J Biosci Vol. 42, No. 3, September 2017, pp. 383–396
DOI: 10.1007/s12038-017-9696-3

Protein complex finding and ranking: An application to Alzheimer's disease

Pooja Sharma[1], Dhruba K Bhattacharyya[1,*] and Jugal K Kalita[2]

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

[2] *Department of Computer Science, University of Colorado, Colorado Springs, CO, USA*

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

MS received 10 December 2016; accepted 24 May 2017; published online 7 July 2017

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 and a few other model organisms. Such protein complex identification methods, when applied to large human PPIs often give poor performance. We introduce a novel method called ComFiR to detect such protein complexes and further rank diseased complexes based on a query disease. We have shown that it has better performance in identifying protein complexes from human PPI data. This method is evaluated in terms of positive predictive value, sensitivity and accuracy. We have introduced a ranking approach and showed its application on Alzheimer's disease.

Keywords. Connectivity; Disease gene; Protein complex; Relevance score

1. Introduction

Proteins are at the core of life on earth. They are known to be the workhorse for every activity occurring in the living cell. A protein is known to coordinate with other proteins to carry out important functions. This coordination creates interactions known as protein–protein interactions (PPIs). A group of interactions leads to specific biological functions at both cellular and system levels, giving rise to stoichiometrically stable compounds known as protein complexes (Bader and Hogue 2003). These complexes assist in cell functions such as cell growth (Szymanski 2005), physiology (Yi and Deng 2005) and metabolism (Winkel 2004). For example, VGII1-TEAD4 complex enhances expression of *IGFBP5* gene, which is responsible for promoting cell proliferation in humans (Pobbati *et al.* 2012). Thus, in order to understand the dynamics of living beings, it is essential to study such compounds. Protein complexes are reciprocal functional units in a protein–protein interaction network (PPIN) (Kashyap et al. 2016). This reciprocality produces robustness of proteins against mutation (Erten *et al.* 2009). In terms of topology, reciprocality (modularity) results in groups of proteins which are densely connected among themselves.

A number of techniques are available for mining protein complexes as dense subgraphs from a PPIN. Some methods such as MCODE (Bader and Hogue 2003), DPClus (Li *et al.* 2008a) and LCMA (Jung *et al.* 2010) solely use statistical properties of the PPIN to identify complexes. MCODE uses a three-step procedure for finding complexes, although the major step involves a vertex weighing technique. A slightly modified version of MCODE, known as DPClus, uses a different vertex weighing strategy to identify complexes. Another method, known as MCL (Van Dongen 2001), uses the concept of random walk to identify protein complexes. It was realized that proteins exhibit added functionality when forming complexes and these methods identified exclusive complex pairs. Subsequently, the non-exclusive complexes (overlapping of one partner in a given pair of complexes) were studied. Several methods which can satisfy this non-exclusiveness criterion for clustering on PPI data are known at this time, including hierarchical agglomerative clustering with overlaps (HACO) (Wang *et al.* 2009), OCG (Becker *et al.* 2012) and ClusterONE (Nepusz *et al.* 2012). HACO is an extended version of canonical hierarchical agglomerative clustering (Kaufman and Rousseeuw 2009) which can effectively find overlapping protein complexes. ClusterONE is the most

Electronic supplementary material: The online version of this article (doi:10.1007/s12038-017-9696-3) contains supplementary material, which is available to authorized users.

competent method at present. It works more or less like the MCODE using a seed expansion procedure except that the expansion is governed by a cohesiveness measure, which indicates the compactness within the cluster as compared to the rest of the network. A few other methods such as CORE (Leung *et al.* 2009), COACH (Wu *et al.* 2009) and MCL-CAw (Srihari *et al.* 2010) make use of additional biological information during complex finding, while there are methods like RNSC (King *et al.* 2004), DECAFF (Li *et al.* 2007) and PCP (Chua *et al.* 2008) which use functional information during the complex identification process. An empirical study was carried out to analyse the performance of some state-of-art methods over the yeast dataset (Sharma *et al.* 2015). All these methods rooted to a particular ground, i.e. their effectiveness, was evaluated on the yeast dataset, but as human PPI information is increasingly becoming available, efforts have moved towards highlighting their performance on the human PPI dataset. The most significant difference between the results on human and yeast is that none of the methods can produce results better than that of yeast (Wu *et al.* 2013). A good complex finding method enables identification of overlapped protein complexes with high functional coherence. However, from the literature, we see that the best precision attained on HPRD dataset (Prasad *et al.* 2009) by ClusterONE is lower than 25% (established experimentally using ClusterONE plugin). Protein complexes are involved in a variety of activities occurring in the living cell. Slight changes in the nucleotide base sequence of amino acids (which combine to form proteins) affect the protein formation process. Mutation in gene coding proteins known to form protein complexes tend to disrupt the whole functional nature of the compound (Rao *et al.* 2013). Many of such mutations give rise to the formation of such complexes which may be linked to one or more diseases. Hence, identifying such complexes is an important extension to protein complex finding. Many researchers have worked in this direction to rank (prioritize) disease-associated protein complexes. In the study by Vanunu *et al.* (2010), the prioritization process is based on the formation of protein complexes (the member proteins are individually ranked first depending on their associative score in causing a particular disease). These complexes are evaluated in terms of functional, expression and conservation coherency (Vanunu *et al.* 2010). Another method called MAXCOM (Chen *et al.* 2014) uses the concept of maximum information flow to prioritize the candidate protein complexes with respect to a given query disease using the information from a heterogeneous network made up of disease-phenotypic similarities, disease–protein links and PPIs. A much more recent method called NBH (Le 2015) prioritizes diseased candidate protein complexes from a protein complex network using a similarity measure. This protein complex network is built using the concept of functional similarity, where two complexes are connected if they either share protein elements or GO terms or are connected by protein interactions. Thus, an appropriate ranking scheme, which is unbiased and easy to use, can help biologists to understand the relevance of a complex for a given disease query. It also helps identify the relationship of disease genes belonging to a given complex with other participating genes in the same complex.

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, to support biologists, PPI data analysis with respect to a given disease such as Alzheimer's is considered a critical research problem for bioinformaticians. This is our motivation to work on the problem from PPI complex finding and ranking points of view.

Protein complex detection helps decipher new/uncharacterized proteins from the functions of other proteins in the complex to which this new protein belongs. In order to effectively interpret the function of such proteins, the protein complexes need to be of high functional relevance. Most existing protein complex prioritization methods are highly sensitive to the protein complexes used prior to the ranking process and a stable complex finding method which could predict biologically significant complexes is still a distant reality (Sharma *et al.* 2015). Thus, the need is to find an effective protein complex detection method which could predict the complexes with high precision. Once highly relevant complexes are found, their association with various diseases can be precisely understood and prioritization would then be more meaningful to the biologists. They can use the information about the complex to design target drugs.

In this article, we focus on finding protein complexes by incorporating some knowledge from already established protein complexes during the initial phase for seed selection. Thereafter, the expansion of seed pair is totally unsupervised using topological and biological features, which leads to the formation of clusters (protein complexes). We ensure that the proposed method detects protein complexes with high biological significance. For a fair comparison, we report the performance of ComFiR and a few other existing methods on the human PPI dataset. We also present a way of ranking these complexes with respect to a given query disease and attempt to support our approach using evidence from valid sources. The following section describes our protein complex finding method followed by our ranking method in detail. The complexity analyses of both these methods are reported in section 3 of supplementary material. The next section then discusses experimental results and finally in the last section, we give the concluding remarks and future direction of research.

