Chapter 5

5 Complex Detection from PPI Data Using Unsupervised Methods

With the advent of high-throughput techniques in molecular biology, a significant amount of physical interaction data has been collected leading to computational approaches to systematically mine complexes from the network of physical interactions among proteins, i.e., from Protein-Protein Interaction (PPI) networks. PPI networks can be used for discovering complexes consisting of proteins that share a common function. This is motivated by the observation that proteins are organized into different protein complexes each performing some specific task in a cell [310,311]. Furthermore, proteins belonging to a specific complex are more related to each other than to the members of other complexes [312] and also proteins interacting with each other often participate in the same biological processes.

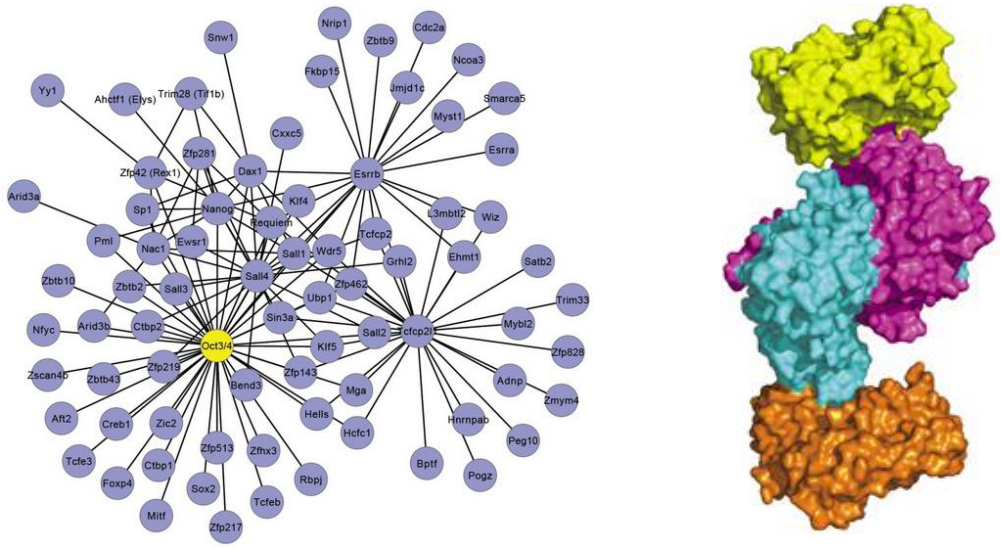
Protein complexes perform many crucial tasks within living beings, including transcription of DNA, translation of mRNA, cell growth and transporting molecules from one place to another. Since proteins perform their tasks by interacting with each other, determining these interactions is an important task. The rapidly growing biomedical literature provides a significantly large and readily available source of interaction data, which can be integrated into the protein networks for better complex detection. Moreover, in-depth study of protein complexes helps to understand how

they are built and how they work, allowing better comprehension of biological systems. It has been well established that the main cause of many diseases are proteins [313,314,315,316]. Therefore, correct classification of a newly discovered protein complex is important as it may guide discovery of appropriate drugs.

In this chapter a review and evaluation of state-of-the-art techniques for computational identification of protein complexes using various evaluation metrics is presented. These techniques adopt different strategies to detect protein complexes and hence obtain different results. A framework for an ensemble method is developed to detect protein complexes from PPI networks by incorporating the challenges that have been uncovered while reviewing the existing protein complex clustering algorithms. The ensemble method is then validated using real life data, with satisfactory results.

5.1 Introduction

Over the years many algorithms have been proposed to detect protein complexes in protein-protein interaction (PPI) networks. A sample PPI network is shown in *Figure* 5.1(a). Proteins are complex organic compounds made up of chains of amino acids used by the body for growth, maintenance and repair of body tissues.



(a) (b)

Figure 5.1:(a) PPI Network (b) Protein Complex

Most proteins form complexes (groups of proteins), as shown in *Figure* 5.1(b), to accomplish biological functions (processes) such as transcription of DNA, translation of mRNA and cell growth. To understand the dynamics of biological processes within an organism, it is necessary to determine the complete map of physical interactions among the proteins (interactome). The result is the PPI network where nodes represent proteins and edges represent interactions between pairs of proteins.

Since an erroneous production of a protein complex can cause diseases by affecting the biological processes in which it is involved, study of protein complexes helps us understand how they are built, allowing us to expand the knowledge of biological systems. New protein complexes can be detected based on the observation that densely connected regions in the PPI networks often correspond to actual protein complexes.

