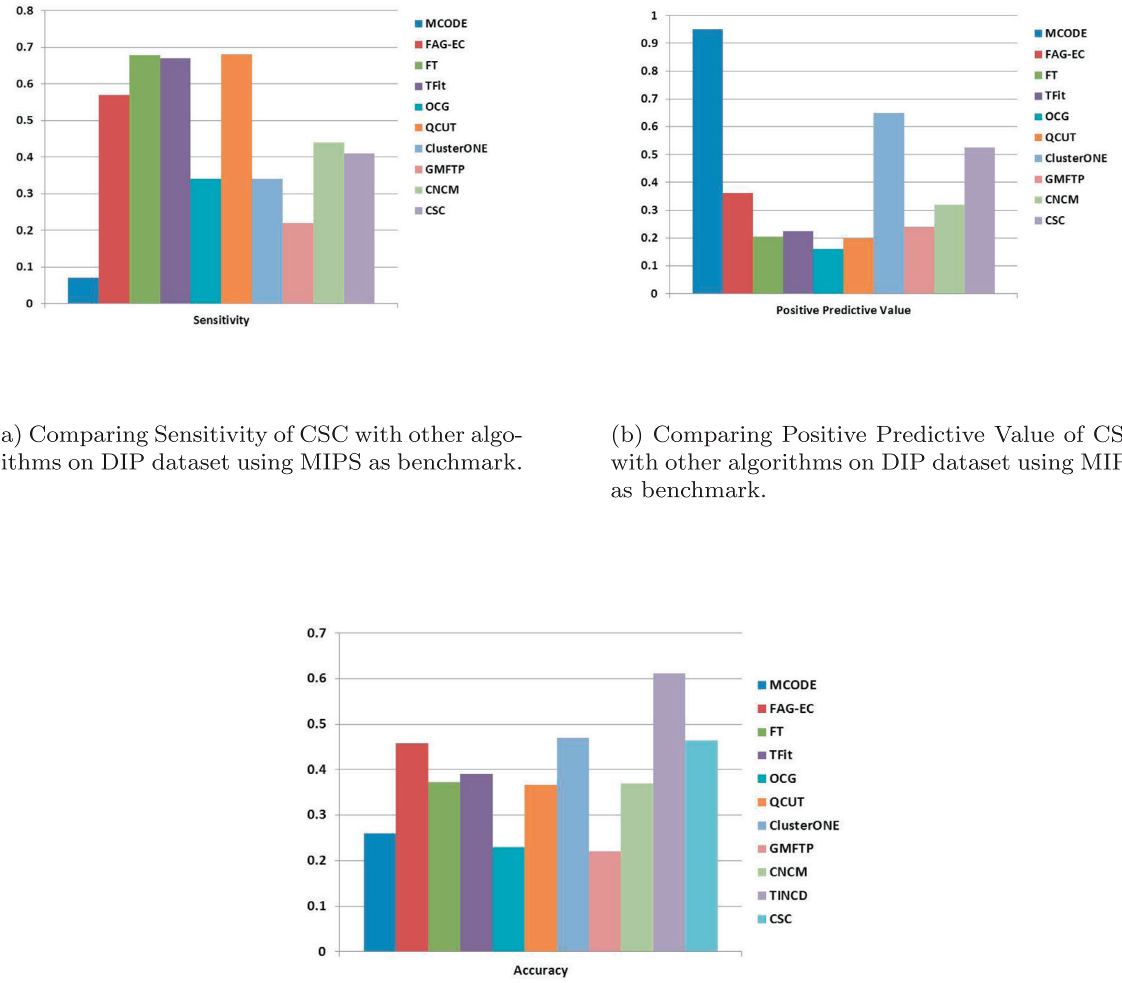
4. Performance evaluation

We now describe the environment used to implement the CSC method and the datasets used for evaluating our method. In order to evaluate our method's performance, we have used two datasets. One is the DIP dataset [20] and the other is the HPRD dataset [21]. DIP is a yeast dataset consisting of 17201 interactions and 4606 proteins whereas HPRD is a human protein interaction dataset comprising of 39209 interactions and 10080 proteins. We have also taken the help of three benchmark complexes namely MIPS [22], CYC2008 [23] and PCDq [24]. The first two are well known benchmark sets for yeast and the third one is for human. We implemented the CSC method in MATLAB running on an HP Z800 workstation with two 2.4 GHz Intel(R) Xeon (R) processors and 12 GB RAM, using the Windows 7 operating system. We have compared the performance of the CSC method with some existing methods for which Cytoscape plugins are available. For GMFTP, we use MATLAB source code provided by the authors and for CNCM [25], we use the MATLAB executable. We also use a very recent method called TINCD for comparing the accuracy of the CSC method. For this method, reported results [15] are used for comparison, as the source code could not be obtained. We have therefore, limited our comparison of CSC with TINCD only for the DIP dataset.