2. Method

The objective of our method is to initially identify protein complexes in close proximity to benchmark complexes for the human PPI dataset, and finally to rank these complexes based on their relevance with respect to a given query disease. To achieve the initial objective, we use some amount of prior knowledge in the beginning of our method during the starting of the process. Once the seed pair is decided, we opt the unsupervised approach for the cluster expansion. The following graph theoretic concepts and definitions are useful in describing our method. For a PPI network defined by $G = (V, E)$, where V represents the set of vertices (proteins) and E represents the edges between them, we define the following terms:

Definition 1 The seed pair to form a complex is a pair of non-selected nodes or proteins $(v_{i,}\ v_j)$ which co-occur in benchmark complexes with significantly high frequency.

Definition 2 Protein benchmark complex matrix is a binary matrix referred to as *PcM*, where entries are computed as follows:

$$\mathrm{PcM_{i,j}} = \begin{cases} 1, & \text{if ith protein is present in the jth benchmark complex} \\ 0, & \text{otherwise} \end{cases}$$

Definition 3 The density of a subgraph G' of G, i.e. $den(G')$, is the ratio of the total number of actual links in G', i.e. $|e_d|$ to the number of maximum possible links, i.e. $|E_d|$ Mathematically,

$$den(G') = \frac{|e_d|}{|E_d|}$$

Definition 4 The connectivity of a node v_i of a subgraph or a *partialCluster* is defined as the ratio of l_i, the number of links v_i possesses with the *partialCluster* and t_i, the total number of links, the node v_i has with the whole network, V. Mathematically,

$$conn(v_i, partialCluster) = \frac{l_i}{t_i}$$

Definition 5 A pair of proteins or nodes (v_i, v_j) is declared an outlier if there is no protein v_k such that $conn(v_k, (v_i, v_j)) \geq$ some connectivity threshold set by the user (default value taken to be 0.4).

Definition 6 A subgraph G' of G is defined as a protein complex if each $v_i \in G'$ satisfies the connectivity threshold (as in Definition 4) and semantic similarity among proteins in $G' \geq SsT$, SsT is the semantic similarity threshold set by the user.

Definition 7 Two protein complexes C_1 and C_2 are said to overlap if and only if $C_1 \cap C_2 \neq NULL$ or $\exists\ v_i \in C_1$ such that $v_i \in C_2$ also.

The following sections describe our method ComFiR using above definitions.

2.1 *Protein complex finding*

We initiate the complex finding process with a pair of proteins which occurs the highest number of times among all the complexes as seed. This is the only phase where we have used the information from existing complexes. The framework of ComFiR is shown in figure 1.

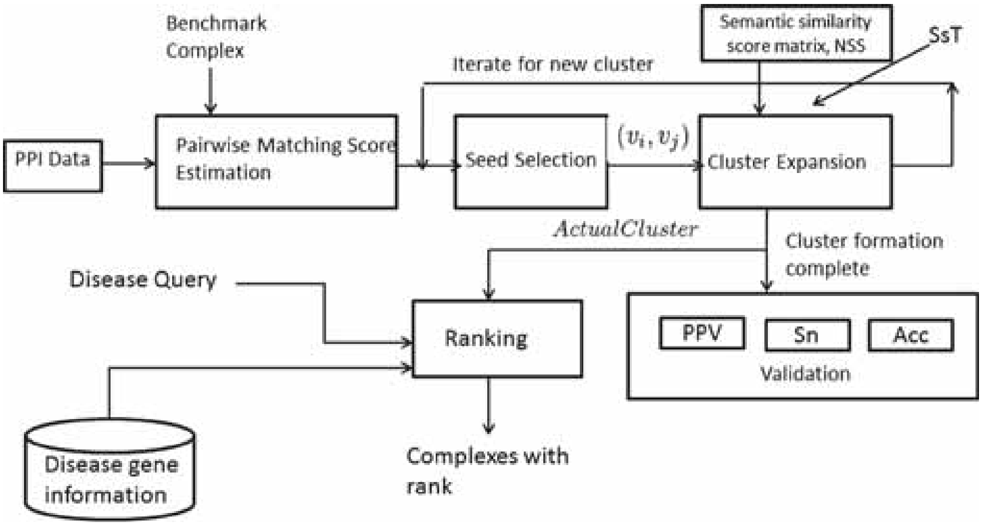

Figure 1. Framework of ComFiR.

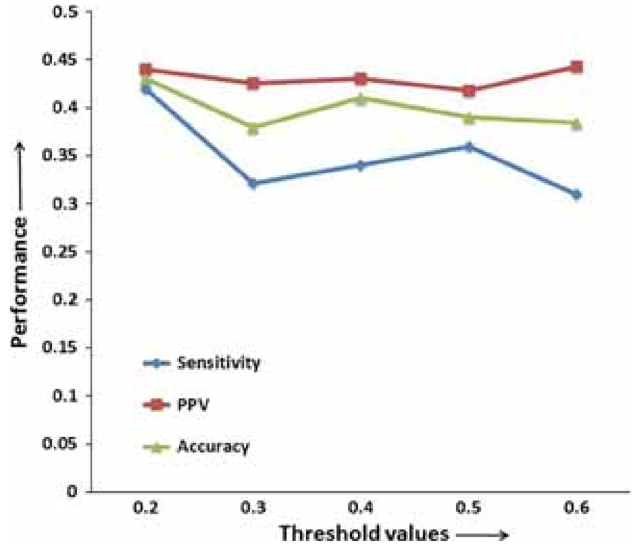

Figure 2. Performance indices obtained at varying thresholds.

To support seed selection, we use the *PcM* matrix as defined in Definition 2. To illustrate the seed selection process, a working example is given here.

Example 1 Suppose, proteins *a* and *b* occur in complexes C_1, C_2 and C_3. Also, assume that protein *c* occurs only in complex C_2. In this case, the *matching scores* of the combination of seeds is as follows.

$$m_{sa,b} = 3, m_{sa,c} = 1 \text{ and } m_{sb,c} = 1.$$

Since the matching score of *(a, b)* is the highest, we start the complex finding procedure with the pair *(a, b)* as the seed nodes.

With the selected pair of seed proteins, the cluster expansion process is started one at a time. This expansion of cluster from the seed pair is guided only by the topological and biological characteristics of the PPI network. The expansion process iterates over nodes satisfying the connectivity criterion of being more than 40% linked. The connectivity threshold is set to 40–50% based on experimental evidence that establishes it as the optimal value. This can be seen in figure 2. We carry out the cluster expansion process for varying thresholds from 0.2 to 0.6. In the figure, we can see that the performance indices show stability at the threshold value of 0.4. Hence, we fix this as a static parameter and proceed with the addition of further information into the cluster expansion process. In order to get biologically enriched clusters, we further constrain cluster expansion by checking if the new node has a semantic similarity score greater than that of the threshold, *SsT* set by the user. This value is the maximum value of all semantic scores the upcoming node shares with elements in the *partialCluster* one at a time. If such a node exists, it is inserted into the *partialCluster* and a new node with more than 40% connectivity is again sought to continue the process. However, if no node exists to further expand the *partialCluster* by more than two elements, these elements are returned as outlier proteins.

Table 1. Symbols used

Symbols	Interpretation
PPIN	A graph corresponding to a PPI Network with *V* vertices and *E* edges
MMS(v_i, v_j)	Maximum Matching Score between a pair of proteins v_i and v_j
conn(v_i, partialCluster)	Connectivity between v_i and *partialCluster*
NSS	Semantic similarity score matrix
SsT	Semantic similarity score threshold
Cluster	Set of clusters
ActualCluster	Set of clusters after eliminating redundant clusters
D_g	Set of disease genes
DGC	Disease gene association network (weighted)
SC_j	Relevance score of complex C_j

Once the expansion criterion is violated, we start with another set of seed nodes and repeat the process. This is done for each potential seed node and in the process we get a number of clusters *Cluster,* referred to as protein complexes. Our method is given in Algorithm 1 and the symbols used are described in table 1.