5.2 Related Work

A wide range of detection methods have been proposed for the purpose of identification and categorization of complexes and graph clustering techniques have been found to be useful to handle the computational challenge, since PPI networks are large-scale graphical data structures consisting of tens of thousands of pair-wise protein-protein interactions.

5.2.1 Taxonomy of Existing Clustering Methods

Clustering methods for complex detection in PPI networks can be broadly categorised into distance-based and graph-based approaches [201]. Distance-based clustering approaches use the concept of distance between two proteins as described by vectors of features [211,317,318,319] whereas graph-based clustering techniques mainly consider the topology of the network. In this work, we mainly focus on methods that use only graph topology for the purpose of detecting complexes.

Graph-based techniques that are used for protein complex detection in PPI networks can be classified into three main types of algorithmic approaches, as shown in *Figure* 5.2.

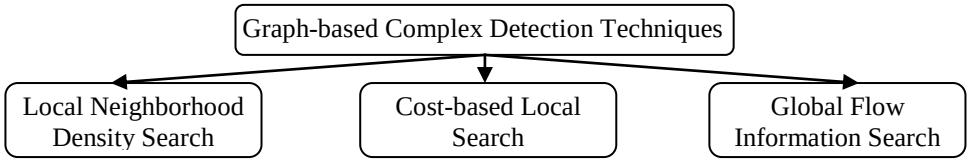


Figure 5.2: Classification of Graph-based Complex Detection Techniques

5.2.1.1 Local Neighborhood Density Search (LD)

The objective of the methods based on local neighbourhood density search is to find dense subgraphs within the input network, i.e., subgraphs where each node is connected to many other nodes within the same subgraph,. Some popular methods for finding complexes in PPI networks based on the LD approach are MCODE [215], DPClus [216], SWEMODE [320], DECAFF [221], CFinder [220], AP [321], ClusterONE [200], CMC [230], PCP [322] and DME [323].

5.2.1.2 Cost-based Local Search (CL)

Methods based on cost-based local search extract complexes from the interaction network by partitioning the graph into connected subgraphs, by using a cost function that guides the search towards the best partition. Typical instances of such an approach are RNSC [210], Qcut [324] and ModuLand [325].

5.2.1.3 Global Flow Information Search (GFI)

Global Flow Information methods for detecting protein complexes in a PPI network are based on an approach that imitates the spread of information in a network. MCL [218] and RRW [217] methods are popular methods based on the concept of random walk. IFB [326] and STM [327] methods exploit biological knowledge for passing information between proteins in the network in order to cluster proteins.

5.2.2 Review of Existing Work

Every new published method compares its performance with a few selected earlier methods. It has been noticed that due to the differences in PPI networks produced, benchmark datasets used for evaluation criteria, threshold settings and parameters used, the results of such comparisons and surveys [226,227,228,229] on complex detection vary and are difficult to reconcile.

Brohee and van Helden [226] performed a comprehensive assessment and empirical comparison among MCODE [215], MCL [218], RNSC [210] and Super-Paramagnetic Clustering (SPC) [211], which was one of the first comprehensive assessments. These algorithms were tested on PPI datasets from high-throughput experiments, and the resultant complexes were evaluated against benchmark complexes from MIPS [328]. The authors concluded that MCL and RNSC outperformed MCODE and SPC in terms of precision and recall. MCL was robust in presence of noise, confirming its superiority over the other three algorithms.

Vlasblom and Wodak [227] compared MCL with Affinity Propagation (AP) [321] on unweighted as well as weighted PPI networks. They concluded that MCL performed considerably better than AP in terms of accuracy and separation of predicted clusters and was able to discover benchmark complexes even at high noise levels.

A detailed comparison of several algorithms: MCODE [215], RNSC [210], MCL [218], DPClus [216], CFinder [220], DECAFF [221], CORE [329], and COACH [222], was performed by Li *et al.* [228] on PPI datasets from DIP [330] and Krogan *et al.* [331]. Based on the Bader overlap score [215], the precision, recall and F-measure values were calculated with the conclusion that MCODE was able to achieve the highest precision, but it produced very few clusters resulting in very low recall.

