Chapter 4

Protein Complex Finding Methods: An Application to Alzheimer's Disease

4.1 Introduction

A number of studies have shown the association between protein complexes and diseases. Mutation in a gene coding protein in a complex can lead to dysfunction of protein complex leading to a disease condition. For example, a complex of proteins named SCR1B, NOS1AP and VANGL1 is known to be associated with progression of breast cancer. Another protein complex consisting of prostaglandin d-synthase (PDS) and transthyretin (TTR) is a biomarker for Alzheimer's Disease [77]. In this chapter, we focus on the problem of detecting quality protein complexes detection from PPI networks and their role in context of Alzheimer's disease.

4.2 Related Work

Researchers believe that incorporating a substantial amount of biological information into topological measures during complex finding may yield biologically enriched complexes. This is because the data used for PPI analysis are obtained from experimental techniques and are often found to be noisy and have a large number of false positives. One can refine such data using biological information and use them for complex finding. A number of researchers have worked in this manner. *RNSC* [66] works using a cost reduction technique. It divides the PPI network by shuffling proteins in between clusters so as to optimize a cost function. The final step involves assigning p-values to clusters based on their functional coherence. This method returns only those clusters with $p - value < 0.001$ as final complexes. Another method called *DECAFF* [81] uses density of local neighborhoods to find complexes. *PCP* [17] uses functional annotations of genes to assign weights to a raw PPI dataset. Using this weighted network, it then clusters proteins based on a clique merging process. A method called *GMFTP* [171] uses a propensity score based on topological network and functional annotations to determine the affinity among proteins in a complex. An extended version of *COACH* [160] called *WCOACH* [68] integrates semantic similarity among proteins to determine complexes. DBGPWN [173] works on a weighted PPI network, where weights between the edges is given by the semantic similarity between proteins. The next step uses a density based method similar to DBSCAN to identify complexes.

Usually, methods which detect protein complexes rely on the topological outlook of the PPI network. However, not all networks correspond to the assumed topology, thereby hampering the performance of unsupervised methods. Some researchers use the knowledge from real complexes as a basis to find new complexes. Qi et al. [114] used the topological, biological and chemical properties of known complexes to train a Bayesian Network model. Subgraphs denoting complexes are scored using a log-likelihood ratio. If this ratio exceeds a certain threshold, the subgraph can be considered a complex. Shi et al. [135] used weighted scores for different properties of a complex. They trained a neural network based on these properties. Any subgraph given to this model was then assigned a score, which ultimately decide if that subgraph could correspond to a complex or not based on a specified threshold. Both these methods required additional effort to determine the properties of already known complexes. Some properties such as protein length and polarity of amino acids used in these methods need chemical knowledge. Hence, the utility of these methods were limited. In addition, they could not be practically used for a human PPI dataset due to the need to have properties of benchmark complexes known apriori. A summary of a few such methods is given in Table 4.1.

Mutations in gene coding proteins are known to disrupt the normal functioning of complexes. A study of such genes can be useful in tracing their role in the disease. Many researchers have analyzed complexes associated with diseases. Topological properties of disease genes can be used to analyze the inherent organization of the interactome and their linkage possibilities. Certain works have also focussed on ranking of complexes associated with diseases. In [148], the prioritization process is

Table 4.1 Summary of protein complex finding techniques based on a combination of both topological and functional information

Method	**Salient feature**	**Datasets Used**	**Availability**
RNSC [66]	Uses a cost function in the initial step, then filters them based on p-value	DIP, Krogan	GIBA tool
DECAFF [160]	Based on local neighborhood density	DIP, BIOGRID	-
PCP [17]	Uses functional annotation of genes for weighting PPI edges	MIPS	-
GMFTP [171]	Uses both topological and functional annotations to decide upon members in a complex	DIP, BIOGRID	Matlab code
Bayesian Network based [114]	Uses topological, biological and chemical properties of known complexes during training phase	MIPS	www.cs.cmu.edu/ gyj/SuperComplex
Neural Network based [135]	Uses a neural network to train a model based on known properties	MIPS, DIP, BIOGRID	-

based on the formation of protein complexes (the member proteins are individually ranked first depending on their association scores in causing a particular disease). These complexes are evaluated in terms of functional, expression and conservation coherency [148]. Another method called MAXCOM [15] uses the concept of maximum information flow to prioritize the candidate protein complexes w.r.t. a given query disease. It uses the information from a heterogeneous network, which is made up of disease-phenotypic similarities, disease-protein links and PPIs. A more recent method called NBH [77] 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. A strongly supported ranking of these complexes would narrow down the analysis of only those genes known to be associated with the disease.

In this chapter, we make the following contributions.

- We propose a method called CSC using topological as well as biological information to find entities that can be used to detect complexes.
- We also propose a semi-supervised method for complex detection. This is known as ComFiR. It uses both topological and biological properties to detect protein complexes.

4.3 Protein complex finding based on topological and functional information: CSC

A protein complex is a group of functionally similar proteins which act together to achieve certain biological functions. Experimental details of Tandem Affinity Purification-Mass Spectrometry revealed that proteins in a complex are arranged in the form of a dense core, which is helped by periphery proteins to achieve certain functions. Certain other researchers have demonstrated the importance of combining functional information with topological informations of PPI network to detect quality complexes.

In this section, we propose a method called CSC which uses a set of topological as well as biological criteria to find complexes. The proposed technique has the following features.

- CSC detects protein complexes of high biological significance compared to other similar published approaches.
- This method has been established as effective considering well-known performance measures such as Sensitivity, Positive Predictive Value and Accuracy for two model organisms, i.e., yeast and human.
- A framework to analyze the influence of disease gene in a complex both biologically as well as topologically in terms of eight association parameters is also discussed.

4.3.1 Proposed algorithm: CSC

The problem of potein complex finding is an unsupervised learning process which involves partitioning the PPI network into parts based on topological and functional similarity. This method uses a combination of both these features to detect biologically significant complexes from the network. The following definitions are used during the complex finding process.

Definition 10 (HConfidence measure). *HConfidence measure between a pair of vertices (v_i, v_j) is given as the ratio of the common elements in the neighbor set, $Ng(v_i)$ and $Ng(v_j)$ of the the two nodes to its minimum connectivity value given by their degrees, d_{v_i} and d_{v_j}. Mathematically,*

$$HC(v_i, v_j) = \frac{Ng(v_i) \cap Ng(v_j)}{min(d_{v_i}, d_{v_j})} \tag{4.1}$$

Definition 11 (Seed pair). *Two nodes, v_i and v_j can be a candidate seed pair, $sd = (v_i, v_j)$ if $HC(v_i, v_j) > HC(v_k, v_l)$, $\forall v_k, v_l \in \{V - sd\}$.*

Definition 12 (Semantic similarity). *The semantic similarity between a protein pair (v_i, v_j) is calculated as the distance between the GO terms with which they are associated in the DAG structure of Gene Ontology.*

$$semsim(v_i, v_j) = sim(GOterms_i, GOterms_j) \tag{4.2}$$

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

Definition 13 (Reachability Index). *The reachability index of a node v_i in a cluster C_i is given as the number of links the direct neigbors of v_i have within C_i to the* *number of edges within C_i. Mathematically,*

$$RbI_{v_i} = \sum_{dN} \frac{l_{wC_i}}{tedges_{C_i}} \tag{4.3}$$

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

Definition 14 (Contribution)**.** *The contribution of a subgraph say G', is the aggregate of the reachability indices of all nodes $v_1, v_2 ... v_k$ in the subgraph.*

$$Cbtn(G') = \sum_{i=1}^{k} RbI_{v_i} \tag{4.4}$$

Definition 15 (Non-reachable proteins)**.** *A pair of protein nodes $sd = (v_i, v_j)$ is considered non-reachable if $\nexists v_k$ such that $Connt(v_k, sd) \geq \alpha$, where α is estimated to be 0.4 based on experimental results (Figure 4.1).*

Definition 16 (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 $Cbtn(G'') \geq Cbtn(G')$ where $G'' = G' \cup \{v_m\}$, $v_m \in v_j$ is a new candidate node to be added to G' .*

Definition 17 (Overlapped complex)**.** *Two protein complexes C_1 and C_2 are said to overlap if $C_1 \cap C_2 \neq \phi$.*

The proposed complex finding method follows the seed selection and expansion approach to extract 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 performeds using the HConfidence score for each pair of node. At every iteration, a seed pair with the highest HConfidence score is chosen as the seed pair for cluster expansion. Once the seed nodes, say, $sd = (v_a, v_b)$ are selected, the pair is inserted into the pC. Then the process of cluster expansion is performed in an unsupervised manner. A node v_c with the highest connectivity (among all nodes) with pC is chosen as the first candidate for cluster expansion. During cluster expansion, the goal is to make the topological and functional contibution (α and β respectively) during cluster formation to be 1, i.e.,

$$\alpha + \beta = 1 \tag{4.5}$$

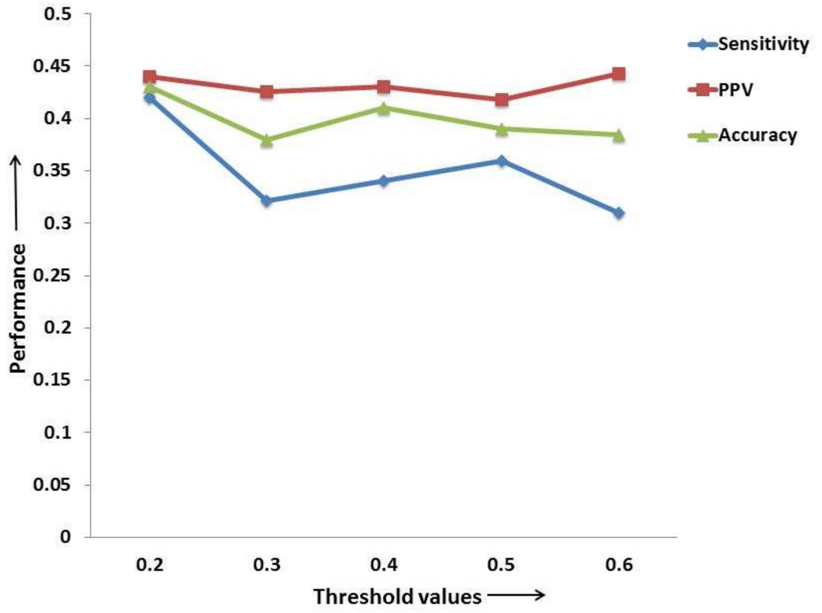


Figure 4.1. Performance indices obtained at varying thresholds using HPRD dataset.