4.1. Results on yeast dataset

Using the above three performance indices, we analyze the performance of our method with nine other methods: MCODE [8], FAG-EC [26], FT [27], TFit [28], OCG [29]), QCUT, ClusterONE [12], GMFTP [13], and CNCM [25]. We use two benchmark complexes, viz., MIPS [22] consisting of 203 complexes and CYC2008 [23] consisting of 408 complexes. Fig. 2 show the performance of CSC with the other algorithms on the DIP dataset using MIPS as the benchmark.

Sensitivity of the CSC method is around 42%, which is better than few other methods such as MCODE, OCG, ClusterONE and GMFTP as shown in Fig. 2(a) whereas Positive Predictive Value of CSC is beaten by MCODE and ClusterONE only as seen in Fig. 2 (b). From Fig. 2(c), we see that accuracy of CSC is higher than all other methods except TINCD [15]. We could not compare our results with TINCD in terms of sensitivity and PPV as these results

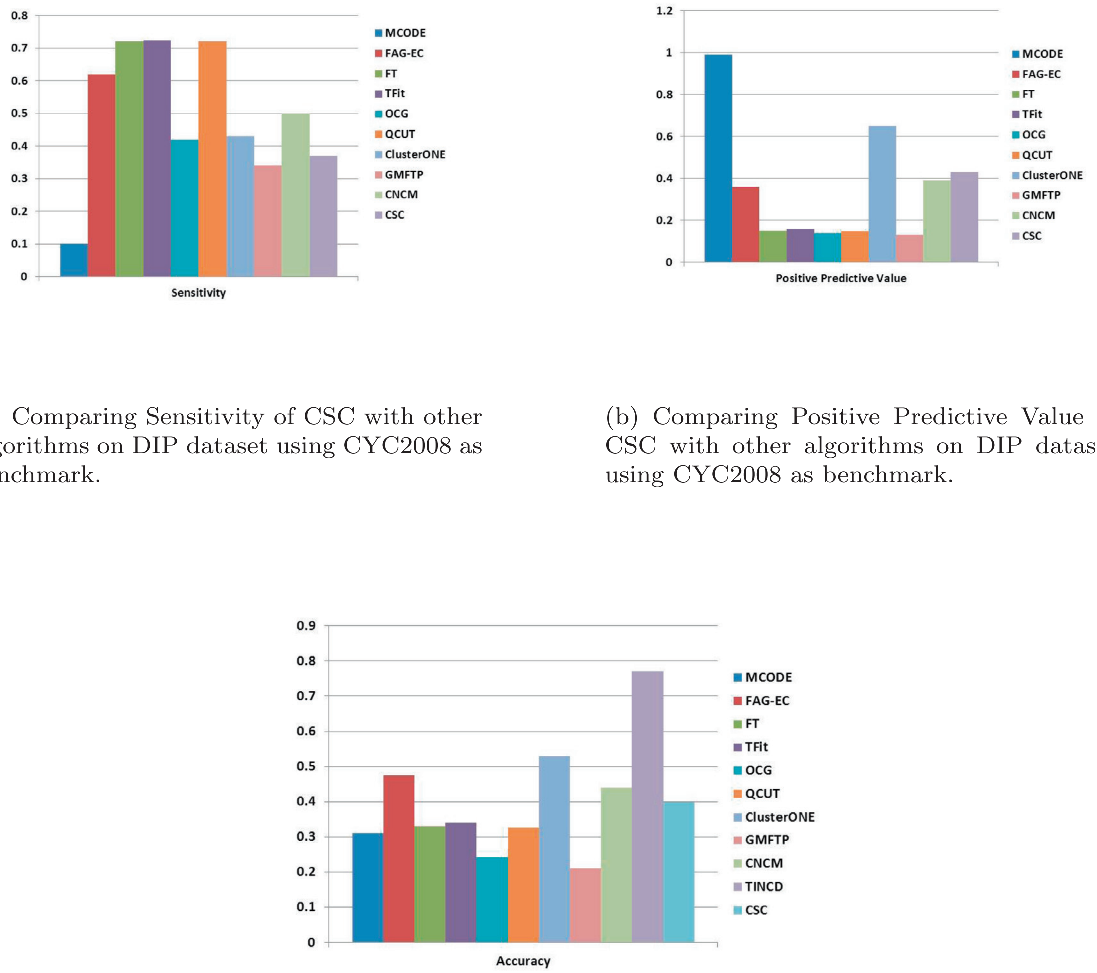


(a) Comparing Sensitivity of CSC with other algorithms on DIP dataset using MIPS as benchmark.