2.2 *Parameter tuning*

Choosing a set of parameters needs careful performance analysis. From figure 2, we see that the performance at 0.2 threshold is better than that at 0.4. This is because if we take just 20% connectivity among nodes, then obviously compact clusters would be generated. But in order to nullify this, we have opted for 40% connectivity threshold. Similarly, if we keep the semantic similarity threshold high, then the cluster generated would be more functionally similar. This is why the sensitivity, positive predictive value and accuracy is higher as we increase the threshold in the x-axis (figure 3). In order to get good quality clusters without compromising with the effect of semantic similarity threshold, we have chosen to report the results at threshold of 0.4.

2.3 *Removal of redundant clusters*

There may be cases when the same cluster set is generated with different starting seed nodes, in which case, it would unnecessarily add to the precision value. We take special care to eliminate redundant clusters and irrelevant clusters with size less than 3. Redundancy can be of two types, either the whole set of cluster is matched with another cluster set generated from different pair of seed nodes or there is very high percentage of overlap between two cluster sets. In the

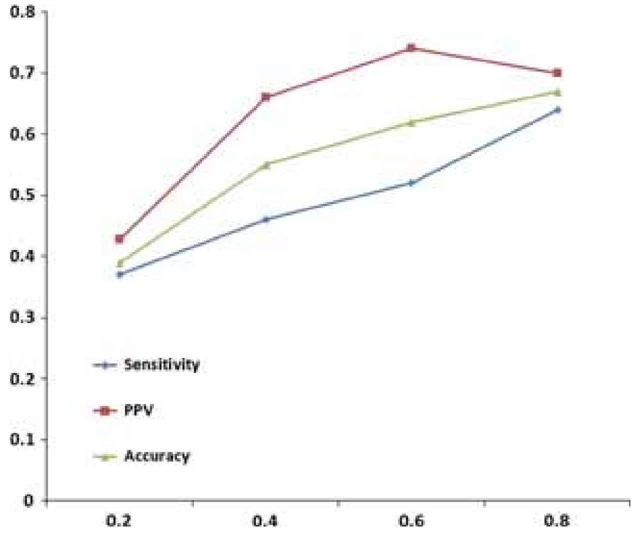

Figure 3. Performance indices of ComFiR obtained at varying thresholds of *SsT*.

first case, if we encounter any cluster with all similar elements, we simply discard such occurrences. We maintain a track of unique clusters in *ActualCluster*. In order to handle the second scenario, where there may be very high overlap between two clusters, we choose an overlapping threshold given by the formula

$$Ov_{th} = \frac{|C_i \cap C_j|^2}{|C_i||C_j|}$$

The value of this Ov_{th} is chosen to be 0.8 as suggested in Nepusz *et al.* (2012). This threshold value suggests 80% overlap between members of two clusters. For clusters, where $Ov_{th} \geq 0.8$, we discard the smaller clusters and take only the larger ones into the *ActualCluster* set. In this way, we get a set of unique clusters from the process.

We present the following proposition to establish that ComFiR is able to detect overlapping complexes.

Proposition 1 *ComFiR detects overlapping protein complexes.*

Explanation ComFiR initiates expansion of a complex with a pair of seed nodes (v_i, v_j) which has the highest frequency of co-occurrence among all benchmark complexes. The process starts with these nodes and expands depending on the connectivity between elements in *partialCluster* and the rest of the nodes. A node v_k gets added to the *partialCluster* only if $conn(v_k, partialCluster) \geq 40$ %, else the expansion process is terminated and a new pair of seed nodes is chosen to begin complex formation. The selected node is further refined using a constraint on its functional similarity, which is given by the user. For a node to be included into the *partialCluster*, it must satisfy both conditions. Assume that a node $\mathrm{v_k}$ is a member of a complex formed with an initial seed pair (v_i,v_j). Again, for another complex initiated with a seed pair (v_l,v_m), the node $\mathrm{v_k}$ satisfies both criteria for cluster expansion. In that case, v_k is a member of both complexes and hence ComFiR can detect overlapping complexes.

Algorithm 1: ComFiR Algorithm steps for complex formation

```
Input :G={V, E} (PPIN); $P_cM$ (Benchmark complex matrix); SsT (Semantic similarity
threshold); NSS(Semantic similarity score matrix)

Output: Cluster = {$C_1, C_2, \ldots C_N$}, (a set of N complexes)

Initialize clusterExpNode = V, Cluster = NULL, count = 1, counts=1, counta=1, i=1;

  // Candidate seed selection

foreach $v_i \in V$
    choose $v_j$ from {V− $v_i$} such that $\forall$ $v_k \in$ {V−$v_i$}, $MMS(v_i, v_k) < MMS(v_i, v_j)$;
    CdS(counts)= $(v_i, v_j)$;
    MaxA(counts)=$MMS(v_i, v_j)$;
    counts++;

end

    // Seed selection procedure

while|MaxA(i)| > 0 do

    choose $sd_i$ from CdS such that $\forall j \in$ {MaxA− i}, MaxA(i) >MaxA(j);
    partialCluster =$sd_i$;
    clusterExpNode=clusterExpNode −$sd_i$;

            // Cluster Expansion process
```

```
                choose $v_m \in clusterExpNode$ such that $\forall\ v_n \in clusterExpNode,$
                $conn(v_m, partialCluster) \geq conn(v_n, partialCluster);$

                while{$v_m$ exists and $conn(v_m, partialCluster) \geq 0.4$} do
                        choose $v_m$ iff $\exists\, v_x \in partialCluster$ such that $NSS(v_m,v_x) \geq SsT$
                        $partialCluster = partialCluster \cup v_m$;
                        $clusterExpNode = clusterExpNode - v_m$;
                        choose next $v_m$;
                end

        Mark $partialCluster$ as $C_{count}$ only when $|partialCluster| \geq 3$;
                $Cluster = Cluster \cup C_{count}$;
                $count{+}{+}$;
                $MaxA = \{MaxA - MaxA(i)\}$;
                $i{+}{+}$;
        end
        Return $Cluster$ ;
```

2.4 *PPI complex ranking*

ComFiR ranks the complexes by exploiting the information available in standard biological databases such as GeneCard (Rebhan *et al.* 1997). For the purpose of ranking, since protein names are similar to gene names (Povey *et al.* 2001), we use the gene names directly from the database. Our ranking is based on the assumption that if a protein complex includes more number of disease genes, then it is likely to have more relevance as a causative entity for that disease. Traditionally, ranking is carried out on diseased complexes based on (i) the coherency of causative genes and (ii) inclusion of the most significant causative gene. However, both the techniques are ruled out as the first one needs the presence of at least three diseased genes in a complex, which may not be always true. The second one also has been ruled out owing to the fact that diseases may be caused by one or more number of genes. Hence, our approach is to rank these complexes according to the number of disease genes. If there occurs a tie between two complexes having the same number of disease genes, higher rank is given to the one having the best p-value. This is supported by the fact that since p-value gives the functional enrichment of a group, and therefore better the p-value of the complex, more will be its coherence. It takes the set of complexes or clusters obtained as input along with a query disease. We identify the genes responsible for causing the disease which is available in a database (Rappaport *et al.* 2013) consisting of human afflictions along with their causative agents. Using this information, we construct a heterogeneous network consisting of protein complexes and disease genes. This network shows the presence of an edge if the gene is found in the complex. Once the whole network is built, we find which are the protein complexes associated with the query disease. The number of disease associated genes in the complex determines its rank. In order to decide upon the priority of ranking, if there occurs a scenario where two of the complexes have the same number of disease genes, we can use the biological criteria of p-value. The p-value of complexes is obtained using BinGO plugin (Maere *et al.* 2005) available in Cytoscape. Members of a complex are fed as input to the BinGO tool, which retrieves the relevant GO annotations for each entry and proliferate it upwards in the complete GO tree. It gives the probability of occurrence of *a* in *A* test genes which is enriched with the same functional group *F* out of *p* in *P* reference set. Complex with the best p-value is given the highest rank. To describe our ranking approach, we need the following definitions.