Our experimental analysis suggests the most suitable connectivity threshold (α) to be 0.4. This is explained by the performance graph shown in Figure 4.1, which shows stable performance at around 40% connectivity. The membership of node v_c is further strengthened by the semantic similarity values between the nodes in pC and v_c. The threshold for semantic similarity (β) is accordingly adjusted to 0.6 for Equation (4.5) 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 pC. However, the decisive role is played by the contribution function calculated for pC 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 pC, else the elements in the pC are returned as outlier proteins. This process is repeated until no further node is left satisying 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. The pseudo code of the method is given in Algorithm 5.

The following propositions are used to establish the effectiveness of the CSC method over other existing methods.

Proposition 4. *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

```
Input : $G = \{V, E\}$ (PPIN); $\beta$ (Semantic similarity threshold); $SSm$ (Semantic similarity score
        matrix), $h_t$ (Hconfidence threshold)
Output: $Cluster = \{C_1, C_2, \cdots, C_N\}$, (a set of $N$ complexes)

Initialize $\mathsf{RemList} = V, \mathsf{NodeExpcluster} = V, \mathsf{secCluster} = NULL, \mathsf{Cluster} = NULL, p = 1, ccount =$
  $1, scount = 1, acount = 1, i = 1, hcount = 1$;
// HConfidence calculation
foreach $v_i \in V$ do
    foreach $v_j \in V$   such that $\forall v_j \in \{V - v_i\}$ do
        $HCS = HC(v_i, v_j)$;
        $hcount + +$;
    end
end
// Seed selection procedure
while $|HCS(p)| > h_t$ do
    choose $sd_p$ from $HCS$  such that $\forall q \in \{HCS - p\}$, $HCS(p) > HCS(q)$ and $sd_p = (v_i, v_j)$
      obtained from $HCS(p)$;
    $pC = sd_p$;
    $NodeExpcluster = NodeExpcluster - sd_p$;
    //Cluster Expansion process
    choose $v_m \in NodeExpcluster$ such that $\forall v_n \in NodeExpcluster, Connt(v_m, pC) \geq Conn(v_n, pC)$;
    $Cbtn(old) = Cbtn(pC)$;
    while $v_m$ exists and $Connt(v_m, pC) \geq 0.4$ do
        choose $v_m$ if and only if $\exists v_x \in pC$ such that $SSm(v_m, v_x) \geq \beta$
        $pC1 = pC \bigcup v_m$;
        $Cbtn(new) = Cbtn(pC1)$;
        if $Cbtn(new) > Cbtn(old)$ then
            $pC = pC \bigcup v_m$;
            $RemList = RemList - v_m$;
            choose next $v_m$;
        else
            Declare seed nodes as outlier proteins ;
        end
    end
    Mark $pC$ as $C_{ccount}$ only when $|pC| \geq 3$;
    $secCluster = secCluster \bigcup C_{ccount}$;
    ccount++;
    AMax={AMax-AMax(i)};
    i++;
end
// Removing Redundant clusters
foreach $c_i \in secCluster$ do
    $Clusteracount = Cluster \bigcup c_i$ ;
    $acount + +$;
end
Return Cluster ;
```

Algorithm 2: CSC Algorithm for complex formation

cluster expansion depending on 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 pC. 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. □

Computational complexity For a given network of n nodes, to compute the maximum value of every combination of seed pairs, CSC requires $O(n_{unique}^2)$ time, where n_{unique} is the number of unique elements in the graph. Choosing a seed node each time requires $O(n_{unique})$ time. Once the seed pair is chosen, cluster expansion needs to traverse all the remaining nodes to find the node which satisfies the connectivity criterion. This requires $O(n_{unique}^2)$ time. Once a node is chosen considering connectivity, it has to satisfy the semantic similarity criterion. In order to reduce time complexity, we consider only those nodes which fulfill the connectivity criterion or are part of the seed node. Let the number of such values be m. Looking for a particular pair of proteins among m nodes requires $O(m)$ time. This node can be added to the *partialCluster* in $O(1)$ time and then the contribution value of the subgraph is calculated. Calculating the contribution requires at most $O(n_{unique})$ computation. The cluster can be expanded if all the conditions satisfy. So the overall complexity for complex finding is $O(n_{unique}^2) + O(n_{unique}) + O(n_{unique}^2) + O(m) + O(1) + O(n_{unique}) \equiv O(n_{unique}^2)$, since m is small compared to n_{unique}.

4.3.2 Experimental Results

The CSC method is implemented 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 use DIP and HPRD datasets for the performance evaluation. For the DIP dataset, we use two manually curated PPI sets–MIPS and CYC2008 as the benchmark whereas for HPRD, we use PCDq as the benchmark set. Details of the dataset and benchmark set are given in Subsections 2.1.6.1 and 2.1.7.1 in Chapter 2.

We compare the performance of CSC with methods such as MCODE [4], FAG-EC [79], FT [121], TFit [33], OCG [11], QCUT [121], ClusterONE [101], GMFTP

[171], CNCM [130] and DCRS [129]. We also compared accuracy of CSC with TINCD [102], that uses a data ensembling approach for finding protein complexes. However, due to unavailability of the source code of TINCD, we used results directly from [102] for the same dataset and benchmark set.

Results on yeast dataset We report the Sensitivity, Positive Predictive Value and Accuracy of CSC with other contemporary methods for thee DIP dataset. A detailed explanation of the performance indices is given in Subsection 2.1.9 of Chapter 2. Figures 4.2-4.4 show the sensitivity, positive predictive value and accuracy of CSC w.r.t. other methods using MIPS as the benchmark set.

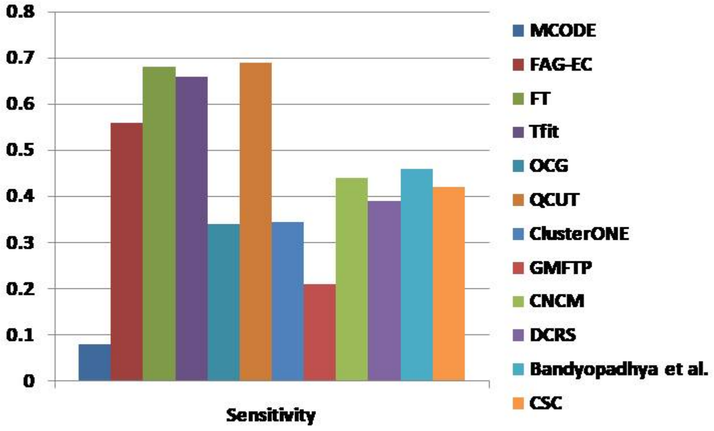


Figure 4.2. Comparing Sensitivity of CSC with other algorithms on DIP dataset using MIPS as benchmark.

Sensitivity of the CSC method is around 42%, which is better than a few other methods such as MCODE, OCG, ClusterONE and GMFTP and DCRS, as shown in Figure 4.2 whereas Positive Predictive Value of CSC is beaten by MCODE and ClusterONE only, as seen in Figure 4.3. In Figure 4.4, we see that accuracy of CSC is higher than all other methods except TINCD [102]. We could not compare our results with TINCD in terms of sensitivity and PPV as these results were not reported in the original paper [102]. 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.

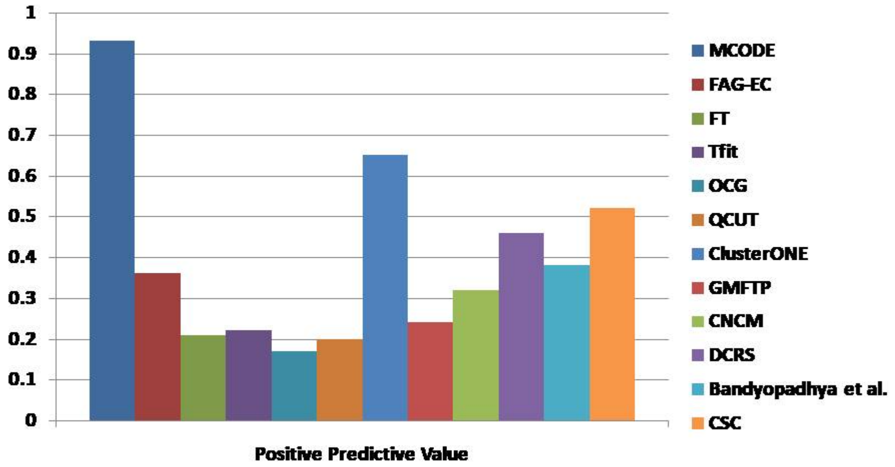


Figure 4.3. Comparing Positive Predictive Value of CSC with other algorithms on DIP dataset using MIPS as benchmark.

We also used another benchmark set called CYC2008 for performance comparison of our method. Figures 4.5-4.7 show the performance on DIP dataset using CYC2008 as the benchmark dataset. In Figure 4.5, we see that the sensitivity of the CSC method is quite low 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 places as seen in Figure 4.6. The accuracy of our method is around 40%, whereas two other methods– ClusterONE and TINCD show an accuracy of 50 - 70% as seen in Figure 4.7.

Parameter tuning The performance indices obtained by CSC can be improved by fine tuning the β threshold in the algorithm. This is justified by Figure 4.8. As seen, increasing the value of β leads to an increase in the value of the performance indices. However, in order to get a fair trade-off for our method, we used $\alpha = 0.4$ and $\beta = 0.6$ as suggested in Subsection 4.3.1.