After studying the analyses and surveys, it has been observed that evaluation of complexes predicted by a method is generally performed using the following approaches.

a) Comparing the predicted complexes against one or more "gold standard" sets of complexes by using performance measures such as *precision* and *recall* [230]. Problems arise due to the reliability of the method if it produces too many predictions (false positives), resulting in high recall but low precision. Since precision and recall have an inverse relationship to each other, a harmonic mean of precision and recall called F-measure is used to obtain a more unbiased performance metric.

b) If a gold standard set is not available, measures such as cluster *cohesiveness* and *separability* are used [215,210]. The topological characteristics of a cluster

such as its size or density is measured by its cohesiveness, while separability measures how separated is the cluster from others [215]. A combination of cohesiveness and separability produces a likelihood of individual clusters representing real complexes although separability does not favour overlapping clusters.

c) Evaluation of the predicted complexes can also be performed by computing a *functional* or *co-localization* score, after incorporating appropriate annotation data [230,210]. This measures how functionally coherent the proteins are within a predicted complex by comparing the GO terms associated with the proteins, and whether they are co-localized within the cell. This evaluation is particularly useful for biological relevance of the predictions.

5.3 Motivation for an Empirical Study

Over the years, the gold standard data used for evaluation have become more enriched, new methods have emerged [230,217,200] and new evaluation measures have been proposed, prompting the need for a new evaluation study of protein complex prediction algorithms. A comparison of eight algorithms, namely RNSC [210], AP [321], MCL [218,332], MCODE [215], CFinder [220], CMC [230], RRW [217] and ClusterONE [200] is carried out. To evaluate the quality and accuracy of predicted complexes, the measures *Frac* [200], *MMR* [200] and *Acc* [226] are used and for the biological relevance, *Co-localization* [333] and *GO semantic similarity* [275] has been used.

5.4 Terminology

In this section some basic terminologies for graphs are introduced.

- PPI data can be represented in the form of a graph $G = (V, E)$, comprising of vertices V, which are proteins and edges E, where each edge represents an interaction between two distinct proteins.

- A walk in graph *G* is defined as a sequence of vertices where there is an edge between two adjacent vertices.

- The set of all the neighbours of a vertex $v \in V$ is denoted as $N_v = \{u | u \in V, (u, v) \in E\}$.

- The degree of v is the cardinality of N_v, written as $deg(v)$.

- The density of G [334], denoted as $den(G)$, is defined as $den(G) = \frac{2 \text{ x } |E|}{|V| \text{ x } (|V|-1)}$.

- The neighbourhood graph of v is defined as $G_v = (V', E')$ where $V' = \{v\} \cup N_v$ and $E' = \{(u_i, u_j) | (u_i, u_j) \in E, u_i, u_j \in V'\}$.

- The neighbourhood graph G_v is the subgraph which consists of all of *v*'s immediate neighbours (including *v*) and all the edges among them.

- A *k*-core is a subgraph in which all the vertices have degrees no less than *k* and the order of a *k*-core is *k* if it is not a (*k* + 1)-core.

- Given two graphs $A = (V_A, E_A)$ and $B = (V_B, E_B)$, the neighbourhood affinity [215] measures the similarity between *A* and *B*, and is defined as

$$\boldsymbol{NA(A,B)} = \frac{|\boldsymbol{V_A} \cap \boldsymbol{V_B}|^2}{|\boldsymbol{V_A}| \text{ x } |\boldsymbol{V_B}|}. \quad (5.1)$$

5.5 Selection of Unsupervised Methods

Conventional graph clustering approaches mine for cliques or densely connected subgraphs, which could correspond to protein complexes. Based on graph theory, there have been various algorithms that have been proposed in literature. Out of the vast plethora of algorithms, a few algorithms for which the software implementations are available have been selected, forming a good representative collection of the existing techniques, and experiments have been performed on them for comparison purposes. The implemented algorithms selected for comparison are MCL [218,332], MCODE [215], RNSC [210], CFinder [220], RRW [217], AP [321], CMC [230] and ClusterONE [200].

The algorithms that support the use of edge weights are AP, MCL, RRW and ClusterONE. In order to run algorithms not supporting directly weights, namely, MCODE, RNSC, CMC and CFinder, the algorithms were pre-binarized using the threshold values originally suggested by the authors of the datasets. Also, algorithms such as MCODE, CFinder, CMC, RRW and ClusterONE are capable of handling overlapping clusters. *Table* 5.1 presents the clustering algorithms evaluated in this review, chronologically ordered, based on the year in which they were developed.

Table 5.1: Clustering algorithms evaluated

Algorithm	Approach adopted	Input Parameter	Overlapping	Weighted
MCL [218,332]	Global flow information	1	no	yes
MCODE [215]	Local neighbourhood	4	yes	no
RNSC [210]	Cost-based Local Search	3	no	no
CFinder [220]	Local neighbourhood (Local cliques)	1	yes	no
RRW [217]	Global flow information (Random Walk)	3	yes	yes
AP [321]	Local neighbourhood	1	no	yes
CMC [230]	Local neighbourhood (Local cliques)	2	yes	no
ClusterONE [200]	Local neighbourhood (Greedy Approach)	2	yes	yes

5.6 Overview of the Approaches

For each algorithm selected, a brief introduction is presented along with the main characteristics.