Definition 8 Agene $g_i \in C_j$, i.e. *jth* complex, is said to be relevant disease gene with respect to a given disease query, if it can be validated using the database (Rappaport *et al.* 2013).

Definition 9 Acomplex C_j is called relevant disease complex with respect to a given disease query if it includes at least one protein encoded by a disease gene or includes at least one relevant disease gene (as in Definition 8).

Algorithm 2 presents the approach followed for ranking.

Algorithm2: ComFiR algorithm steps for ranking complexes for a given disease query

Input: $ActualCluster = \{C_1, C_2, \ldots, C_N\}$, (a set of N protein complexes obtained using Algorithm1); $D_g = \{g_1, g_2, \ldots, g_m\}$(a set of genes responsible for the query disease); $pvalue = \{pvalue_{C_1}, pvalue_{C_2}, \ldots, pvalue_{C_N}\}$ (p-value for each cluster obtained from *ActualCluster*).

Output: $rank = \{rank_{C_1}, rank_{C_2}, \ldots, rank_{C_N}\}$, (the rank of each disease associated complex.)

```
        Initialize $DGC = NULL$, $SortedDgComplex=NULL$,
        $rank = 0$, $Dgcount=0$, $DgComplex=0$, $swap=0$, $r=1$, $swapdp=0$;

        foreach $g_i \in D_g$ do
        {
                foreach $C_j \in ActualCluster$ do
                {
                 //Check if disease gene $g_i$ is present in complex $C_j$
                        if $g_i \in C_j$ then
                          {
                                $DGC_{g_i,c_j}=s_{g_i}$;
                          }
        }
}
//Calculate number of disease genes in each disease associated cluster
foreach $C_j \in DGC$ do
{
        foreach $v_i \in C_j$ do
        {
            if $v_i \in D_g$ then
              {
                 $Dgcount(x)=v_i$;
                 $x=x+1$;
              }
        }
$DgComplex_{C_j}= |Dgcount|$;
}
// Sort the disease associated complexes in decreasing order of the number of disease genes it
contains
foreach $C_j \in DgComplex$ do
{
        foreach $C_k \in DgComplex$ do
         {
            if $DgComplex_{C_j}<DgComplex_{Ck}$ then
              {
                 $swap=DgComplex_{C_j}$;
                 $DgComplex_{C_j}=DgComplex_{C_k}$;
                 $DgComplexC_k=swap$;
                 Interchange the complex indices accordingly.
              }
         }
}

$SortedDgComplex=DgComplex$;

$r=1$;

foreach $C_j \in SortedDgComplex$ do
   {
        if $DgComplex_{C_j}=DgComplex_{C_k}$
         {
           $rank_{C_j}=r$;
          $r=r+1$;
         }

// Get p-value of each complex, $C_i$ in $pvalue_{C_i}$ and sort them in decreasing order

foreach $C_j \in DgComplex$ do
  {
        foreach $C_k \in DgComplex$ do
           {
                if $pvalue_{C_j} > pvalue_{C_k}$ then
                 {
                        $swapdp=pvalue_{C_j}$;
                        $pvalue_{C_j}=pvalue_{C_k}$;
                        $pvalue_{C_k}=swapdp$;
```

```
                    Interchange the complex indices accordingly.
                }
            }
        }
        Sortedpvalcomplexes=pvalue;
        foreach $C_j \in$ Sortedpvalcomplexes
        {
            $rank_{C_j}$=r;
            r=r+1;
        }
        SortedDgComplex={SortedDgComplex - $C_j$};
    }
    Return rank for each cluster $rank_{C_j}$, where $j=\{1, 2, ..., N\}$, $j \in DGC$.
```

3. Computational results

We implemented the ComFiR method in MATLAB running on an HP Z 800 workstation with two 2.4 GHz Intel(R) Xeon (R) processors and 12 GB RAM, using the Windows 7 operating system. We carried out the experiment on three datasets: DIP dataset (Salwinski *et al.* 2004) comprising of 17202 interactions and 4903 proteins, Gavin_2006 (Gavin *et al.* 2006; Moschopoulos *et al.* 2009) consisting of 6531 interactions and 1430 proteins, and a human PPI extracted from HPRD (Prasad *et al.* 2009) consisting of 39237 interactions and 9088 proteins. In order to evaluate the quality of predicted complexes, we used a benchmark complex set available in literature. For the yeast dataset (DIP and Gavin_2006), we used MIPS and CYC2008 as the benchmark which consists of 203 and 408 complexes respectively, and for human PPI dataset, we used a bonafide complex set available in (Kikugawa *et al.* 2012; H-InvDB 2012). The semantic similarity for a given pair of proteins is found using the DaGO-Fun tool (Mazandu and Mulder 2013). We used Wang's semantic similarity (Wang *et al.* 2007) for our purpose. Wang's semantic similarity between two terms X and Y represented by $DAG_X = (X,\ T_{X,} E_X)$ and $DAG_Y = (Y,\ T_{Y,}\ E_Y)$ is given by

$$\boldsymbol{SS(X,Y) = \frac{{}_{pT_XT_Y}(S_X(p)S_Y(p))}{S_V(X)S_V(Y)}}$$

where $S_X(p)$ and $S_Y(p)$ are S-values of term p related to X and Y respectively, and $S_v(X)=\sum_{p\in TX} S_X(p)$ and $S_v(Y)=\sum_{p\in TY}$-$S_Y(p)$are semantic values of terms X and Y respectively.

A detail explanation of this semantic similarity measure is given in section 2 of the supplementary material. This is justified by its better performance over all other existing measures (Wang *et al.* 2007; Pesquita *et al.* 2009). The executable for ComFiR is available at *http://agnigarh.tezu.ernet.in/~dkb/resources.html*. The complexity analysis for our complex finding and ranking method is given in section 3 of the supplementary material.

3.1 *Protein complex finding results*

In order to compare our protein complex finding results with state-of the-art algorithms such as such as MCODE (Bader and Hogue 2003), FAG-EC (Li *et al.* 2008b), FT (Guenoche 2011), TFit (Gambette and Guénoche 2011), OCG (Becker *et al.* 2012), QCUT (Ruan and Zhang 2008), ClusterONE (Nepusz *et al.* 2012) and GMFTP (Zhang *et al.* 2014), we used performance indices such as Sensitivity (Sn), Positive Predictive Value (PPV) and Accuracy (Acc). These metrics are broadly discussed in (Brohee and Van Helden 2006). From our experimental study, it has been observed that for the range of *SsT* values (0.4–0.6), our method shows better performance, especially for *SsT*=0.4. It is shown in figure 3, we see that with increase in the *SsT* cutoff value, the results keep on improving. However, as we wanted to make limited use of known information in our method and we did not want to suppress the quality of results obtained with 40% connectivity, we used *SsT*=0.4 in computing our results.