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

Results on HPRD dataset The performance of the CSC method is also

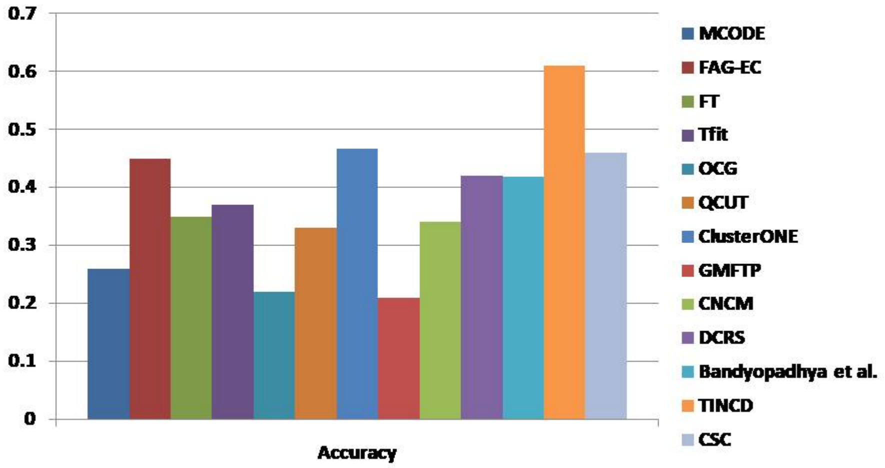


Figure 4.4. Comparing Accuracy of CSC with other algorithms on DIP dataset using MIPS as benchmark.

analyzed in terms of a bigger HPRD dataset [104], which is the Human Protein Reference Dataset comprising of 39,209 interactions. The literature [15, 75, 148] has shown that the knowledge of protein complexes can be used in disease diagnosis, so it is our keen interest to analyze the accuracy of our method over the human dataset. A more accurate method would aid biomedical scientists in developing a better understanding of complexes and would prove helpful in finding their association with diseases. We have compared the performance values of the CSC method with nine other methods: MCODE [4], FAG-EC [79], FT [40], TFit [33], OCG [11], QCUT [121], ClusterONE [101], GMFTP [171], CNCM [130] and DCRS [129] as shown in Figure 4.9-4.11. As seen in Figure 4.9, the sensitivity of CSC is around 23%, which is much higher than other methods except CNCM and DCRS. The PPV of our method is at the third position after ClusterONE and DCRS(Figure 4.10). In Figure 4.11, we see that accuracy of our method emerges as the second winner after DCRS in this dataset.

4.3.3 CSC: An Application to Alzheimer's Disease

The effectiveness of a protein complex can be evaluated from both topological and biological points of view. As the literature points out [34], mutation in gene coding

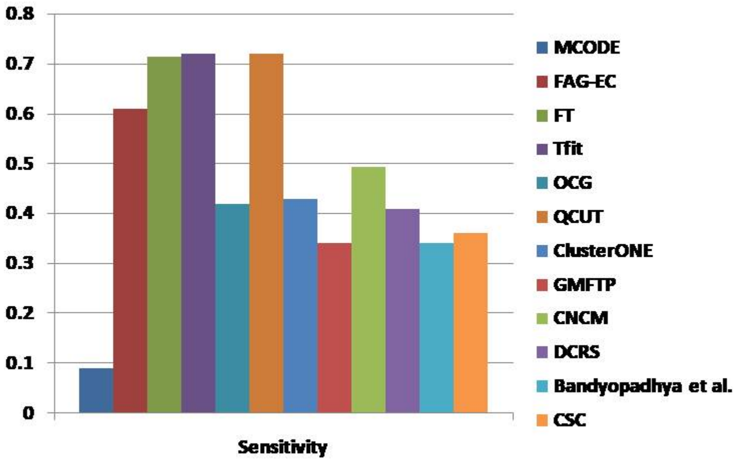


Figure 4.5. Comparing Sensitivity of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

proteins in a complex may lead to diseases. For example, the SWI/SNF complex is known to be associated with Coffin-Siris syndrome and plays a role in causing cancer [15, 159]. We analyze a subset of complexes based on a few query diseases. In order to find this subset of complexes, we use the disease related gene information given in GeneCard [116]. We use the gene names given in GeneCard as there is one-to-one correspondence between genes and proteins and proteins are named the same way as genes [108]. In this work, only a single disease is considered, so the number of disease genes found from GeneCard is not 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 generic framework.

The Disease Gene-Central Gene Analysis Framework Now, we present a conceptual framework to analyze the associations of a disease gene with the central gene(s) (chosen to represent a complex based on connectivity) of complexes. A visual represenation of the framework is shown in Figure 4.12. The following definition and illustration are useful for further description of the application.

Definition 18 (Central gene of a complex)**.** *A gene g_j is defined to be a central gene of a complex C_i, if $\nexists g_k$ such that its associations in C_i is higher than that of*

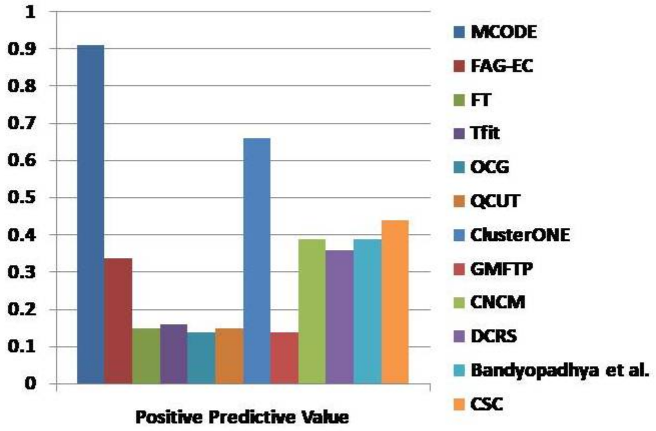


Figure 4.6. Comparing Positive Predictive Value of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

$g_j \in C_i$.

For illustration, we use the example graph shown in Figure 4.13, representing a complex given by CSC. 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 this figure.

From the set of complexes given by CSC, the central gene representing each complex is identified. 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, the disease genes are mapped 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, the disease associated complexes are identified quickly. This process outputs results of the form *<GeneID, Genename, Complexlist>*. The *Complexlist* is a set of indexed complexes which have $GeneID$ as a member. 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, associations of central gene(s) and other genes with the disease gene(s) can be further explored.

To support our analysis, we use two online tools, GeneMania [156] and STRING [139]. GeneMania is a web based tool which features functions such as analyzing

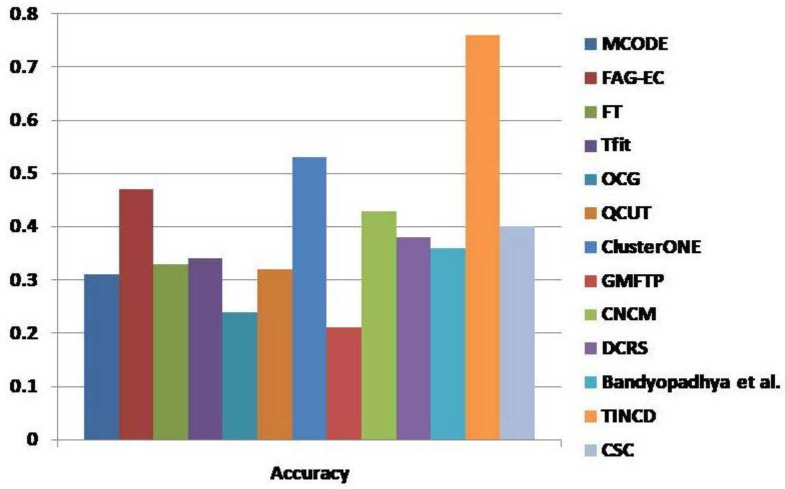


Figure 4.7. Comparing Accuracy of CSC with other algorithms on DIP dataset using CYC2008 as benchmark.

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 also 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. In addition, it also shows edge information obtained using *text mining* considering 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.

Analyzing complexes w.r.t. Alzheimer's Disease We consider an example in the form of Alzheimer's Disease to analyze the complexes found by the CSC method. Among all forms of mental illnesses, Alzheimer's Disease is devastatingly common. It is the sixth leading cause of death, especially among the elderly.

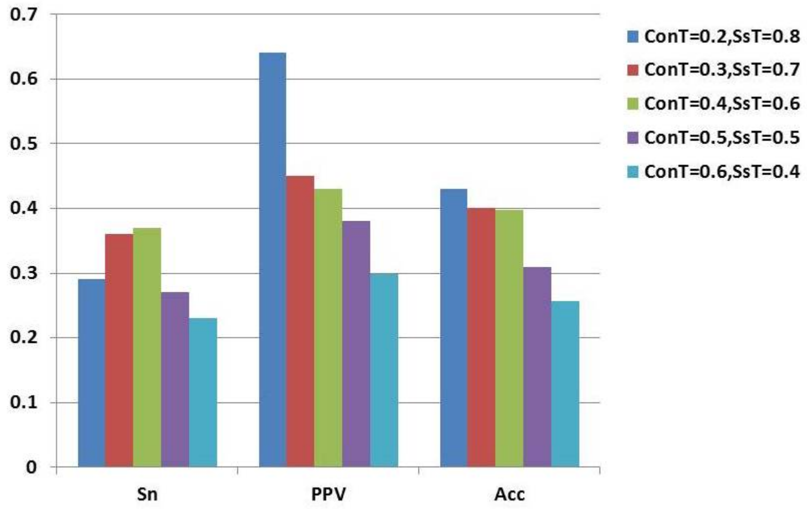


Figure 4.8. Comparing Sensitivity, Positive Predictive Value and Accuracy of CSC with other algorithms on DIP dataset using CYC2008 as benchmark with varying α and β thresholds.

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 criterion during complex finding leads to a reduced search space for the protein complexes to form. Due to decrease in the size of the search domain, the CSC method can find very few complexes associated with the disease. The members of two of the complexes are analyzed using the tools discussed above. The two disease associated complexes along with their member proteins and associations among them are given in Table 4.2. In Table 4.2, columns 5-11 show the association of the disease gene with the central gene as well as other complex members w.r.t. the 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 [152]. Table 4.3 gives the pathways with which each member of the two disease associated complexes are associated. In Table 4.3, we observe 75% similarity in the pathway information in complex 1 and 100% similarity in pathway in complex 2. Therefore, we can say that CSC is able to extract high quality complexes both from statistical and biological point of view.

Table 4.2 Alzheimer associated complex (Association of disease gene with other genes in the complex)

S.No	**Disease gene(s) in complex**	**Other members of complex**	**Whether Central gene**	**Physical interaction**	**Co-expression**	**Predicted interaction**	**Pathway**	**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 4.3 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	75
	APH1A	No	Notch signaling pathway	
	TMED2	No	Pre-notch expression and processing	
	TMED10	No	mRNA processing	
2	TOMM22	No	Mitochondrial protein import	100
	TOMM7	No	Mitochondrial protein import	
	TOMM40	Yes	Mitochondrial protein import	

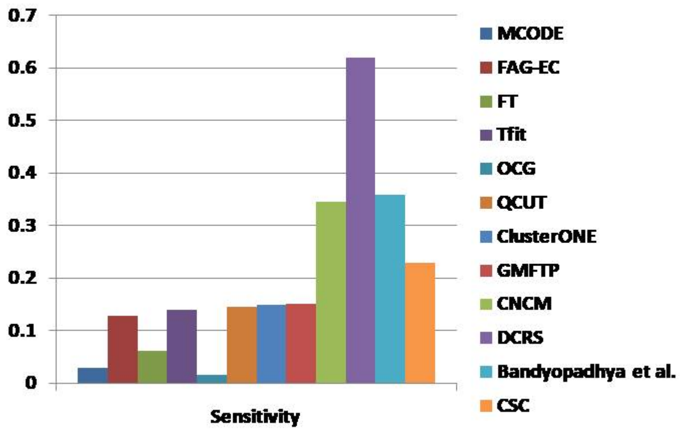


Figure 4.9. Comparing Sensitivity of CSC with other algorithms on HPRD dataset.