(b) Comparing Positive Predictive Value of CSC with other algorithms on DIP dataset using MIPS as benchmark.

(c) Comparing Accuracy of CSC with other algorithms on DIP dataset using MIPS as benchmark.

Fig. 2. Senstivity, PPV and Accuracy of CSC and other methods on DIP dataset with MIPS benchmark set.



(a) Comparing Sensitivity of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

(b) Comparing Positive Predictive Value of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

(c) Comparing Accuracy of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

Fig. 3. Sensitivity, PPV and Accuracy of CSC and other algorithms on DIP dataset using CYC2008 as benchmark.

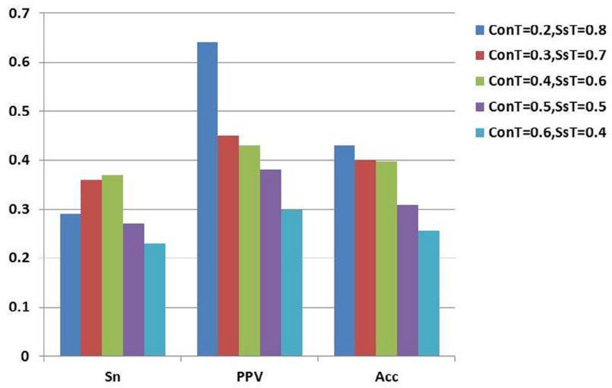
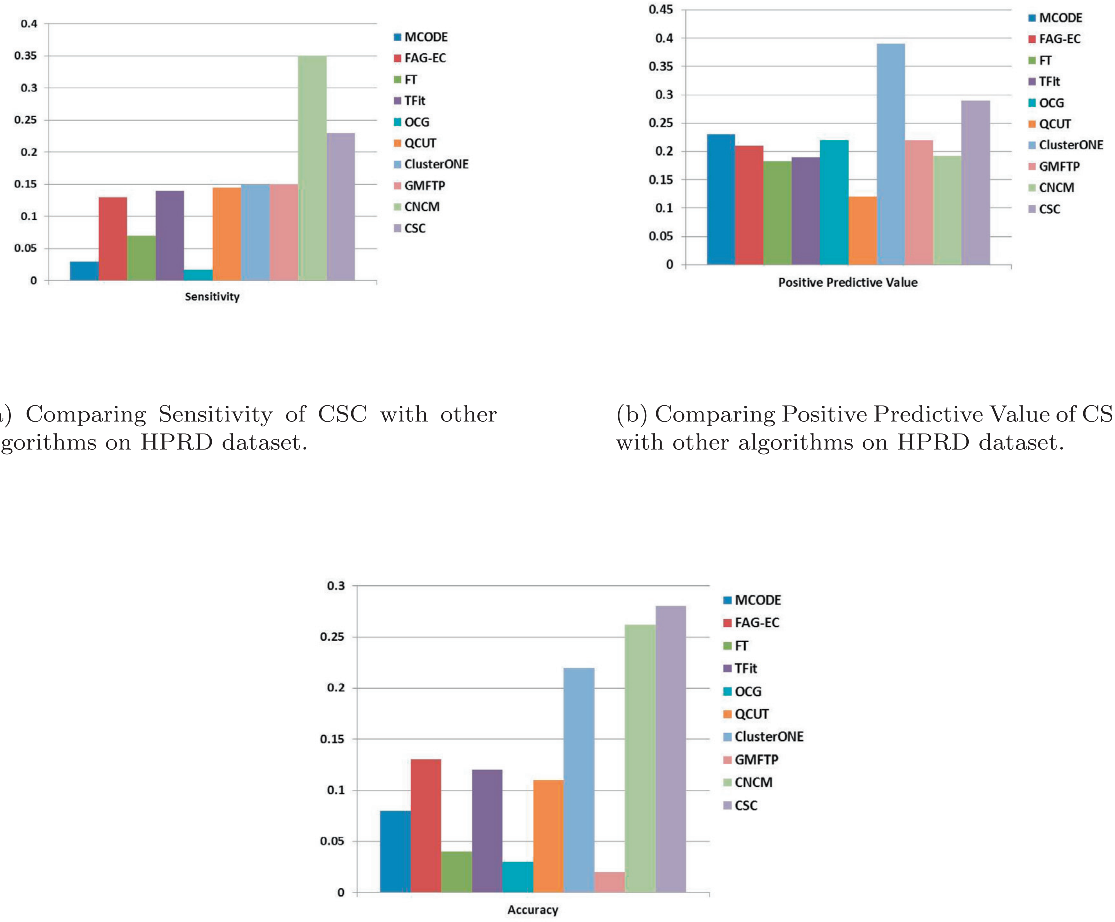