5.6.1 MCL (Markov Clustering)

MCL [218,332] attempts to find dense regions in the input graph by exploiting the concept that there are many links within a cluster and few links between clusters. Following this perspective, MCL detects functional modules and protein complexes by simulating random walks in PPI networks. It manipulates the weighted or un-weighted adjacency matrix with two operators called *expansion* and *inflation*. The expansion operator assigns new probabilities for all pairs of nodes while the inflation operator changes the probabilities for all these walks in the graph, increasing the probability of intra-cluster walks and demoting inter-cluster walks. This step enhances

the contrast between regions of strong and weak flow in the graph. Another parameter required by MCL is called inflation and it tunes the granularity of the clustering. Larger inflation values result in smaller clusters, while smaller inflation values generate only a few large clusters.

5.6.2 MCODE (Molecular Complex Detection)

The MCODE [215] algorithm is able to detect overlapping protein complexes but as with most algorithms exploiting the notion of cliques, it cannot handle weighted input networks. It initially weighs every node based on local neighborhood densities. Then starting from the node with the highest degree, a protein complex is grown from each node, regulated by a parameter called *depth limit*, which regulates how far it can continue from the seed node to other nodes. MCODE controls the amount of difference allowed between the score of each node in a particular complex using another parameter, the *vertex weight percentage.*

Two post-processing steps applied at the end of the algorithm to refine the protein complexes are *haircut*, which iteratively removes nodes that are connected by a single edge to the rest of protein complex, and *fluffing* that tries to expand a protein complex by adding highly connected nodes outside the cluster. Even though MCODE produces overlapping complexes during the fluffing phase, the experiments have shown that the algorithm performs better when fluffing is turned off. The number of predicted complexes is generally small and the *size* of many predicted complexes is often too large.

5.6.3 RNSC (Restricted Neighbourhoods Search Clustering)

RNSC [210] detects protein complexes based on both graph-theoretical and gene-ontological properties but cannot detect overlapped clusters. Since it is unable to deal with weighted edges, it calculates the value of the cost function by computing a summarized value of two scores, called *naïve* score and *scaled* score for each node, to detect sub-graphs. It starts with an initial random clustering and then searches for a better clustering with the minimum cost by moving a vertex from one cluster to another. Therefore it is a cost-based local search algorithm and is prone to finding poor local minima.

RNSC discards unpromising clusters based on their size, density and functional homogeneity (smallest *p*-value). It also predicts relatively fewer complexes and its results depend on the quality of the initial clustering (random or user-defined). As a result, different runs of this algorithm on the same input data can result in different clusters.

5.6.4 CFinder (Clique Finder)

It is one of the first methods to detect overlapping clusters in PPI networks. Taking as input a parameter *k*, CFinder [220] detects all the *k*-cliques of the input network. A *k*-clique is a complete sub-graph of *k* nodes. It builds a *k*-clique accessibility graph where two *k*-cliques are connected if they are adjacent, i.e., if they share (*k* - 1) common nodes. A *k*-clique percolation cluster is then constructed by linking all the adjacent *k*-cliques into a bigger sub-graph.

Since the original version of the algorithm operated on undirected, unweighted networks, an improvement proposed was to substitute the *k*-clique search with the enumeration of the maximal cliques with at least *k* vertices of the input network. Though it took care of weighted networks, it still could not detect overlapping clusters and was more time consuming than the original algorithm.

5.6.5 RRW (Repeated Random Walk)

The RRW [217] clustering algorithm is able to handle weighted and unweighted graphs, enabling the detection of overlapping clusters. Given a cluster of nodes, the algorithm tries to expand it to include proteins with high proximity to the cluster using an affinity function. Random walks with restart are used to find the set of proteins near a certain cluster. Starting from a cluster of size one, it iterates this expansion at most *k* times (which is an input parameter) or until a stopping condition related to the number of nodes in the cluster is reached, producing a cluster of size < *k*. The process is applied to all the nodes followed by a ranking step that removes clusters with an overlap score above a given threshold.

The restart probability of the random walk at each step is given to the RRW algorithm along with the threshold overlap parameters and the early cut-off. The maximum cluster size can be kept at eleven as recommended by the authors.