4.3.4 Discussion

We have presented a method which gives accurate results for protein complex finding. The accuracy of our method CSC, for the human dataset is significantly higher than existing methods. We have established the biological significance of our method empirically. We have also introduced 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. Further, improvements to the accuracy of complex detection methods can be considered, as more accurate detection would lead to higher reliability of prediction of the functions of disease genes. In order to enhance the accuracy of complex finding method, we propose a semi-supervised technique for this purpose, discussed next.

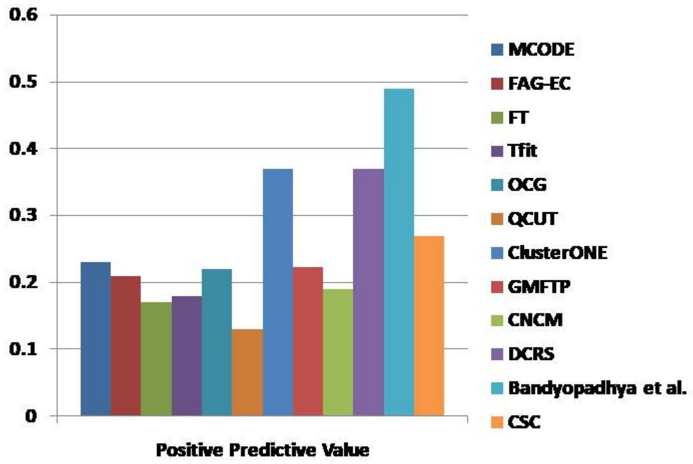


Figure 4.10. Comparing Positive Predictive Value of CSC with other algorithms on HPRD dataset.

4.4 Protein complex finding using semi-supervised technique: ComFiR

The task of analyzing protein complexes in the context of diseases requires a clear-cut demarcation among diseased and non-diseased complexes. One cannot randomly assign a protein associated with a disease gene to a complex simply based on a similarity criteria. Therefore, it is better to use existing knowledge to classify such complexes. This approach is a semi-supervised classification technique to detect protein complexes from PPI networks. Using some amount of established data leads to more reliable complex prediction and hence more accurate analysis.

In this section, we therefore incorporate a little knowledge from the existing benchmark complexes along with a simple and established form of knowledge called GO semantic similarity between proteins to develop our complex finding method called ComFiR. The following are the notable features of ComFiR.

- It uses topological and biological features of a PPI network during expansion of already established seed pairs.
- It has been validated in terms of yeast and human PPI datasets.

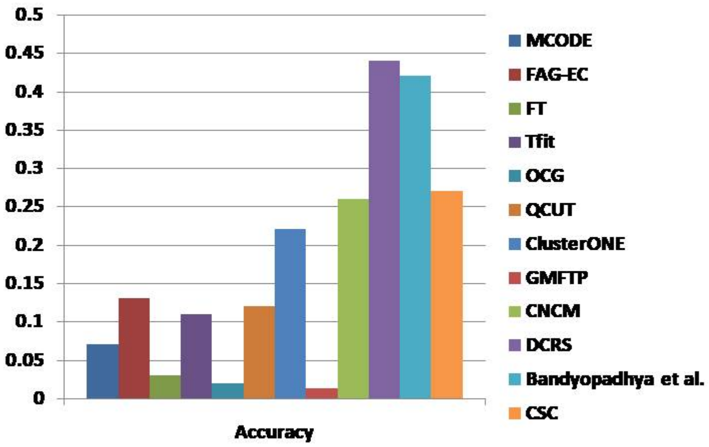


Figure 4.11. Comparing Accuracy of CSC with other algorithms on HPRD dataset.

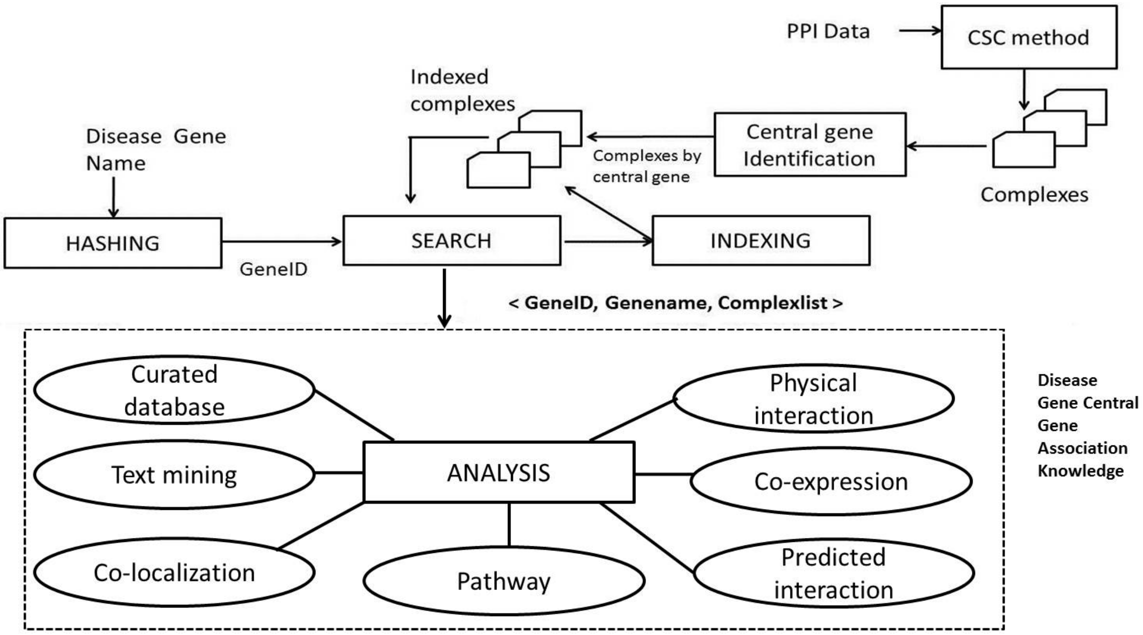


Figure 4.12. Disease gene-Member genes analysis framework

- It shows better performance in terms of accuracy for both datasets, hence justifying the use of a semi-supervised approach.
- Biologically enriched complexes are analyzed w.r.t. a disease.

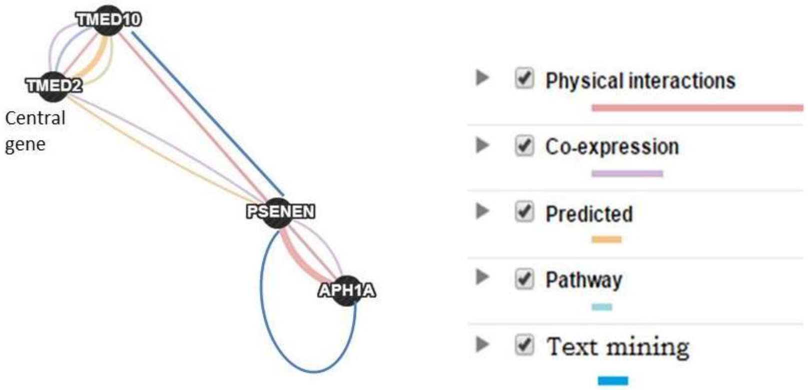


Figure 4.13. Protein complex members along with its association links

- The disease associated complexes are ranked based on properties of these complexes.

4.4.1 Proposed Method : ComFiR

The main reason for using a semi-supervised approach for complex finding is to enhance the accuracy of complexes w.r.t. benchmark complexes. Once the accuracy of the process is found to be satisfactory, one can analyze the complexes w.r.t. a query disease. We use the information available in benchmark complexes to find seed pairs required for cluster expansion. Once the seed pair is available, the expansion process is completely unsupervised based on topological and functional properties of the PPI network. The following definitions are used in this context.

Definition 19 (Seed pair)**.** *A seed pair during complex formation is a pair of nodes $S_{sd} = (v_i, v_j)$ which co-occur in benchmark complexes with significantly high frequency compared to other pairs.*

Definition 20 (Protein benchmark complex matrix)**.** *This corresponds to a binary matrix, BCM, where entries are computed as follows:*

$$BCM_{i,j} = \begin{cases} 1 & \textit{if protein } i \textit{ is found in } j^{th} \textit{ benchmark complex} \\ 0 & \textit{otherwise} \end{cases}$$

Definition 21 (Density)**.** *The density of a subgraph $G' \subseteq G$, where $G' = \{v_1, v_2, ...v_p\}$ is given by the actual number of links in G' i.e., $|e_{al}|$ divided by the maximum number of possible links i.e., E_{ml}. Mathematically,*

$$density(G') = \frac{|e_{al}|}{E_{ml}} \tag{4.6}$$

where $E_{ml} = \sum_{c=1}^{p} d_{v_c} \times ((d_{v_c} - 1)/2)$ *and* d_{v_c} *is the degree of node* v_c.

Definition 22 (Outlier proteins). *A pair of nodes* $S_{sd} = (v_i, v_j)$ *corresponds to an outlier, if* $\nexists v_l$ *such that* $Connt(v_l, S_{sd}) \geq \alpha$ *(default value taken to be 0.4).*

Definition 23 (Protein complex). *A subgraph* $G' \subseteq G$ *is defined as a protein complex if* $\forall v_i \in G'$*, the connectivity threshold* (α) *and the semantic similarity threshold* (β) *are satisfied.*

Definition 24 (Maximum matching score). *The maximum matching score is the maximum frequency with which any pair of nodes co-occur in a complex. This is obtained using the* BCM *matrix.*

A working example is discussed here.

Example 1: Suppose, proteins v_a and v_b occur in complexes C_1, C_2 and C_3. Also, assume that protein v_c occurs only in complex C_2. In this case, the *matching scores* of the combination of seeds is as follows. $ms_{v_a,v_b} = 3$, $ms_{v_a,v_c} = 1$ and $ms_{v_b,v_c} = 1$. Since the matching score of (v_a, v_b) is the highest, we start the complex finding procedure with the pair (v_a, v_b) as the seed nodes.