In order to get a clear picture of these indices, we used a cross-tabulation matrix. This matrix P is of the order $p \times q$, where p represents the number of predicted complexes and q represents the number of benchmark complexes. Each entry P_{ij} in the matrix corresponds to the number of matched proteins among the benchmark i and predicted complex j. Now we discuss the performance indices.

3.1.1 *Positive predictive value:* The positive predictive value (PPV) is a measure of how much of the predicted complex set matches that of the real complex set. It is computed for every complex set and is defined as

$$PPV_{cmplxj} = max_{i=1}^{q} PPV_{i,j},$$

where $PPV_{i,j} = \frac{P_{i,j}}{P_j}$ and P_j being the marginal sum of the j^{th} complex. The overall *PPV* of a method is the average of PPVs of all complexes, i.e.

$$PPV = \frac{{}_{j=1}^{q} P_j PPV_{cmplxj}}{{}_{j=1}^{q} P_j}$$

PPV gives the odds of a predicted complex to be a benchmark complex. This implies higher the value of PPV the higher are the chances of the predicted complex set to correspond to known biological functions.

We report the PPV value for some existing methods along with our proposed method in figure 7(a).

3.1.2 *Sensitivity:* Sensitivity gives a measure of how much of the benchmark complex set is contained in the predicted set. The sensitivity of a complex Sn_{cmplxj} is defined as the maximum value of sensitivity obtained for complex *j* over all *p* real complexes. Mathematically, it is given as

$$Sn_{cmplxj} = max_{j=1}^{p} Sn_{i,j},$$

where $Sn_{i,j} = \frac{P_{i,j}}{N_i}$, with N_i representing the cardinality of complex *i*. The overall sensitivity is the weighted average of the individual ones and is defined as

$$Sn = \frac{{}_{i=1}^{p} N_i Sn_{cmplxi}}{{}_{i=1}^{p} N_i}.$$

A high sensitivity value on the performance scale indicates wider coverage of the method in detecting real complexes.

3.1.3 *Accuracy:* In order to take into account the tradeoff between the two indices, their geometric mean is computed as the *accuracy (Acc)* of the method.

Mathematically,

$$Acc = \sqrt{}(PPVXSn).$$

We tuned the parameter *SsT* used in ComFiR in order to obtain maximum accuracy and report the results at *SsT=0.4.*

3.2 *Complex finding results on yeast dataset*

We report the PPV and Sn value for some existing methods along with our proposed method for the DIP dataset in figure 4(a) and 4(b) respectively. The accuracy obtained using ComFiR and other methods for this dataset is shown in figure 5.

From figure 4(a), it can be seen that PPV value obtained using ComFiR is at the third position. It is beaten by both MCODE and ClusterONE. However, from figure 4(b), we see that *Sn* value for this dataset is at par with the top performing methods. A significant conclusion can be drawn from figure 5 which shows that ComFiR emerges as the winner among the contemporary methods for DIP dataset. We also validated the performance of our method in case of Gavin_2006 dataset. The results are reported in section 1 of the supplementary material.

3.2.1 *Performance in terms of Precision and F-measure:* Precision and F-measure are also used to analyse the performance of the method. These indices can be explained using the formula

$$Precision = \frac{EffC}{ActualCl}$$

$$Recall = \frac{MatchC}{BenchmarkCl}$$

$$F - meausre = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where *EffC* is the number of clusters found by the method, *ActualCl* is total number of clusters predicted by the method, *MatchC* is the number of complexes that match those in the benchmark set and *BenchmarkCl* is the total number of

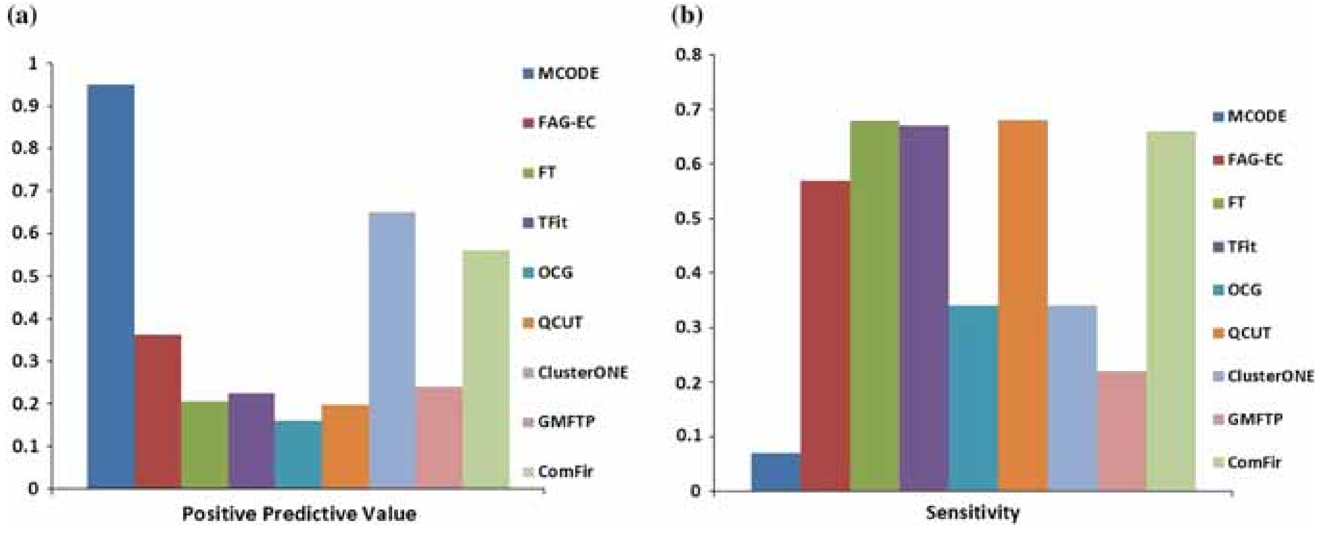

Figure 4. (**a**) Positive predictive value and (**b**) sensitivity of ComFiR and other methods on DIP dataset.

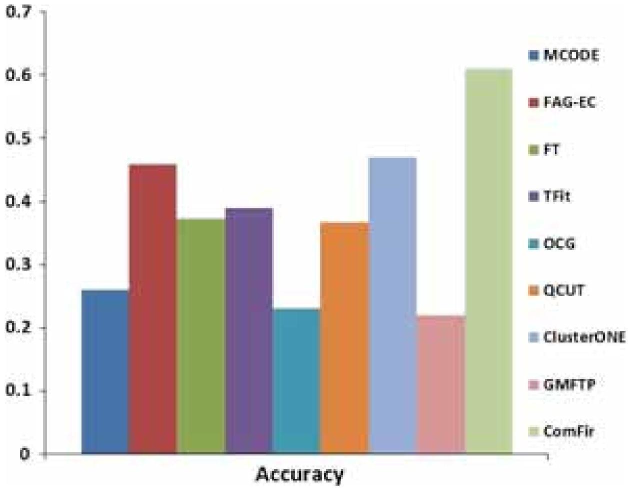

Figure 5. Accuracy of ComFiR and other methods on DIP dataset.

benchmark complexes. Determining *EffC* and *MatchC* requires analysing the overlapping score between elements in the predicted cluster set and the benchmark set. This overlapping threshold has been set up as 0.2 as given in (Sharma *et al.* 2015). We have reported Precision and F-measure of some of the methods against ComFiR over DIP dataset in figure 6.

From figure 6(a), it can be seen that ComFiR is the winner among all other methods in terms of Precision. In terms of F-measure, it holds the second position being beaten by MIPCE only as seen in figure 6(b).