5.6.6 AP (Affinity Propagation)

AP [321] is a clustering algorithm that can handle real-valued similarities between the input objects but cannot detect overlapping clusters. It is a *k*-centers clustering technique, which uses the input data to learn a set of centers such that the sum of squared errors between each input object and its nearest centre is minimum. The centres are referred to as *exemplars*. This method begins with an initial set of exemplars, randomly selected. The exemplars are refined iteratively with an aim to decrease the sum of squared errors.

Since the clustering is sensitive to the initial selection of exemplars, Affinity Propagation deals with this limitation by considering each data point as a node in a network. Each node iteratively sends a message to all the other nodes communicating its relative attractiveness to them and receives a response about their relative availability. By using this message passing procedure, the nodes try to identify the best exemplar with respect to a particular function. The name of this algorithm comes from the fact that the magnitude of each message exchanged at a certain time point resembles the affinity that a particular data point has for another data point as its exemplar.

5.6.7 CMC (Clustering based on Maximum Cliques)

The CMC [230] algorithm is a clustering algorithm that assesses the probability whether two proteins are in the same protein complex by using an iterative scoring algorithm, followed by a maximal clique finding process. As with most clique-based algorithms, it is not able to handle weighted networks. However it can output overlapping protein complexes.

During the search process the cliques found are merged to build the final set of protein complexes. Two cliques are considered sufficiently overlapping using an overlap threshold, while a merge threshold determines when two cliques should be

merged together. Two cliques are merged when the network between them is denser than the merge threshold otherwise the smaller clique is discarded. A low overlap threshold implies the detection of only a few big protein complexes, while a high value results in a large number of redundant complexes. The degenerate case occurs when none of the complexes is able to merge with others.

5.6.8 ClusterONE (Cluster Overlapping Neighbourhood Expansion)

ClusterONE [200] is a method for detecting overlapping protein complexes from protein-protein interaction data. It can handle weighted PPI data and overlapping clusters simultaneously. This algorithm uses a greedy approach to calculate a score called *cohesiveness* which measures how likely it is for a group of proteins to form a protein complex. The procedure starts by selecting the protein with the highest degree as the first seed and grows a cohesive group from it using a greedy approach. When the growth process finishes the algorithm selects the next seed from the proteins having the highest degree, but not already included in any protein complex. The process ends when there are no remaining proteins to consider.

The next step merges the cohesive groups based on their overlap score. For a given a set of clusters, an overlap score is computed for each pair of clusters and a graph in which each node represents a cluster is constructed. If two nodes overlap more than a certain threshold, they are connected and the clusters connected to each other by a path of edges are merged, resulting in protein complexes.

5.7 Evaluation of Protein Complexes

To evaluate the performance of the prediction of protein complexes, a comparison of how well a predicted protein complex matches an actual complex in a set of gold standard protein complexes, is made. Unfortunately, one of the main issues that arise during comparison is that the match between predicted complexes and a gold standard complex is often only partial. Furthermore, a protein in a gold standard complex can match proteins contained in more than one predicted complex and vice versa.

5.7.1 Evaluation Measures

Three quality measures are used to independently assess the similarity between a set of predicted complexes and a set of reference complexes: (i) the fraction of protein complexes [200] matched by at least one predicted complex, (ii) a geometric accuracy measure [226], and (iii) the Maximum Matching Ratio [200].

It may be noted that all these measures assess the quality of a predicted protein complex comparing it with protein complexes present in a specific gold standard. Hence, it is not possible to establish the quality of a predicted complex if it does not match, even partially, at least one gold standard complex.

5.7.1.1 Frac

The first measure used is the fraction of pairs (*Frac*) [200] between predicted and reference complexes (gold standard) that match with an overlap score ω greater than 0.25, since this threshold indicates that the intersection of two complexes having the same size is at least half of the complex size [200]. The overlap score between a predicted complex *p* and a real complex *b* is given by

$$NA(p,b) = \frac{|V_p \cap V_b|^2}{|V_p| \times |V_b|}. \quad (5.2)$$

If $NA(p,b) > \omega$, the complexes are considered to match.

5.7.1.2 Geometric Accuracy

The second measure, introduced by Brohee and van Helden [226], is geometric accuracy [226], which is the geometric mean of clustering-wise sensitivity (*Sn*) [226], and the positive predictive value (*PPV*) [226] of a clustering. Both can be computed from the confusion matrix $T = [t_{ij}]$ of the complexes.