The process starts with a seed pair $S_{sd} = (v_i, v_j)$ that has the maximum matching score among all node pairs. This calculation is explained with the help of Example 1. The expansion process then passes through the connectivity test, which decides the strength between the new node v_a and the pC (now containing the seed pair). This new node is considered a probable candidate for expansion only if $Connt(v_a, pC) \geq \alpha$. Choice of α is made by varying this parameter from 0.2-0.6. An optimal choice is found at $\alpha = 0.4$ as shown in Figure 4.14. The final decision for a node v_a to become a member of pC is made by using the functional property. We use the Wang's semantic similarity to decide whether two nodes can belong to the same group or not. A node v_a can be a member of pC if $semsim(v_a, pC) \geq \beta$. Once both the criteria are satisfied, v_a gets added to pC and the process continues by choosing a new node considering connectivity. The expansion process continues as long as nodes satisfying the connectivity criterion exist. Once this criterion fails, the next seed pair is chosen and the expansion process begins. We consider clusters with more than three elements as complexes, as smaller size clusters lack informativeness and unnecessarily add to the precision value. Another issue here is the formation of redundant clusters from different seed nodes. This is taken care of by removing such clusters from the original set. Redundancy may arise in two forms–either the whole cluster set matches another cluster or there is a very high percentage of overlap between two clusters. The second form of redundancy can be handled by using an overlapping threshold given by Equation 4.7. Clusters with more than 80% overlap [101] are taken as redundant clusters and only the larger one of the two is taken into the unique cluster set, $ActualCluster$.

$$Ov_{th} = \frac{|C_i \cap C_j|^2}{|C_i||C_j|} \tag{4.7}$$

The algorithm is given in Algorithm 3.

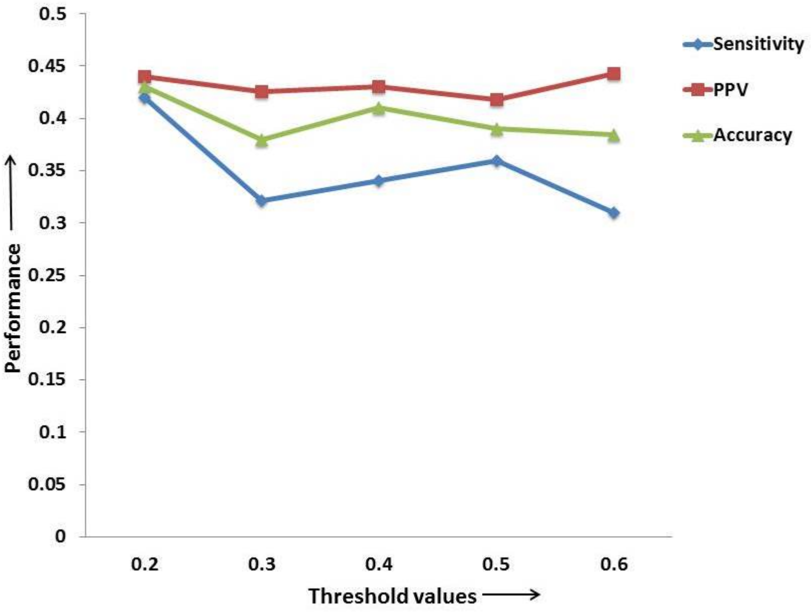


Figure 4.14. Performance indices obtained by varying α threshold on HPRD dataset.

Computational complexity For a given network of n nodes, to compute the maximum value of every combination of seed pair, ComFiR requires $O(n_{unique}^2)$, where n_{unique} is the number of unique elements in the graph. Choosing seed node each time requires $O(n_{unique})$. Once the seed pair is chosen, cluster expansion needs to traverse all the remaining nodes to find the node which satisfies the connectivity criterion. This requires $O(n_{unique}^2)$. Once a node is chosen from connectivity point of view, it has to satisfy the semantic similarity criterion. Here, in order to reduce time complexity, we have considered only those nodes which fulfills the connectivity criterion or are part of seed node. Let the number of such values be m. Looking for a particular pair of proteins among m nodes requires $O(m)$. So the overall complexity for complex finding in ComFiR is $O(n_{unique}^2) + O(n_{unique}) + O(n_{unique}^2) + O(m) \equiv O(n_{unique}^2)$, since m is small as compared to n_{unique}.

To establish the effectiveness of ComFiR, we introduce the following proposition.

```
Input  : $G = \{V, E\}$ (PPIN); $BCM$ (Benchmark complex matrix); $\beta$
         (Semantic similarity threshold); $NSSm$ (Semantic similarity
         score matrix)
Output: $Cluster = \{C_1, C_2, \cdots, C_N\}$, (a set of $N$ complexes)

Initialize $RemList = V$, NodeExpcluster $= V$, Cluster $= NULL, ccount =$
  $1, scount = 1, acount = 1, i = 1$;
// Candidate seed selection
foreach $v_i \in V$ do
    choose $v_j from \{V - v_i\}$ such that
      $\forall v_k \in \{V - v_i\}, MMgS(v_i, v_k) < MMgS(v_i, v_j)$ ;
    $CndS(scount) = (v_i, v_j)$;
    $AMax(scount) = MMgS(v_i, v_j)$;
    scount++;
end
// Seed selection procedure
while $|AMax(i)| > 0$ do
    choose $S_{sd_i} fromCndS$ such that
      $\forall j \in \{AMax - i\}, AMax(i) > AMax(j)$;
    $pC = S_{sd_i}$;
    $NodeExpcluster = NodeExpcluster - S_{sd_i}$;
    //Cluster Expansion process
    choose $v_m \in NodeExpcluster$ such that
      $\forall v_n \in NodeExpcluster, Connt(v_m, pC) \geq Connt(v_n, pC)$;
    while $v_m$ exists and $Connt(v_m, pC) \geq 0.4$ do
        choose $v_m$ if and only if $\exists v_x \in pC$ such that $NSSm(v_m, v_x) \geq \beta$
        $pC = pC \bigcup v_m$;
        $NodeExpcluster = NodeExpcluster - v_m$;
        choose next $v_m$;
    end
    Mark $pC$ as $C_{ccount}$ only when $|pC| \geq 3$;
    $Cluster = Cluster \bigcup C_{ccount}$;
    ccount++;
    AMax={AMax-AMax(i)};
    i++;
end
Return $Cluster$ ;
```

Algorithm 3: ComFiR Algorithm steps for complex formation

Proposition 5. *ComFiR detects overlapping protein complexes.*

Explanation: ComFiR is a two-step process involving seed selection followed by cluster expansion. A new node v_n is used to expand pC iff it satisfies the connectivity criterion and the functional similarity threshold. Assume a node v_n is a member of a complex formed with an initial seed pair $S_{sd} = (v_i, v_j)$. For another complex initiated with a seed pair $S_{sd} = (v_l, v_m)$, the node v_n satisfies both criteria for cluster expansion. In this case, v_n is a member of both complexes and hence ComFiR is able to detect overlapping complexes. □

4.4.2 Experimental 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 performed experiments on two datasets–DIP and HPRD dataset. Details of these datasets are given in Subsection 2.1.6.1 of Chapter 2. The semantic similarity for a given pair of proteins is found using the DaGO-Fun tool [93], explained in Chapter 2. We compared the performance of ComFiR with a few state-of the-art algorithms such as MCODE [4], FAG-EC [79], FT [40], TFit [33], OCG [11], QCUT [121], ClusterONE [101], GMFTP [171], CNCM [130], DCRS [129] and CSC [134] in terms of Sensitivity, Positive Predictive Value and Accuracy.

From the experiments, we observe that the optimal range of β is 0.4-0.6 for our method. At $\beta = 0.4$, ComFiR shows the best performance. In Figure 4.15, we see that with increase in the β cutoff value, the results keep on improving. However, we report the results at $\beta = 0.4$, so as to maintain a fair balance between both topological and functional properties of PPI network.

Figures 4.16 and 4.17 show the Positive Predictive Value, Sensitivity and Accuracy for the DIP dataset using MIPS as the benchmark set.

In Figure 4.16(a), we see that the PPV value of MCODE and ClusterONE are at the top two positions. The PPV of ComFiR is at the third position. The Sn value of our method is at par with the top performing methods shown in Figure 4.16(b). However, considering accuracy, ComFiR emerges as the winner for the DIP dataset. This shown in Figure 4.17.

Precision and F-measure are also used to evaluate the performance of ComFiR on the DIP dataset. Details of these measures are discussed in Subsubsection 2.1.9.1 of Chapter 2. Figure 4.18 shows these two indices for the DIP dataset. In Figure