Fig. 4. Comparing Sensitivity, Positive Predictive Value and Accuracy of CSC with other algorithms on DIP dataset using CYC2008 as benchmark with varying α and β thresholds.

were not reported in the original paper [15]. It is evident from the figure that CSC gives an accuracy of 46% whereas TINCD, the most recent approach gives an accuarcy of 61% on the DIP dataset using MIPS as the benchmark.

We also use CYC2008 as the benchmark dataset for comparing the performance of our method. Fig. 3 shows the performance on DIP dataset using CYC2008 as the benchmark dataset. In Fig. 3(a), we see that the sensitivity of the CSC method is quite less as compared to other methods except MCODE and GMFTP. The PPV of CSC is in the third position for this benchmark set with MCODE and ClusterONE occupying the first and second place as seen in Fig. 3 (b). The accuracy of our method is around 40%, whereas two other methods-ClusterONE and TINCD show an accuracy 0f 50–70% as seen in Fig. 3(c). We can fine tune these performance values by tuning the SsT/β threshold. This is justified by Fig. 4. However, we have used $\alpha = 0.4$ and $\beta = 0.6$ for our computation as suggested in Section 3.



(a) Comparing Sensitivity of CSC with other algorithms on HPRD dataset.

(b) Comparing Positive Predictive Value of CSC with other algorithms on HPRD dataset.

(c) Comparing Accuracy of CSC with other algorithms on HPRD dataset.

Fig. 5. Sensitivity, Positive Predictive Value and Accuracy of CSC with other algorithms on HPRD dataset.

Our method performs significantly well over other methods as can be seen in case of the DIP dataset using MIPS benchmark. Although, it could not beat TINCD and few other methods in case of the CYC2008 benchmark dataset with the used parameters, we can still justify scope of improvements in these indices by tuning the parameters.