3.3 *Complex finding results on HPRD dataset*

We report the *PPV* and *Sn* value for some existing methods along with our proposed method for HPRD dataset in figure 7(a) and 7(b) respectively. We could not compare our method with MIPCE as its results were not available for this dataset. The accuracy obtained using ComFiR and other methods is shown in figure 8.

In figure 7(a), we see that a large fraction of predicted complexes match with those of the bona fide complexes leading to a comparatively higher value of PPV than the rest of the methods.

In figure 7(b), we also see that the sensitivity obtained using ComFiR is far higher than the existing methods. This shows that a lot of complexes detected through ComFiR corresponds to those of the benchmark set.

The final conclusion is drawn from figure 8 where a high value on the accuracy scale makes ComFiR the winner over the existing methods. This is because it takes into account both sensitivity and PPV for computing accuracy of the method. The accuracy of ComFiR is around 55% which is more than double that of the most promising method discussed in literature till date. The steep rise in accuracy owes to using a semantic similarity constraint during the cluster expansion process. Without this constraint, the accuracy would be around 42% which is lower than that given in a most recent paper (Bandyopadhyay *et al.* 2015). Thus, we went ahead with using this constraint in order to get more compact and biologically relevant complexes.

The ROC obtained at different thresholds ranging from 0.1 to 0.8 for the HPRD dataset is given in figure 9.

3.4 *Protein complex ranking results for Alzheimer's disease*

In order to establish the significance of our ComFiR method, we experimented and analysed results for real human diseases. We used Alzheimer's disease (*OMIM Id-*

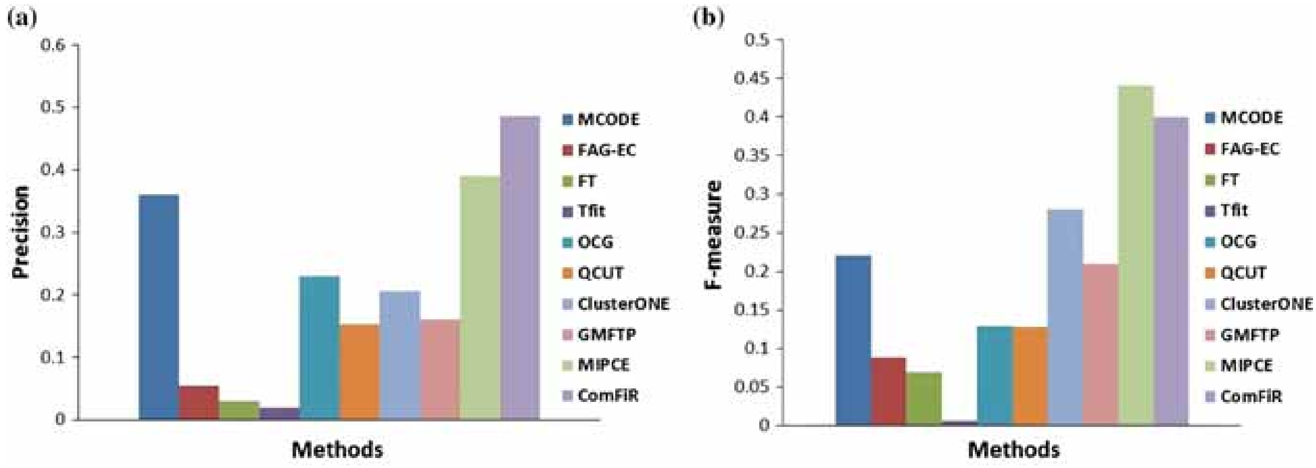

Figure 6. Comparison of ComFiR and other methods on DIP dataset in terms of (**a**) Precision and (**b**) F-measure at overlapping threshold of 0.2.

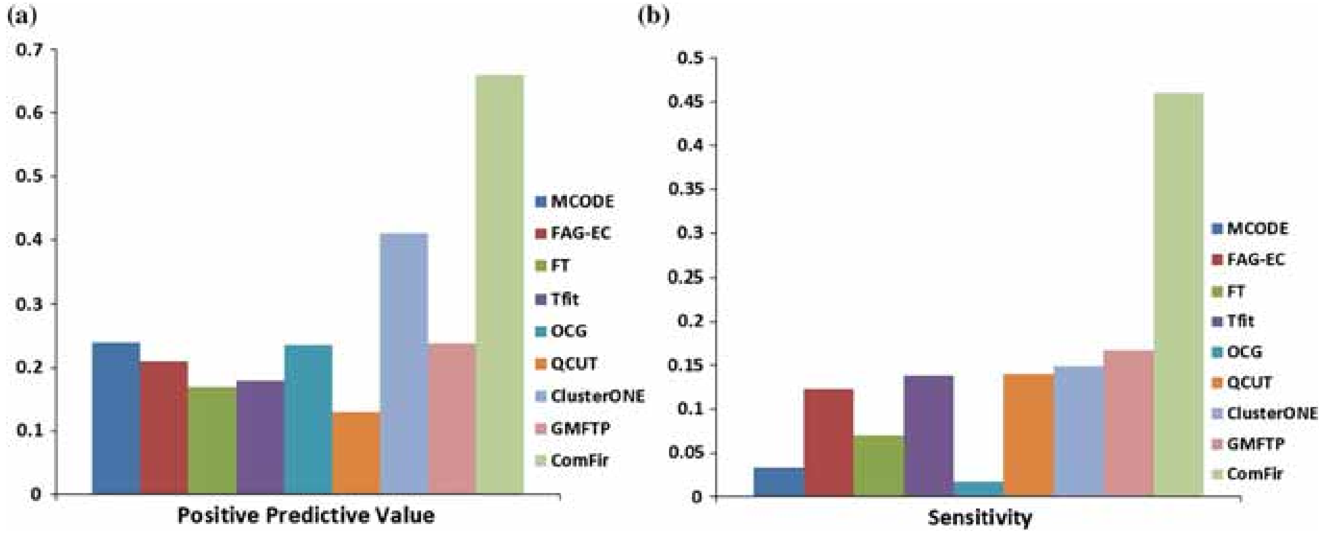

Figure 7. (**a**) Positive predictive value and (**b**) sensitivity of ComFiR and other methods on HPRD dataset.

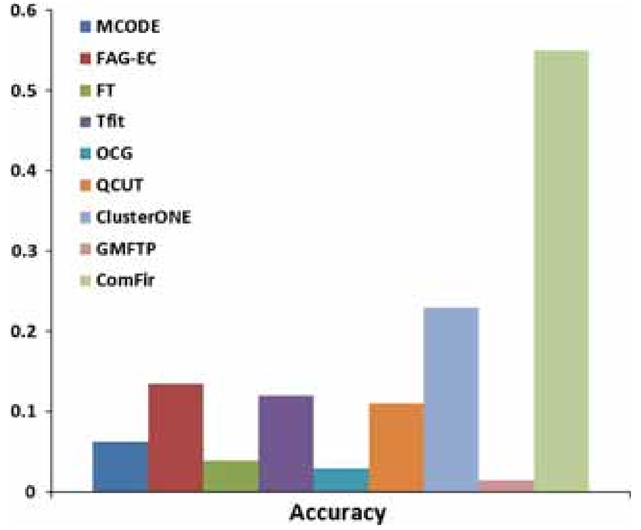

Figure 8. Accuracy of ComFiR and other methods on HPRD dataset.