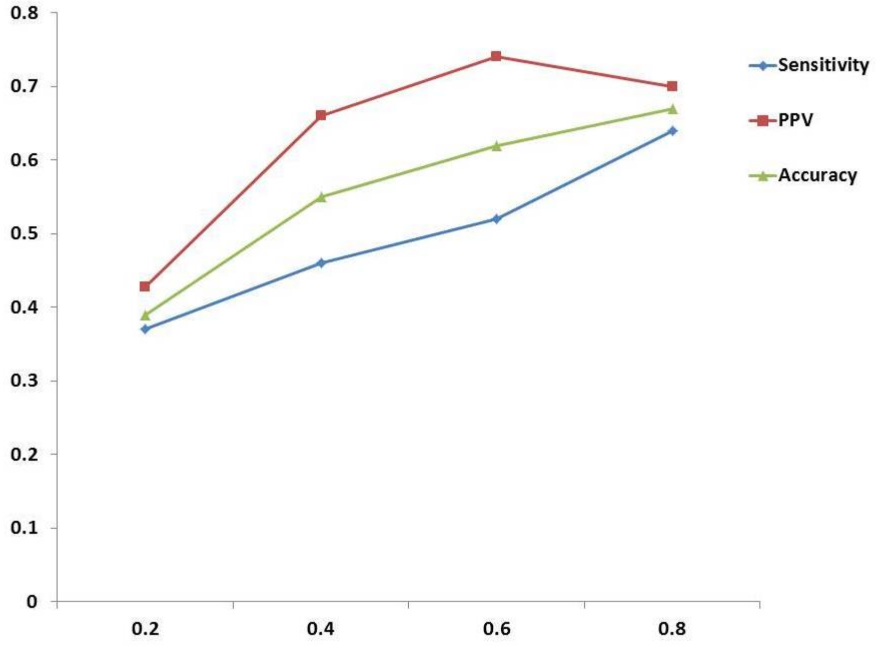
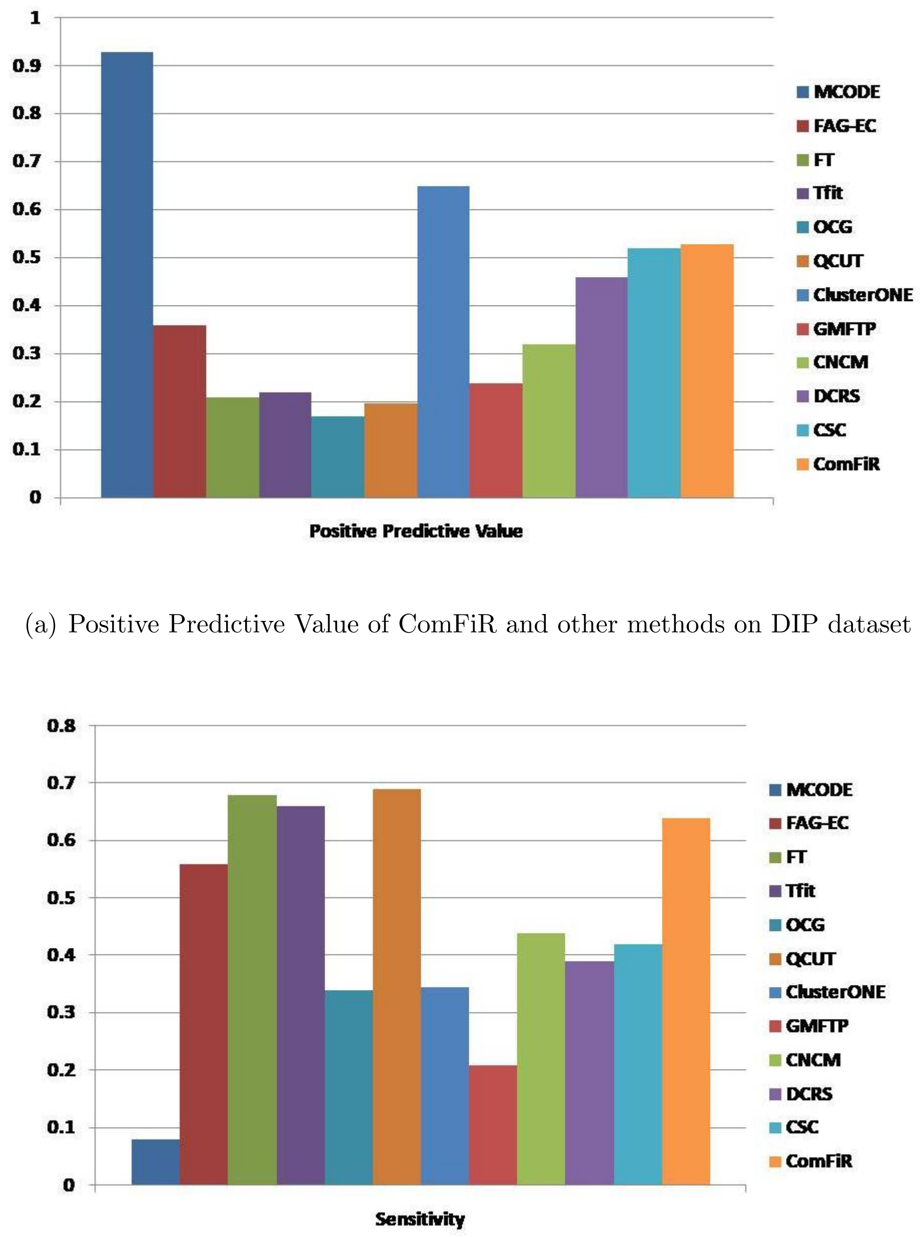


Figure 4.15. Performance indices of ComFiR obtained at varying thresholds of β on HPRD dataset.

4.18(a), we observed that ComFiR is the winner in terms of Precision. In terms of F–measure, only MIPCE is seen to beat the performance of ComFiR as seen in Figure 4.18(b).

Next, we report the results obtained using ComFiR on the HPRD dataset along with the results of several other algorithms in Figures 4.19 and 4.20. In Figure 4.19(a), a high PPV corresponding to ComFiR implies that a large fraction of predicted complexes match with those of the bonafide complexes. Similarly, we see that the sensitivity obtained using ComFiR is far higher than the values obtained by the existing methods except DCRS as seen in Figure 4.19(b). This implies that a lot of complexes detected by ComFiR correspond to those in the benchmark set. In Figure 4.20, we observe that ComFiR emerges as the winner in terms of accuracy among all existing methods. The accuracy of ComFiR for the HPRD dataset is around 55%, which is almost double that of the most promising method discussed in the literature till date. The rise in accuracy is attributed to the use of semantic similarity during the cluster finding process. Even without using this constraint, the accuracy was found to be around 42%, far higher than most other methods except [5]. We therefore use this constraint to get more biologically meaningful clusters.

Protein Complex Ranking w.r.t. query disease ComFiR uses the infor-



(a) Positive Predictive Value of ComFiR and other methods on DIP dataset

(b) Sensitivity of ComFiR and other methods over DIP dataset

Figure 4.16. Positive Predictive Value and Sensitivity of ComFiR and other methods on DIP dataset.

mation available in a repository called GeneCard [116] to rank the disease associated complexes. This database stores a list of causal genes associated with a number of diseases. During our ranking approach, we use the gene names directly as there is ample evidence [109] for the existence of homonyms between gene names and protein names. Traditionally, ranking of diseased complexes is usually performed done based on *(i)* the coherency of causative genes or *(ii)* inclusion of the most

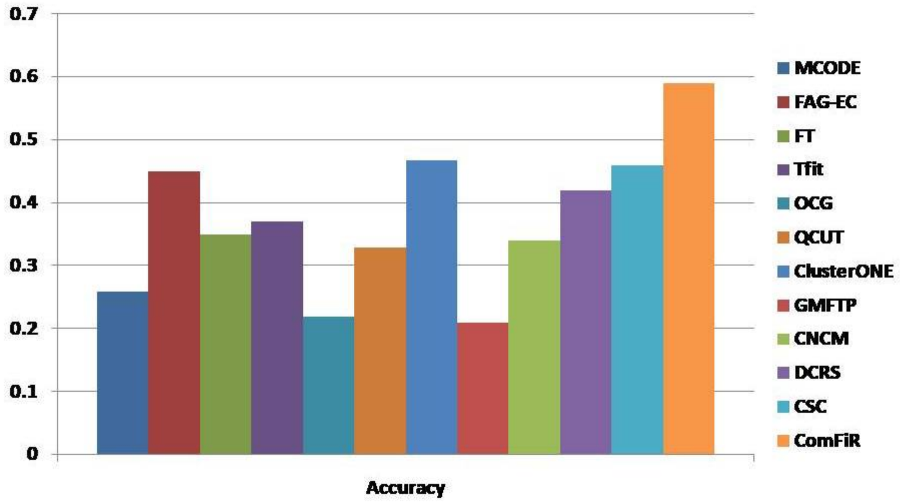
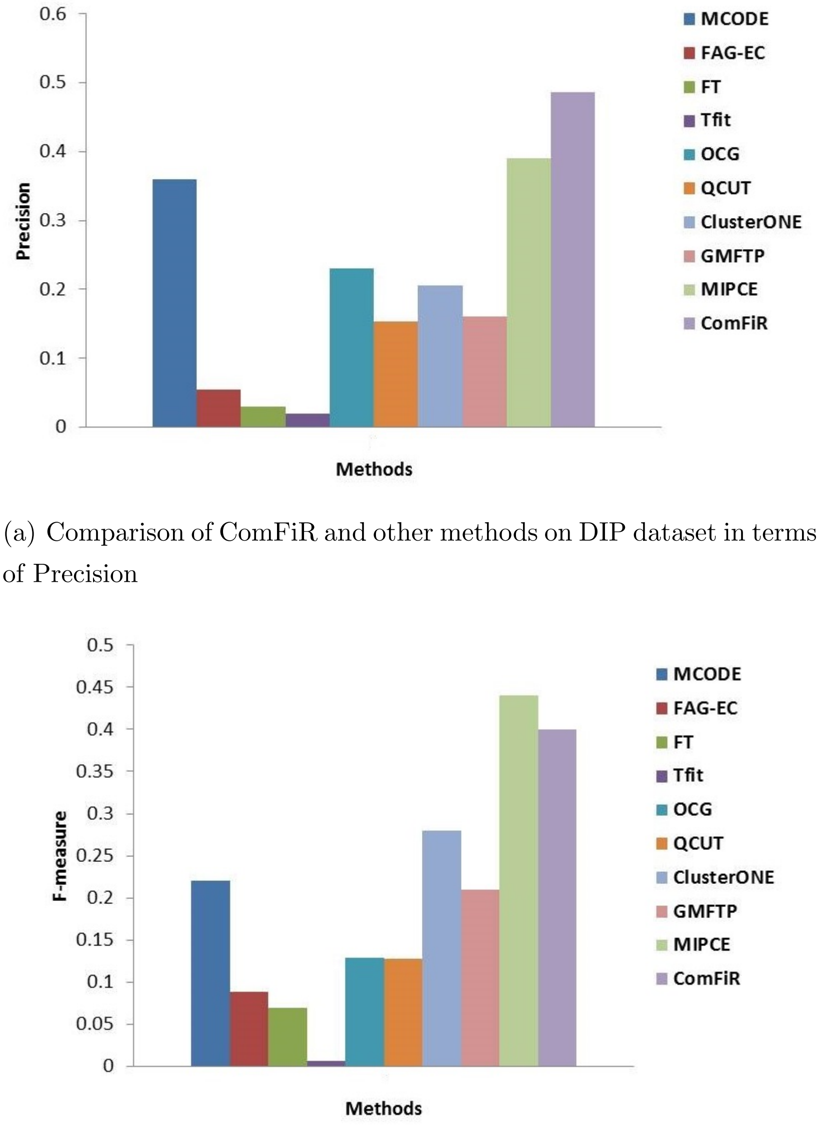


Figure 4.17. Accuracy of ComFiR and other methods on DIP dataset

significant causative gene. However, there are certain limitations of both these methods. The first one fails if the diseased complex has fewer than three elements, which might be possible in case of sparse complexes, while the second one fails if the disease belongs to a non-Mendelian class. Our assumption of ranking diseased complexes is associated with the number of disease genes in a complex. If there are more than one complex with the same number of disease genes, we use p-value to decide the ranking. This is supported by the fact that p-value gives the functional enrichment of a set of genes and lower the p-value, better is its functional enrichment.

Our ranking approach takes the set of complexes along with a user defined query disease. The first step involves finding a demarcation line between diseased and non-diseased complexes using the information available in GeneCard. A composite network is formed comprising of complexes and the causal genes. An edge between a gene and a complex represents the presence of this gene in the corresponding complex. The set of diseased complexes is then separated from the set of complexes and ranking begins by counting the number of disease genes in each of these complexes. There might be a situation when two complexes have the same number of disease genes. In such a case, the p-values of such complexes are calculated using the BinGO tool [89] of Cytoscape. The p-value then decides the priority of the complexes. The lower the p-value of a complex, the higher is its rank. The following definitions are used during our ranking approach.