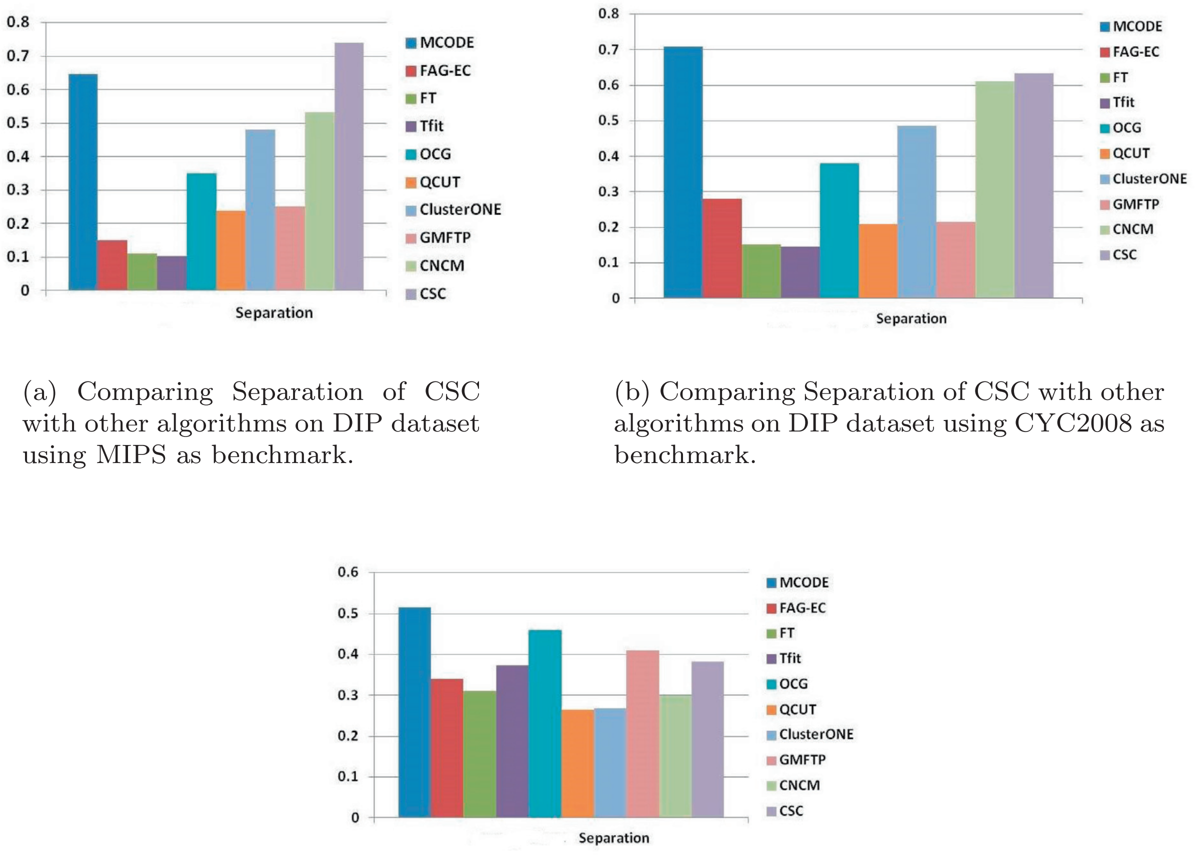
4.2. Results on HPRD dataset

We analyze the performance of the CSC method using the bigger HPRD dataset [21], which is the Human Protein Reference Dataset comprising of 39,209 interactions. Literature [30,7,31,6] has shown that the knowledge of protein complexes can be used in disease diagnosis, so we are keen to analyze the accuracy of our method over the human dataset. A more accurate method would aid the biomedical scientists in developing a better understanding of complexes and would prove helpful in finding their association with diseases. We compare the performance values of the CSC method with nine other methods: MCODE, FAG-EC, FT, TFit, OCG, QCUT, ClusterONE, GMFTP and CNCM as shown in Fig. 5.

As seen in Fig. 5(a), the sensitivity of CSC is around 23%, which is much higher than other methods except CNCM. The PPV of our method is at the second position after ClusterONE (Fig. 5(b)). From Fig. 5(c), we see that accuracy of our method emerges as the winner in this dataset.

Apart from the three performance indices, we have also used another measure called separation, which gives the isolation factor within the predicted complexes. It is given as the product of the fraction of complexes obtained in the predicted cluster with that of the fraction of predicted elements found in the complex. A higher value of separation indicates a two-way correspondence between the predicted clusters and the complexes. Fig. 6 shows the separation obtained from CSC and other methods using DIP and HPRD dataset.

From Fig. 6(a), it can be seen that CSC shows better separation value than other methods in case of MIPS whereas it is in the second position using CYC 2008 dataset as seen in Fig. 6(b). It is mainly because of the smaller number of larger complexes generated by MCODE in comparison to ours. From Fig. 6(c), it is seen that CSC is at the fourth position in terms of HPRD dataset. The main reason behind this reduced separation value is the occurrences of high overlaps among the complexes extracted by our method. As stated in [12], proteins tend to perform multiple functions and hence are usually grouped into multiple complexes. For example,



(a) Comparing Separation of CSC with other algorithms on DIP dataset using MIPS as benchmark.

(b) Comparing Separation of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

(c) Comparing Separation of CSC with other algorithms on HPRD dataset

Fig. 6. Separation of CSC and other algorithms on DIP and HPRD dataset.

the CYC 2008 complex set has around 207 proteins which participate in more than one complex. The role of overlapping protein complexes has also been established in [12].