With *n* reference and *m* predicted complexes, the element t_{ij} of the confusion matrix refers to the number of proteins that are found both in the reference complex *i* and the predicted complex *j*. If n_i is defined as the number of proteins in reference complex *i*, then

$$Sn = \frac{\sum_{i=1}^{n} max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i}, \text{ and} \quad (5.3)$$

$$PPV = \frac{\sum_{j=1}^{m} max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} t_{ij}}. \quad (5.4)$$

Clustering-wise sensitivity (*Sn*) can give inflated values if every protein is placed in the same cluster, whereas the positive predictive value can be maximized by putting every protein in its own cluster. In other words, if a method predicts a giant complex which covers many proteins in a known real complex set, this method will get a very high *Sn* score whereas the *PPV* value tends to be lower as it does not evaluate overlapping clusters properly. For these reasons the two measures are balanced by computing their geometric mean, known as the geometric accuracy (*Acc*), which is defined as:

$$Acc = \sqrt{Sn \text{ x } PPV}. \quad (5.5)$$

5.7.1.3 MMR

The Maximum Matching Ratio (*MMR*) [200] is used to express how well the predicted complexes represent reference complexes, i.e., how well the prediction compares with a gold standard. MMR represents the two sets of predicted complexes as a bipartite graph where on one side are the predicted complexes and on the other the reference ones. Each node of this graph represents a protein complex and each edge connecting two nodes has a weight reflecting the overlap between these complexes, where the overlap score between two protein complexes is computed by Equation (5.2). MMR [200] is obtained by dividing the total weight of the *maximum matching* complexes by the number of *reference* complexes. Given *n* standard complexes and *m* predicted complexes, let *j* be a member of the predicted complexes. MMR then defined as follows:

$$MMR = \frac{\sum_{i=1}^{n} max_{j}^{m} NA(p,b)}{n}. \quad (5.6)$$

The *PPV* scores tend to be lower if predicted complexes overlap significantly with each other. Thus clustering algorithms supporting overlapped clusters may return a

lower positive predicted value than a not so efficient algorithm which places every protein in a separate cluster.

Moreover, a predicted complex that may not match any reference complex is penalized by the geometric accuracy measure. However, a predicted complex that does not match any reference complex is not necessarily an undesired result and may be an undiscovered complex. Therefore, trying to optimize the geometric accuracy may discourage detection of such new complexes in a PPI network. This is one motivation behind the use of MMR as a metric.

5.7.2 Gold Standard for Protein Complexes

As mentioned earlier, to assess the performance of a clustering algorithm on PPI networks, it is necessary to compare protein complexes it predicts with a set of known interactions called gold standards. The Saccharomyces Genome Database (SGD) [335] and the Munich Information Centre for Protein Sequences (MIPS) [328] *S. cerevisiae* Protein-Protein Interaction dataset have been considered because they have been used in several analyses as gold standard reference due to their quality and comprehensiveness. General properties of these datasets are shown in *Table* 5.2.

Table 5.2: General properties of the Gold Standard datasets

General Properties	SGD	MIPS
No. of Proteins	1279	1189
No. of Complexes	323	203
Overlapping Proteins	296	401

5.8 Comparison of Selected Algorithms

It is important to avoid biases when well-known graph clustering algorithms methods are applied to a new scenario while evaluating their performance. Substantial care has been taken to avoid over-optimization of algorithm parameters to a given dataset or to a given quality score following guidelines given by Boulesteix [336].

a) Each of the algorithms have been tested on three publicly available high-throughput benchmark yeast PPI datasets, namely Gavin [337], Krogan Core [331] (referred to as Krogan in the thesis) and Krogan Extended [331] (referred in the

thesis as Krogan+). Some details of the datasets are given in *Table* 5.3.

b) More than one quality score has been used to assess the quality of each set of protein complexes detected by the algorithms. The scores are the fraction of matched complexes [200] with a given overlap score threshold, the geometric accuracy [226] and the maximum matching ratio [200].

c) Two different gold standards: a set derived from the Gene Ontology annotations of the Saccharomyces Genome Database [335] and the MIPS compendium of protein complexes [328] have been used.

d) For each clustering algorithm tested, the input parameters are tuned in order to achieve the best performance for protein complexes detection in Protein-Protein Interaction Network. Tuning was performed separately for each measure, for each dataset and with respect to each gold standard. As a result, different input parameters have been obtained for each case with the purpose to obtain the most optimistic score.

Table 5.3: Initial datasets

Details	Gavin	Krogan	Krogan +
Number of proteins	1855	2708	3672
Number of interactions	7669	7123	14317
Weighted	yes	yes	yes

5.8.1 Parameter Settings for Each Algorithm

The parameters for each algorithm were set after experimenting with a different combination of settings unless the authors of the original algorithm had explicitly suggested particular settings. The values of the parameters for optimal performance for each algorithm are reported in *Table* 5.4.