Definition 25 (Relevant disease gene)**.** *A gene $d_{g_r} \in C_j$, i.e., j^{th} complex, is a*



(a) Comparison of ComFiR and other methods on DIP dataset in terms of Precision

(b) Comparison of ComFiR and other methods on DIP dataset in terms of F –measure

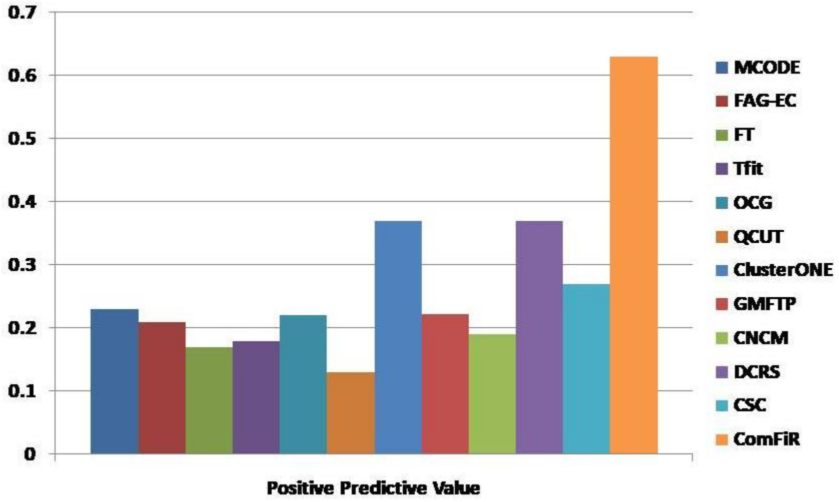
Figure 4.18. Comparison of ComFiR and other methods on DIP dataset in terms of Precision and F –measure at overlapping threshold of 0.2.

relevant disease gene if its association with the disease can be supported by the findings available in a database [116].

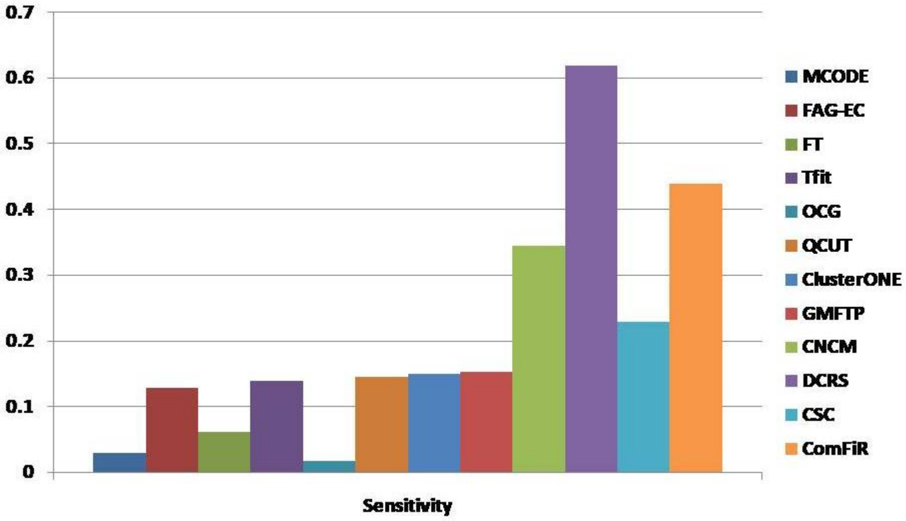
Definition 26 (Relevant disease complex)**.** *A complex C_j is a relevant disease complex with respect to a given disease query if $\exists d_{g_r} \in C_j$ w.r.t. the query disease. This d_{g_r} is the disease gene coding protein, which forms a member of C_j.*

Figure 4.21 presents the approach followed for ranking.

Protein complex ranking results for Alzheimer's Disease To understand



(a) Positive Predictive Value of ComFiR and other methods on HPRD dataset



(b) Sensitivity of ComFiR and other methods over HPRD dataset

Figure 4.19. Positive Predictive Value and Sensitivity of ComFiR and other methods on HPRD dataset.

the implication of our ranking approach, we experimented with the Alzheimer's Disease (*OMIM Id-104300*). It is a chronic disease of the nervous system, which mainly affects middle aged persons (40-60 years of age). For this disease, we found out a set of 129 causal genes from GeneCard. Using this information, a set of disease associated complexes were found and finally ranking of these complexes was performed using the algorithm given in Figure 4.21. A few additional genes (not available among the 129 genes in GeneCard) were also found in this process

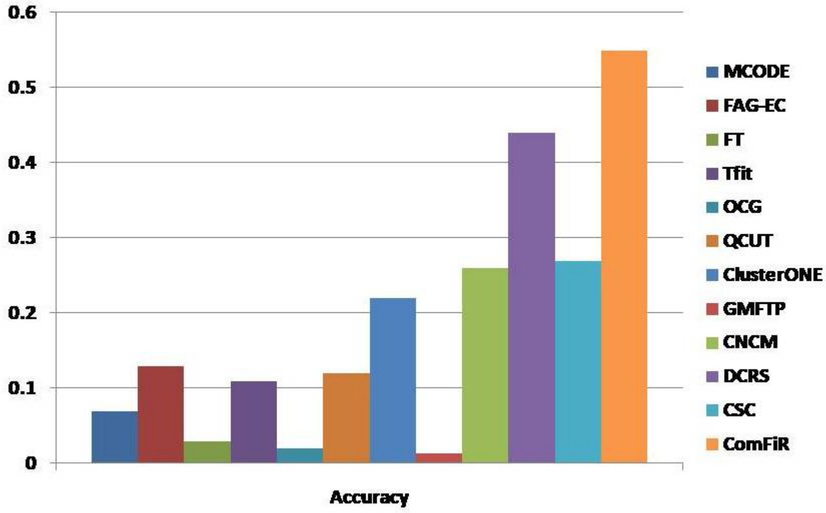


Figure 4.20. Accuracy of ComFiR and other methods on HPRD dataset

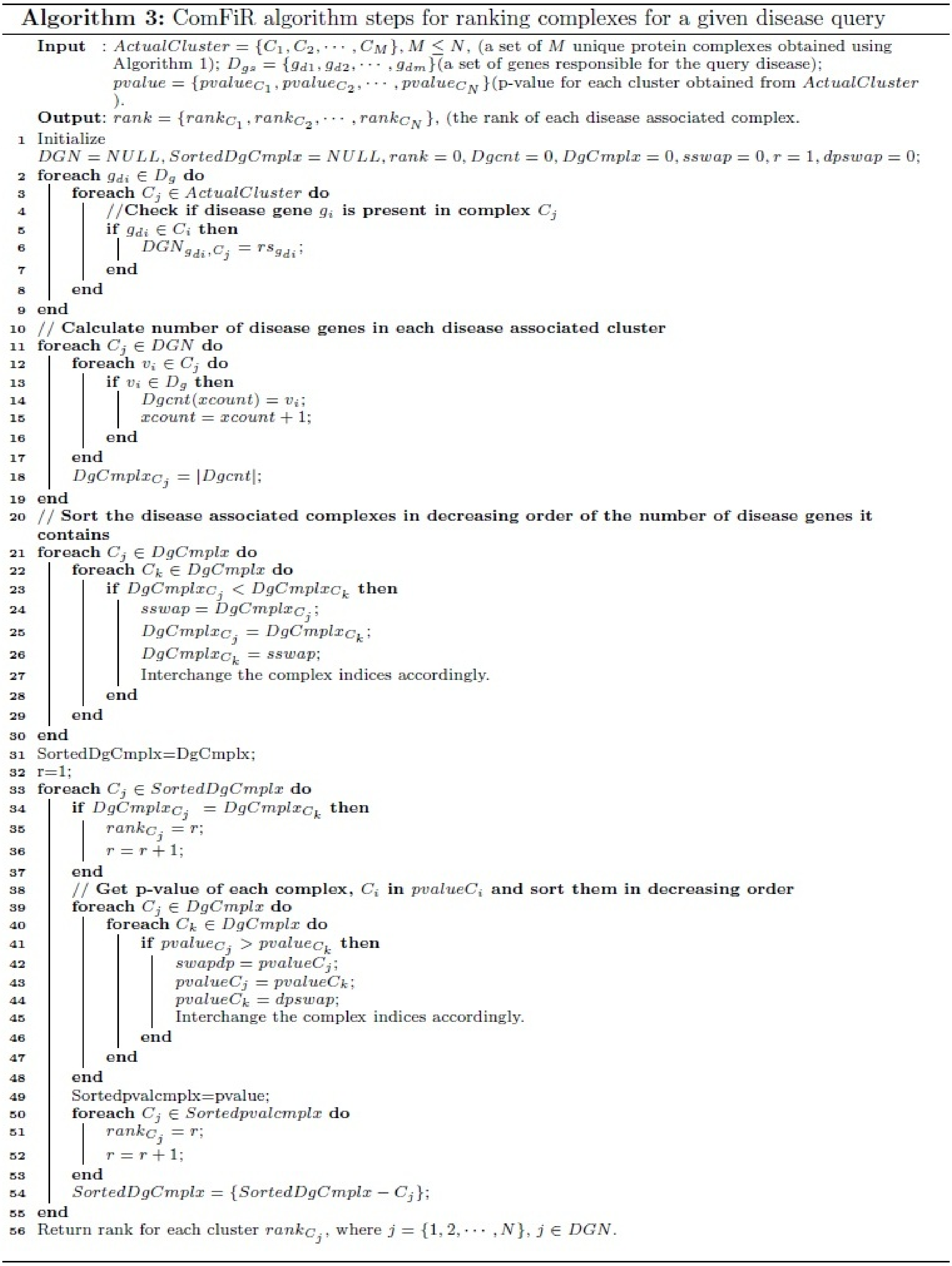
considering connectivity (which might be associated with the disease). These genes were members of top ranked complexes and later these results were supported from various literarature sources.

Table 4.4 lists the top five complexes associated with Alzheimer's Disease along with their member proteins. Let us consider the top ranked complex, i.e., complex number 262 which shows two genes, DLST and MPO as reported by [116]. A third gene, UMPS is also known to be associated with the disease in [19], but it was not present among the 129 causal genes as reported in [116]. The same trend is followed in most complexes found by our method. The top ranked complex has two disease genes as members, therefore, it is given a higher rank, while the other four have only one associated disease gene, and hence the ranking is determined using their p-values. Thus, we could say that by using the existing knowledge to get the rank of the complexes, we could find relevance of some genes in causing the disease which were not available in the repository. This finding can be utilized for diseases for which whose information is already available and can also be extended for unknown diseases with the help of domain experts.