5. Analyzing PPI complexes: a conceptual framework

In this section, we will demostrate the effectiveness of CSC from both biological as well as topological points of view w.r.t. a given disease query. A protein complex is nothing but a group of similar proteins which collaboratively participate in rendering significant cellular functions. As the literature points out [32], any mutation in genes coding proteins in a complex may lead to certain diseases. For example, the SWI/SNF complex is known to be associated with Coffin-Siris syndrome and plays a role in causing cancer [6]. We analyze a subset of complexes based on some query diseases. In order to find this subset of complexes, we use the disease related gene information given in Genecard [33]. We directly use gene names given in GeneCard as there is one-to-one correspondence between genes and proteins are named the same way as genes [34]. We use only a single disease, so the number of disease genes found from GeneCard is not too high. However, if we had chosen a whole class of diseases, the number of genes would be large and as a result, the identification of disease associated complexes would likely be a lengthy process. In order to handle such a scenario, we propose a framework.

5.1. The disease gene-central gene analysis framework

This section presents a conceptual framework as shown in Fig. 8 to analyze the associations of a disease gene with the central gene (s) (chosen to represent a complex based on connectivity) of complexes. The following definition and illustration are useful for further description of the application.

Definition 10 (*Central gene of a complex*). A gene g_i is referred to as a central gene of a complex C_i, iff the associations of g_i with rest of the genes $g_j \in C_i$ is highest.

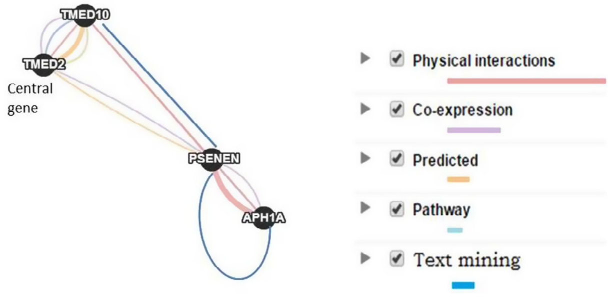


Fig. 7. Protein complex members along with its association links.

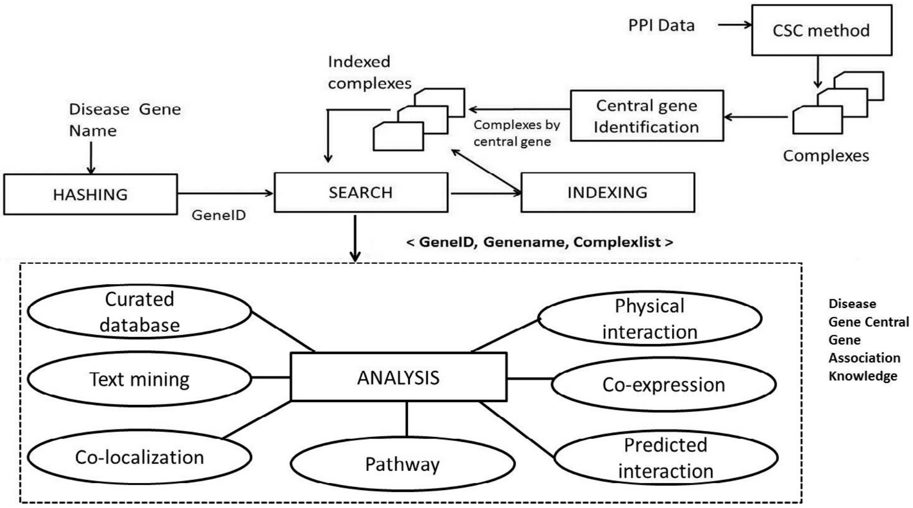


Fig. 8. Disease gene-Member genes analysis framework.

For illustration, an example graph shown in Fig. 7, representing a complex given by CSC is used. Here the nodes represent the genes and the edges represent the associations, which may be of seven distinct types viz., (i) physical interaction, (ii) co-expression, (iii) predicted interaction, (iv) pathway, (v) co-localization, (vi) curated database and (vii) text mining. However, for this complex, only five types of associations are present as shown in the figure.

From the set of complexes given by CSC, we identify the central gene representing each complex. Identification of the central gene is important in order to understand the association of the disease gene with the central gene in the complex. In order to reduce the time taken during string comparison for finding disease associated complexes, we map the disease genes to unique numbers by means of a hashing technique (one-to-one mapping). This unique number is referred to as the GeneID. Using an index search, we identify the disease associated complexes quickly. This process outputs results of the form < *GeneID, Genename, Complexlist*>. The *Complexlist* is a set of those indexed complexes which have *GeneID* as one of the members. This list is dynamic in nature as one disease gene can be present in one or more complexes. Once the *GeneID* along with the *Complexlist* is obtained, we can further analyze the associations of central gene(s) and other genes with the disease gene(s).