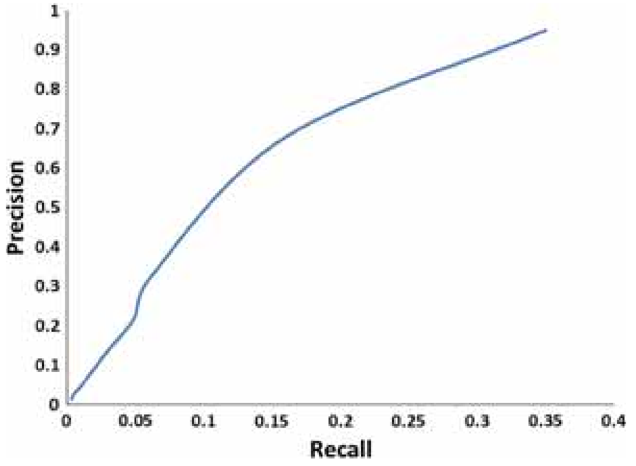

Figure 9. ROC of ComFiR for HPRD dataset.

104300), which is a chronic neurodegenerative disease affecting between age group 40–60 years. We obtained a list of 129 genes associated with this disease along with their relevance scores (Rappaport *et al.* 2013). We then used it to find which of the protein complexes obtained from Algorithm 1 are associated with the disease and ranked them according to the number of disease genes present along with its p-value using Algorithm 2. We also tried to find few additional genes (not available among 129 genes in GeneCard) from connectivity point of view which might be associated with the disease. These genes were members of top ranked complexes and later we supported our finding using some other sources of literature. Table 2 lists the top five complexes associated with Alzheimer's disease along with its member proteins. If we consider the top ranked complex, i.e. complex number 262, which shows two genes, *DLST* and *MPO*, as reported by Rappaport *et al.* (2013), and another gene, *UMPS*, which was reported in Cohen *et al.* (2010), it was not recorded among the 129 genes reported in Rappaport *et al.* (2013) known to cause the disease. The same is the case with most of the complexes found by our method. Only the top ranked complex has two disease genes as its members, while the other four has only one associated disease gene, and therefore the ranking is done based on their p-value. Higher rank is given to the complex with

Table 2. List of complexes associated with Alzheimer's disease

Complex No	Complete set	Subset of proteins associated with disease with reference	Associated chromosome number	Number of disease genes that are common from the repository we used during ranking	p-value
262	DLST, MPO,UMPS	DLST (Rappaport *et al.* 2013), **MPO** (Rappaport *et al.* 2013), **UMPS** (Cohen *et al.* 2010)	DLST (14), MPO(17), UMPS(3)	2	6.290E−4
152	BCL2, BCL2L14, SPNS1, STX18, ITM2B, CPT1A, TP53AIP1	**BCL2** (Cohen *et al.* 2010), **BCL2L14** (Cohen *et al.* 2010), **STX18** (Cohen *et al.* 2010), **ITM2B** (Rappaport *et al.* 2013), **CPT1A** (Cohen *et al.* 2010)	BCL2 (18), BCL2L14 (12), STX18 (4), ITM2B (13), CPT1A (11)	1	7.767E−5
82	TP53, TNNT1, SHISA5, DHCR24, SMYD2	**TP53** (Cohen *et al.* 2010), **SHISA5** (Maulik *et al.* 2013), DHCR24 (Rappaport *et al.* 2013)	TP53 (17), DHCR24 (1), SHISA5 (3)	1	9.642E−5
69	GSK3B, NDC80, FRAT2	GSK3B (Rappaport *et al.* 2013)	GSK3B(3)	1	2.4437E−4
50	NCOA3, GSK3B, ESRRB, ATAD2	**NCOA3** (Cohen *et al.* 2010), GSK3B (Rappaport *et al.* 2013), **ESRRB** (Inoue *et al.* 2013), **ATAD2** (Muller *et al.* 2011)	GSK3B(3), NCOA3 (20), ESRRB(14), ATAD2 (8)	1	1.118E−3

the most significant p-value. Thus, we could say that although we used the existing knowledge to get the rank of the complexes, we could still find relevance of some genes in causing the disease which were not available in the repository that we considered at the time of ranking. This can be done for diseases whose information is already available and can also be extended for unknown diseases with the help of some domain experts.

Table 2 gives a list of the genes associated with Alzheimer's Disease along with the chromosome numbers in which they reside. Studies have shown that the early-onset Alzheimer's disease is caused by mutations in genes present in chromosomes 21, 14 and 1. However, people who are infected by the disease at a later stage (referred to as late onset) tend to show causative mutations in chromosome 19 as well (National Institute on Aging 2016). Studies suggest that chromosomes 11 and 12 are somehow related to familial Alzheimer's disease (Mayo Clinic 2016).

This has also been evidenced by genes like CPT1A and BCL2L14 (found in chromosome 11 and 12, respectively) labeled as relevant genes among the top rated relevant complexes. A visual representation of the genes identified in the top complexes is shown in figure 10.

In figure 10, we can see that node DLST and MPO are isolated, but we have suggested them as members of a complex. This can be supported by the fact that the figure has been obtained using GeneMania tool which considers only a few associations, viz. physical co-expression, predicted, pathway, co-localization and shared protein domains to construct the network. However, in our complex finding method, we also used semantic similarity as an attribute to decide upon the complex membership. Thus, there is a possibility to get the two elements in the same complex.

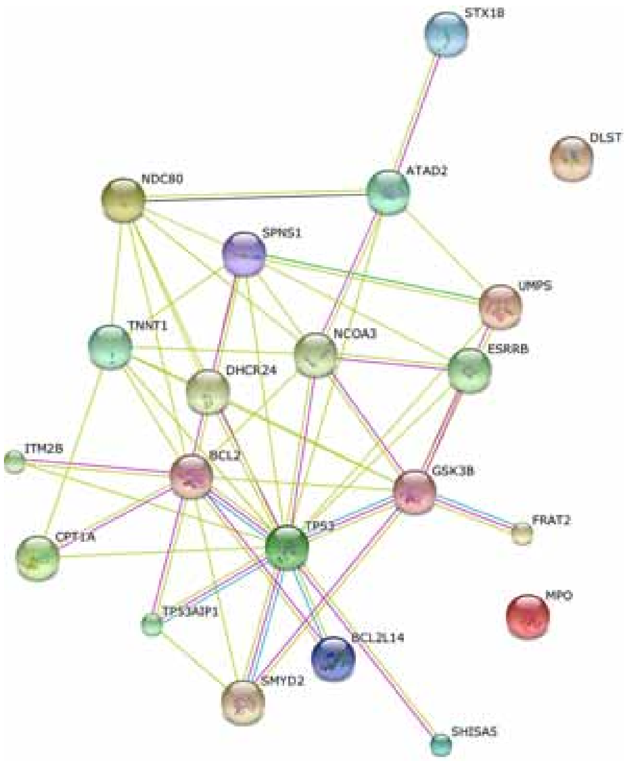

Figure 10. Genes found in top five complexes using ComFiR method on Alzheimer's disease.

4. Discussion

In this work, we introduced an effective method for protein complex detection and ranking, emphasizing its performance on a human PPI dataset. An application of PPI complex ranking while responding to a given disease query (i.e., Alzheimer's disease) has also been established. The complex finding technique has shown significant performance improvement compared to the other contemporary methods. Also, the ranking technique can predict complexes with disease genes effectively and rank them very well. We have validated both the techniques of ComFiR using benchmark datasets. We can further strengthen the protein complex ranking technique so that it is able to prioritize complexes with greater biological accuracy to help the biomedical scientists in drug design.