Table 4.4 gives a list of the genes associated with and chromosome numbers. The early onset of Alzheimer's Disease is hypothesized to happen at mutating genes residing in chromosomes 21, 14 and 1. Late onset of the disease can be perceived on chromosome 19. In addition, studies have shown that chromosomes 11 and 12 are also known to be associated with familial history of Alzheimer's Disease. This has been supported by our findings as well. Genes like CPT1A and BCL2L14 (found in

Table 4.4 List of complexes associated with Alzheimer' s Disease

Complex No.	**Complete set**	**Disease associated proteins**	**Chromosome number**	**Common number of disease genes**	**p-value**
262	DLST, MPO,UMPS	DLST [116], **MPO** [116], **UMPS** [19]	DLST (14), MPO(17), UMPS(3)	2	6.290E-4
152	BCL2, BCL2L14, SPNS1, STX18, ITM2B, CPT1A, TP53AIP1	**BCL2** [19], **BCL2L14** [19], **STX18** [19], **ITM2B** [116], **CPT1A** [19]	BCL2 (18), BCL2L14 (12), STX18 (4), ITM2B (13), CPT1A (11)	1	7.767E-5
82	TP53, TNNT1, SHISA5, DHCR24, SMYD2	**TP53** [19], **SHISA5** [92], DHCR24 [116]	TP53 (17), DHCR24 (1), SHISA5 (3)	1	9.642E-5
69	GSK3B, NDC80, FRAT2	GSK3B [116]	GSK3B(3)	1	2.4437E-4
50	NCOA3, GSK3B, ESRRB, ATAD2	**NCOA3** [19], GSK3B [116], **ESRRB** [55], **ATAD2** [100]	GSK3B (3), NCOA3 (20), ESRRB(14), ATAD2 (8)	1	1.118E-3



```
Algorithm 3: ComFiR algorithm steps for ranking complexes for a given disease query
Input : $ActualCluster = \{C_1, C_2, \cdots, C_M\}, M \leq N$, (a set of $M$ unique protein complexes obtained using
        Algorithm 1); $D_{gs} = \{g_{d1}, g_{d2}, \cdots, g_{dm}\}$(a set of genes responsible for the query disease);
        $pvalue = \{pvalue_{C_1}, pvalue_{C_2}, \cdots, pvalue_{C_N}\}$(p-value for each cluster obtained from $ActualCluster$
        ).
Output: $rank = \{rank_{C_1}, rank_{C_2}, \cdots, rank_{C_N}\}$, (the rank of each disease associated complex.
Initialize
$DGN = NULL, SortedDgCmplx = NULL, rank = 0, Dgcnt = 0, DgCmplx = 0, sswap = 0, r = 1, dpswap = 0;$
foreach $g_{di} \in D_g$ do
    foreach $C_j \in ActualCluster$ do
        //Check if disease gene $g_i$ is present in complex $C_j$
        if $g_{di} \in C_i$ then
            $DGN_{g_{di}, C_j} = rs_{g_{di}};$
        end
    end
end
// Calculate number of disease genes in each disease associated cluster
foreach $C_j \in DGN$ do
    foreach $v_i \in C_j$ do
        if $v_i \in D_g$ then
            $Dgcnt(xcount) = v_i;$
            $xcount = xcount + 1;$
        end
    end
    $DgCmplx_{C_j} = |Dgcnt|;$
end
// Sort the disease associated complexes in decreasing order of the number of disease genes it
contains
foreach $C_j \in DgCmplx$ do
    foreach $C_k \in DgCmplx$ do
        if $DgCmplx_{C_j} < DgCmplx_{C_k}$ then
            $sswap = DgCmplx_{C_j};$
            $DgCmplx_{C_j} = DgCmplx_{C_k};$
            $DgCmplx_{C_k} = sswap;$
            Interchange the complex indices accordingly.
        end
    end
end
SortedDgCmplx=DgCmplx;
r=1;
foreach $C_j \in SortedDgCmplx$ do
    if $DgCmplx_{C_j} = DgCmplx_{C_k}$ then
        $rank_{C_j} = r;$
        $r = r + 1;$
    end
    // Get p-value of each complex, $C_i$ in $pvalueC_i$ and sort them in decreasing order
    foreach $C_j \in DgCmplx$ do
        foreach $C_k \in DgCmplx$ do
            if $pvalue_{C_j} > pvalue_{C_k}$ then
                $swapdp = pvalueC_j;$
                $pvalueC_j = pvalueC_k;$
                $pvalueC_k = dpswap;$
                Interchange the complex indices accordingly.
            end
        end
    end
    Sortedpvalcmplx=pvalue;
    foreach $C_j \in Sortedpvalcmplx$ do
        $rank_{C_j} = r;$
        $r = r + 1;$
    end
    $SortedDgCmplx = \{SortedDgCmplx - C_j\};$
end
Return rank for each cluster $rank_{C_j}$, where $j = \{1, 2, \cdots, N\}, j \in DGN$.
```

Figure 4.21. Complex Ranking steps for a given disease query.

chromosome 11 and 12, respectively) are labeled as relevant genes among the top rated relevant complexes. A graphical representation of genes in the top complexes is shown in Figure 4.22.

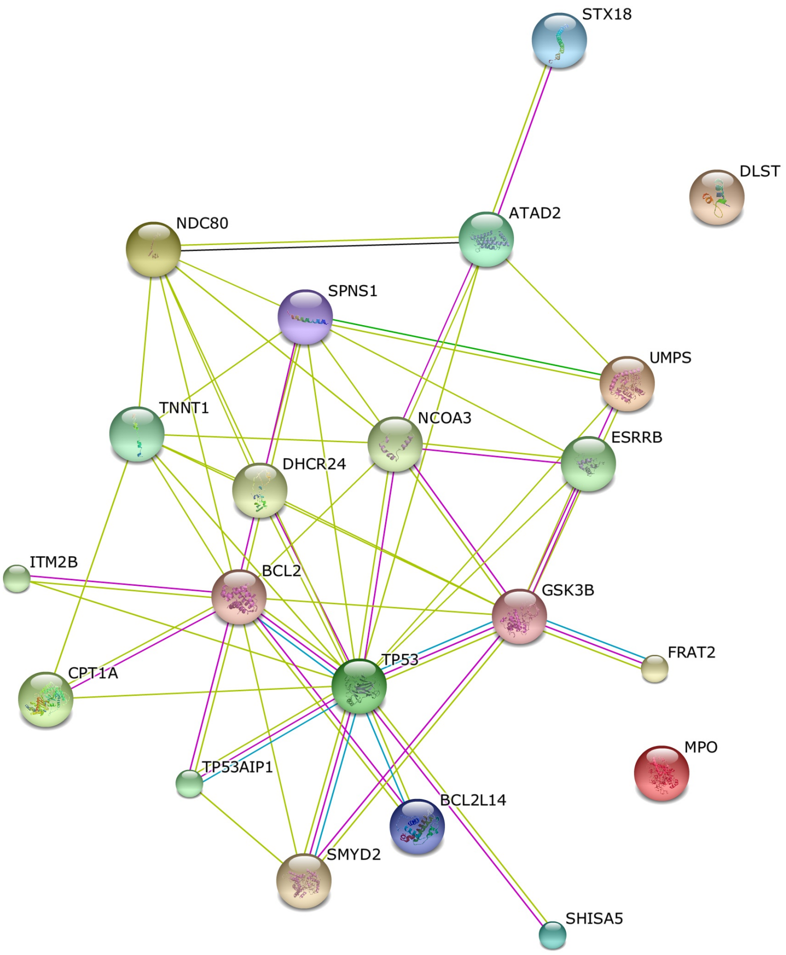


Figure 4.22. Genes found in top five complexes using ComFiR method on Alzheimer's disease (source-STRING tool)

In Figure 4.22, nodes DLST and MPO are isolated, but we find them to be members of the same complex. This can be explained by the limited use of information (physical co-expression, predicted, pathway, co-localization and shared protein domains) in GeneMania while constructing the network. Other genes which are semantically related to the disease gene can be a part of the same complex, as determined by our complex finding approach.

4.4.3 Discussion

ComFiR is a semi-supervised approach for finding protein complexes. In addition to, from the information from already established complexes, it uses both the topological and functional properties of a PPI network to find complexes. The incorporation of some amount of knowledge at the beginning stage has proved effective in raising the accuracy of the overall method. ComFiR has been validated with the DIP as well as the HPRD dataset. We also extended this work to rank the complexes associated with certain diseases. The ranking scheme has been discussed for Alzheimer's Disease here. During the ranking process, a few other genes were discovered to be associated with the disease. Although these genes were individually reported in certain works, they were not yet fully established, hence they were missing from GeneCard. Analyzing properties of such genes could be a remarkable discovery in this field.

4.5 Conclusion

To enhance the accuracy of complex finding for the human PPI dataset, we proposed a method called CSC, which uses a combination of topological as well as biological properties to detect complexes. The performance of CSC was evaluated using yeast as well as human PPI dataset. We also explored the relation between a disease gene in a complex with other non-disease genes based on the information available in GeneMania. This analysis is explained using Alzheimer's Disease. To further enhance the accuracy of complexes obtained, we incorporated some amount of knowledge in our complex finding approach. We proposed a method called ComFiR, which gave an accuracy of around 55% for the HPRD dataset. The performance of ComFiR is unbeaten by any existing methods for this dataset. However, results of both the proposed methods are reported for the optimum threshold. Tuning the parameters would lead to change in the number of complexes and hence would result in the formation of set of new complexes.

We used the complexes obtained using ComFiR to find and rank the disease associated complexes. The ranking scheme proposed was based on the number of disease genes and the p-values. The ranking approach is shown for Alzheimer's Disease here but it can be extended to any other disease. The contribution of this chapter is shown as Publication No.4 & 5 under Publication section.

One important purpose of finding quality complexes from a PPI network is to infer reliable information associated with diseases. Since proteins are the byproducts of genes and very limited knowledge is available regarding interactions among them, it may be difficult to explore certain possibilities which may be of relevance when studying a disease. A suitable analysis of gene expression microarray or Next-Generation Sequencing (NGS) data would come to the rescue of biologists when inferring disease related information. In the next chapter, we explore a few possible ways for finding groups of genes (known as modules) which may be strongly related to certain diseases and further analyze their association among themselves.