To support our analysis, we use two online tools, GeneMANIA [18] and STRING [19]. GeneMania is a web based tool which features many functions such as analyzing a gene list, prioritization of genes and determining gene functions. A very useful function of this tool is the visual representation of a set of genes. This graphical representation has nodes which correspond to genes and edges which correspond to attributes such as (i) physical interactions, (ii) co-expressions, (iii) predicted interactions, (iv) pathways, (v) co-localization, (vi) genetic interactions and (vii) shared protein domains. We use the first five attributes i.e., (i)-(v) for our purpose. The other two options are not used as they mainly focus on the 3D-structure of proteins, which is beyond the scope of this work. We use another tool called STRING (Search Tool for the Retrieval of Interacting Genes), which is an online database resource for annotating functional interactions among proteins. This tool also gives a visual representation of genes in a network with edges corresponding to known interactions, corresponding to those experimentally determined and those which are obtained from curated databases. It also predicts interactions, if at all, they exist using neighborhood information or co-occurrence information among the genes. Moreover, it also shows edge information obtained using *text mining* from different literature sources and from *homology* considerations. Among all these attributes, we use only (i) edge information from curated databases and (iii) text mining for our purpose.

5.2. An application to Alzheimer's disease

In this section, we consider an example disease query from the class of neuro-degenerative diseases for analysis of complexes given by CSC. Among all forms of mental illnesses, Alzheimer's Disease is devastatingly common. It is the sixth leading cause of death, especially among the elderly. Although there has been significant development in drug design to protect people from this deadly disease, effective treatment of this form of dementia does not exist. Therefore, PPI data analysis w.r.t. such a disease is considered a critical research problem for bioinformaticists.

The use of a series of criteria in the CSC method have lead to a reduced search space for the formation of clusters (complexes). Due to this constraint, we find only two complexes associated with the disease. We analyze the members of these two complexes using the tools discussed above. The two disease associated complexes along with their member proteins and associations among them are given in Table 1. In Table 1, columns 5–11 show the association of the disease gene with the central gene as well as other

Table 1
Alzheimer associated complex (Association of disease gene with other genes in the complex).

S. No	Disease gene(s) in complex	Members of complex (except disease gene)	Whether Central gene	Physical interaction	Co-expression	Predicted interaction	Path way	Colocalization	Curated database	Text mining
1	PSENEN	APH1A	No	Yes	Yes	No	Yes	No	No	Yes
		TMED2	Yes	No	Yes	Yes	No	No	No	No
		TMED10	No	Yes	No	No	No	No	No	Yes
2	TOMM40	TOMM22	No	Yes	Yes	No	No	No	Yes	Yes
		TOMM7	Yes	Yes	Yes	No	No	No	Yes	Yes

Table 2
Pathway associated with each member of Alzheimer associated complexes.

S.No	Complex members	Whether disease gene	Pathway in which involved	Percentage of match (belonging to same pathway)
1	PSENEN	Yes	Notch signaling pathway	
	APH1A	No	Notch signaling pathway	
	TMED2	No	Pre-notch expression and processing	75
	TMED10	No	mRNA processing	
2	TOMM22	No	Mitochondrial protein import	
	TOMM7	No	Mitochondrial protein import	100
	TOMM40	Yes	Mitochondrial protein import	

complex members w.r.t. our seven chosen attributes. Another significant characteristic of genes is determined by the pathways in which they are involved during any cellular activity. Pathway information can be used for analyzing the contribution of each member within a complex. Two genes belonging to the same pathway are functionally more similar than those belonging to different pathways [35]. Table 2 gives the pathways with which each member of the two disease associated complexes is associated.

We observe in Table 2 that 75% similarity is seen in the pathway information in complex 1 and 100% similarity in pathway is observed in complex 2. Therefore, we can say that CSC is able to extract high quality complexes both from statistical and biological points of view.

6. Conclusion and future work

In this work, we present a method which gives more accurate results during the protein complex finding process. The accuracy obtained using our method is at par with the existing methods such as MCODE, FAG-EC, FT, TFit, OCG, QCUT, ClusterONE, GMFTP, CNCM and TINCD in case of DIP dataset. In case of human dataset, our method gives the best performance against the other existing methods in terms of accuracy. We establish the biological significance of our method empirically. We also introduce a conceptual framework to analyze the associations of complex members with the disease gene w.r.t. eight significant parameters. Although the framework introduced supports analysis for a neuro-degenerative disease, it can be extended for other diseases as well. A relational database tool is being developed to support a large number of disease queries related to such diseases.