It has been noticed that the optimal parameter values for non-overlapping algorithms seem to vary from one dataset to another whereas algorithms that allow overlapping clusters seem to show more stable performance if their parameters remain within a certain range.

Since the MIPS and SGD gold standard datasets are not entirely consistent with respect to the membership of some proteins in some complexes, it is decided to test these two gold standards separately as in [200].

Table 5.4: Parameters settings for the algorithms being evaluated

Algorithm	Parameter	Using Gold Standard MIPS			Using Gold Standard DIP		
		Optimal Value for			Optimal Value for		
		Gavin	Krogan	Krogan+	Gavin	Krogan	Krogan+
MCL	Inflation	3.2	2.3	2.3	4.7	2.0	2.6
MCODE	Depth Limit	3	3	3	3	3	3
	Vertex Weight Percent	10	20	10	10	20	10
	Haircut	True	True	True	True	True	True
	Fluff	False	False	False	False	False	False
RNSC	No. of Experiments	3	3	3	3	3	10
	Tabu Length	100	10	50	100	50	50
	Scaled stopping tolerance	5	5	1	15	1	5
CFinder	k-clique template size	4	3	3	4	3	4
RRW	Restart Probablity	0.6	0.5	0.5	0.6	0.5	0.5
	Overlap Threshold	0.1	0.2	0.2	0.1	0.2	0.2
	Early cutoff	0.6	0.7	0.7	0.6	0.7	0.7
AP	Preference	-0.15	0.35	0.4	-0.6	0.35	0.3
CMC	Overlap Threshold	0.7	0.7	0.7	0.7	0.7	0.7
	Merge Threshold	0.5	0.4	0.5	0.5	0.4	0.3
ClusterONE	Merging Threshold	0.8	0.8	0.8	0.8	0.8	0.8
	Density Threshold	0.3	0.3	0.3	0.3	0.3	0.3

5.8.2 Quality Scores for MIPS Gold Standard

The results obtained by comparing the protein complexes predicted by each clustering algorithm with the gold standard dataset obtained from the MIPS catalogue of protein complexes are reported in this section.

Table 5.5 shows for each dataset the true number of complexes, the matched number of complexes detected by the algorithms, the fraction of protein complexes matched by at least one predicted complex, geometric accuracy and maximum matching ratio using the MIPS gold standard dataset. ClusterONE detects the highest number of complexes that match at least one real complex, for each of the three quality measures, *Frac*, *Acc* and *MMR*, for the three datasets used for comparison. The Maximum Matching Ratio was explicitly designed to assess the quality of overlapping protein complexes and the results achieved by ClusterONE show a significantly better MMR value than the ones obtained by the other approaches. The best scores are highlighted in bold for easy comparison.

Table 5.5: Results of the protein complex detection algorithms on PPI datasets using the MIPS complex dataset

Algorithm	Gavin					Krogan					Krogan+				
	#c	#m	Frac	Acc	MMR	#c	#m	Frac	Acc	MMR	#c	#m	Frac	Acc	MMR
MCL	251	82	0.711	0.511	0.338	378	85	0.617	0.450	0.276	489	72	0.448	0.421	0.196
MCODE	137	67	0.585	0.457	0.289	77	39	0.289	0.332	0.122	67	26	0.155	0.279	0.102
RNSC	139	73	0.648	0.495	0.323	88	58	0.400	0.394	0.179	96	58	0.362	0.375	0.153
CFINDER	139	68	0.592	0.500	0.286	118	50	0.358	0.380	0.170	124	38	0.225	0.324	0.112
RRW	238	79	0.684	0.457	0.353	323	71	0.518	0.370	0.251	233	77	0.481	0.365	0.224
AP	249	77	0.686	0.454	0.346	225	57	0.411	0.355	0.177	238	55	0.350	0.343	0.169
CMC	345	78	0.675	0.484	0.342	151	53	0.373	0.378	0.174	428	61	0.382	0.346	0.180
ClusterONE	198	**84**	**0.738**	**0.513**	**0.390**	526	**95**	**0.692**	**0.451**	**0.329**	531	**94**	**0.593**	**0.435**	**0.294**

Abbreviations: #c = no. of complexes, #m = no. of matched complexes

The values of the evaluation metrics for all clustering algorithms are along the *x-axis* and the individual quality scores of the predicted complexes for the MIPS catalogue are along the *y-axis* in *Figure* 5.3.