References

Bader GD and Hogue CW 2003 An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **4** 1

Bandyopadhyay S, Ray S, Mukhopadhyay A and Maulik U 2015 A multiobjective approach for identifying protein complexes and studying their association in multiple disorders. *Algorithms Mol. Biol.* **10** 1

Becker E, Robisson B, Chapple CE, Guenoche A and Brun C 2012 Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* **28** 84–90

Brohee S and Van Helden J 2006 Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinform.* **7** 1

Chen Y, Jacquemin T, Zhang S and Jiang R 2014 Prioritizing protein complexes implicated in human diseases by network optimization. *BMC Syst. Biol.* **8** S2

Chua HN, Ning K, Sung WK, Leong HW andWong L 2008 Using indirect protein-protein interactions for protein complex prediction. *J. Bioinform. Comput. Biol.* **6** 435–466

Cohen D, Chumakov I, Nabirochkin S, Guerassimenko O and Graudens E 2010 New diagnostic tools for alzheimer disease. US Patent App. 13/387,174

Erten S, Li X, Bebek G, Li J, Koyuturk M 2009 Phylogenetic analysis of modularity in protein interaction networks. *BMC Bioinform.* **10** 333

Gambette P and Guénoche A 2011 Bootstrap clustering for graph partitioning. *RAIRO Oper. Res.* **45** 339–352

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, *et al.* 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440** 631–636

Guenoche A 2011 Consensus of partitions: a constructive approach. *Adv. Data Anal. Classif.* **5** 215–229

H-InvDB 2012 PCDq Protein-Protein, Protein-Complex and Complex-Complex Interaction Viewer with Integrative Annotation. http://hinvitational.jp/hinv/pcdq/

Inoue H, Takahashi R, Kondo T, Iwata N and Asai M 2013 Method for screening therapeutic and/or prophylactic agents for Alzheimer's Disease. US Patent App.14/385,990

Jung SH, Hyun B, Jang WH, Hur HY and Han DS 2010 Protein complex prediction based on simultaneous protein interaction network. *Bioinformatics* **26** 385–391

Kashyap H, Ahmed HA, Hoque N, Roy S and Bhattacharyya DK 2016 Big data analytics in bioinformatics: architectures, techniques, tools and issues. *Netw. Model. Anal. Health Inform. Bioinform.* **5** 28

Kaufman L and Rousseeuw PJ 2009 Finding groups in data: an introduction to cluster analysis, vol **344** (John Wiley & Sons, New York)

Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, Kanaya S and Imanishi T 2012 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** 1

King AD, Pržulj N and Jurisica I 2004 Protein complex prediction via cost-based clustering. *Bioinformatics* **20** 3013–3020

Le DH 2015 A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. *Algorithms Mol. Biol.* **10** 1

Leung HC, Xiang Q, Yiu SM and Chin FY 2009 Predicting protein complexes from PPI data: a core-attachment approach. *J. Comput. Biol.* **16** 133–144

Li M, Chen Je,Wang Jx, Hu B and Chen G 2008a Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinform.* **9** 1

Li M, Wang J and Chen J 2008b A fast agglomerate algorithm for mining functional modules in protein interaction networks; in BioMedical Engineering and Informatics, 2008, BMEI 2008, International Conference on, IEEE, vol **1**, pp 3–7

Li XL, Foo CS and Ng SK 2007 Discovering protein complexes in dense reliable neighborhoods of protein interaction networks in Comput Syst Bioinformatics Conf (Citeseer) vol **6,** pp 157–168

Maere S, Heymans K and Kuiper M 2005 BinGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21** 3448–3449

Maulik M, Thinakaran G and Kar S 2013 Alterations in gene expression in mutant amyloid precursor protein transgenic mice lacking niemann-pick type c1 protein. *PloS one* **8** e54, 605

Mayo Clinic 2016 Alzheimer's genes: Are you at risk? http://www.mayoclinicorg/diseases-conditions/alzheimers-disease/indepth/alzheimers-genes/art-20046552

Mazandu GK and Mulder NJ 2013 DaGO-fun: tool for gene ontology-based functional analysis using term information content measures. *BMC Bioinform.* **14** 284

Moschopoulos CN, Pavlopoulos GA, Schneider R, Likothanassis SD and Kossida S 2009 Giba: a clustering tool for detecting protein complexes. *BMC Bioinform.* **10** S11

Muller S, Filippakopoulos P and Knapp S 2011 Bromodomains as therapeutic targets. *Expert Rev. Mol. Med.* **13** e29

National Institute on Aging 2016 Alzheimer's disease research centers. https://www.nianihgov/alzheimers/alzheimers-disease-research-centers

Nepusz T, Yu H and Paccanaro A 2012 Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9** 471–472

Pesquita C, Faria D, Falcao AO, Lord P and Couto FM 2009 Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* **5** e1000,443

Pobbati AV, Chan SW, Lee I, Song H and Hong W 2012 Structural and functional similarity between the vgll1-tead and the yap-tead complexes. *Structure* **20** 1135–1140

Povey S, Lovering R, Bruford E, Wright M, Lush M and Wain H 2001 The HUGO gene nomenclature committee (HGNC). *Hum. Genet.* **109** 678–680

Prasad TK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, *et al.* 2009 Human protein reference database-2009 update. *Nucleic Acids Res.* **37** D767–D772

Rao VS, Srinivas K, Kumar GS and Sujin G 2013 Protein interaction network for Alzheimer's disease using computational approach. *Bioinformation* **9** 968

Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, *et al.* 2013 Malacards: an integrated compendium for diseases and their annotation. Database:bat018

Rebhan M, Chalifa-Caspi V, Prilusky J and Lancet D 1997 Genecards: integrating information about genes, proteins and diseases. *Trends Genet.* **13** 163

Ruan J and ZhangW2008 Identifying network communities with a high resolution. *Phys. Rev. E* **77** 016,104

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D 2004 The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32** D449–D451

Sharma P, Ahmed HA, Roy S and Bhattacharyya DK 2015 Unsupervised methods for finding protein complexes from PPI networks. *Netw. Model. Anal. Health Inform. Bioinform.* **4** 1–15

Srihari S, Ning K and Leong HW 2010 MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating coreattachment structure. *BMC Bioinform.* **11** 1

Szymanski DB 2005 Breaking the wave complex: the point of arabidopsis trichomes. *Curr. Opin. Plant Biol.* **8** 103–112

Van Dongen SM 2001 Graph clustering by flow simulation University Thesis University of Utrecht

Vanunu O, Magger O, Ruppin E, Shlomi T and Sharan R 2010 Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6** e1000,641

Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ and Koller D 2009 A complex-based reconstruction of the saccharomyces cerevisiae interactome. *Mol. Cell. Proteom.* **8** 1361–1381

Wang JZ, Du Z, Payattakool R, Philip SY and Chen CF 2007 A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23** 1274–1281

Winkel BS 2004 Metabolic channeling in plants. *Annu Rev Plant Biol* **55** 85–107

Wu M, Li X, Kwoh CK and Ng SK 2009 A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinform.* **10** 1

Wu M, Yu Q, Li X, Zheng J, Huang JF and Kwoh CK 2013 Benchmarking human protein complexes to investigate drug-related systems and evaluate predicted protein complexes. *PloS ONE* **8** e53,197

Yi C and Deng XW 2005 Cop1 from plant photomorphogenesis to mammalian tumorigenesis. *Trends Cell Biol.* **15** 618–625

Zhang XF, Dai DQ, Ou-Yang L and Yan H 2014 Detecting overlapping protein complexes based on a generative model with functional and topological properties. *BMC Bioinform.* **15** 186

Corresponding editor: Hampapathalu Adimurthy Nagarajaram