Conflict of interest

The authors have no conflict of interest.

Acknowledgement

This work was supported by Centre Of Excellence under FAST (Dy. No. 394/14 IF.I dated 26.08.2014) and SAP DRS II of UGC (Dy. No. 28796 dated 15.04.2015).

References

[1] Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415(6868):141–7.
[2] Moschopoulos CN, Pavlopoulos GA, Iacucci E, Aerts J, Likothanassis S, Schneider R, et al. Which clustering algorithm is better for predicting protein complexes? BMC Res Notes 2011;4(1):1.
[3] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinform 2006;7(1):207.
[4] Quinn J, Fyrberg AM, Ganster RW, Schmidt MC, Peterson CL. DNA-binding properties of the yeast SWI/SNF complex. Lett Nat.
[5] Safari-Alighiarloo N, Taghizadeh M, Rezaei-Tavirani M, Goliaei B, Peyvandi AA. Protein-protein interaction networks (PPI) and complex diseases. Gastroenterol Hepatol from Bed to Bench, vol. 7, 1.
[6] Chen Y, Jacquemin T, Zhang S, Jiang R. Prioritizing protein complexes implicated in human diseases by network optimization. BMC Syst Biol 2014;8(Suppl 1):S2.
[7] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLOS Comput Biol 2010;6(1):e1000641.
[8] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform 2003;4(1):1.
[9] King AD, Pržulj N, Jurisica I. Protein complex prediction with RNSC. Bacter Mol Networks: Methods Protoc 2012:297–312.
[10] Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 2006;22 (8):1021–3.
[11] Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics 2009;25(15):1891–7.
[12] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods 2012;9(5):471–2.
[13] Zhang X-F, Dai D-Q, Ou-Yang L, Yan H. Detecting overlapping protein complexes based on a generative model with functional and topological properties. BMC Bioinform 2014;15(1):1.
[14] Kouhsar M, Zare-Mirakabad F, Jamali Y. WCOACH: protein complex prediction in weighted PPI networks. Genes Genet Syst 2015;90(5):317–24.
[15] Ou-Yang L, Wu M, Zhang X-F, Dai D-Q, Li X-L, Yan H. A two-layer integration framework for protein complex detection. BMC Bioinform 2016;17(1):1.
[16] Shi L, Lei X, Zhang A. Protein complex detection with semi-supervised learning in protein interaction networks. Proteome Sci 2011;9(1):1.
[17] Brohee S, Van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinform 2006;7(1):1.
[18] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucl Acids Res 2010;38(suppl 2):W214–20.
[19] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucl Acids Res 2011;39(suppl 1): D561–8.
[20] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucl Acids Res 2004;32(suppl 1): D449–51.
[21] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13 (10):2363–71.
[22] Mewes H-W, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucl Acids Res 2004;32(suppl 1):D41–4.
[23] Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucl Acids Res 2009;37(3):825–31.
[24] Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, 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(2):1.
[25] Sharma P, Ahmed HA, Roy S, Bhattacharyya DK. Detecting protein complexes using connectivity among nodes in a PPI Network. Network Model Anal Health Inform Bioinform 2015;4(1):1–17.
[26] Li M, Wang J, Chen J. A fast agglomerate algorithm for mining functional modules in protein interaction networks. In: 2008 International conference on biomedical engineering and informatics, Vol. 1, IEEE; 2008. p. 3–7.
[27] Ruan J, Zhang W. Identifying network communities with a high resolution. Phys Rev E 2008;77(1):016104.
[28] Gambette P, Guénoche A. Bootstrap clustering for graph partitioning. RAIRO-Oper Res 2011;45(4):339–52.
[29] Becker E, Robisson B, Chapple CE, Guénoche A, Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 2012;28(1):84–90.

[30] Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 2007;25(3):309–16.

[31] Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. Nat Biotechnol 2006;24(4):427–33.

[32] Ganegoda GU, Wang J, Wu F-X, Li M. Prediction of disease genes using tissue-specified gene-gene network. BMC Syst Biol 2014;8(Suppl 3):S3.

[33] Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, et al. MalaCards: an integrated compendium for diseases and their annotation. Database 2013; 2013, bat018.

[34] Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO gene nomenclature committee (HGNC). Human Genet 2001;109(6):678–80.

[35] Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11(12):843–54.