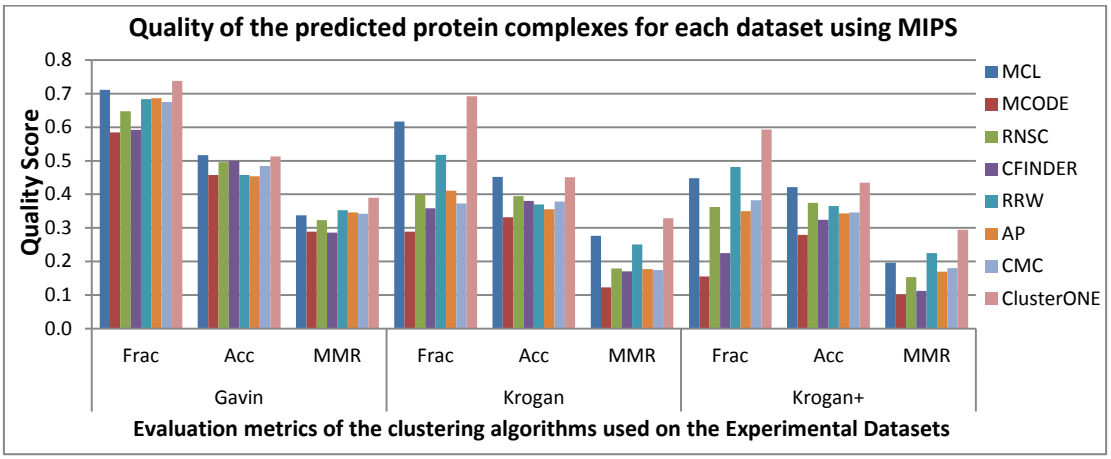


Figure 5.3: Graph showing the comparative analysis of the evaluation metrics of the clustering algorithms for MIPS

It is clear from *Figure* 5.3 that ClusterONE outperforms the other algorithms in all datasets by achieving the highest score in each of the three categories used for comparison.

5.8.3 Quality Score for SGD Gold Standard

Table 5.6 reports the scores for the number of complexes detected, the number of matched complexes, *Frac*, *Acc* and *MMR* for the three datasets achieved by the algorithms using the SGD gold standard dataset. Again ClusterONE achieves the best results in protein complex detection using all three quality measures, for all three datasets.

Table 5.6: Results of the protein complex detection algorithms on PPI datasets using the SGD complex dataset

Algorithm	Gavin					Krogan					Krogan+				
	#c	**#m**	**Frac**	**Acc**	**MMR**	**#c**	**#m**	**Frac**	**Acc**	**MMR**	**#c**	**#m**	**Frac**	**Acc**	**MMR**
MCL	251	96	0.776	0.710	0.447	375	105	0.662	0.659	0.357	487	92	0.509	0.612	0.258
MCODE	134	74	0.598	0.627	0.367	76	58	0.364	0.474	0.202	69	39	0.216	0.394	0.102
RNSC	136	91	0.736	0.715	0.434	88	78	0.486	0.556	0.265	93	78	0.432	0.552	0.241
CFINDER	132	78	0.630	0.688	0.367	114	69	0.433	0.509	0.248	124	49	0.271	0.484	0.158
RRW	233	97	0.785	0.687	0.480	324	100	0.627	0.578	0.368	235	101	0.559	0.545	0.317
AP	248	99	0.800	0.676	0.446	221	81	0.507	0.525	0.306	235	97	0.537	0.529	0.291
CMC	348	95	0.768	0.661	0.453	156	69	0.433	0.526	0.245	423	78	0.432	0.540	0.257
ClusterONE	198	**102**	**0.817**	**0.727**	**0.486**	526	**112**	**0.690**	**0.683**	**0.426**	531	**110**	**0.615**	**0.647**	**0.371**

Abbreviations: #c = no. of complexes, #m = no. of matched complexes

From *Table* 5.5 and *Table* 5.6, it is observed that the number of complexes detected by ClusterONE from the Gavin dataset (198 complexes) is closer to the actual number of MIPS complexes, i.e., 203, whereas the results from the Krogan and Krogan+ datasets seem to contain a large number of extra clusters, i.e., 526 and 531. This may be explained by the fact that some of the MIPS categories are not real protein complexes but groups of related complexes. The same seems to be the case for the rest of the algorithms, except in the case of RNSC and MCODE (*Table* 5.5) as these algorithms tend to detect fewer clusters.

Figure 5.4 shows the plot of the evaluation metrics of all the clustering algorithms for SGD and it is noticed that ClusterONE algorithm outperforms the other approaches for all datasets.

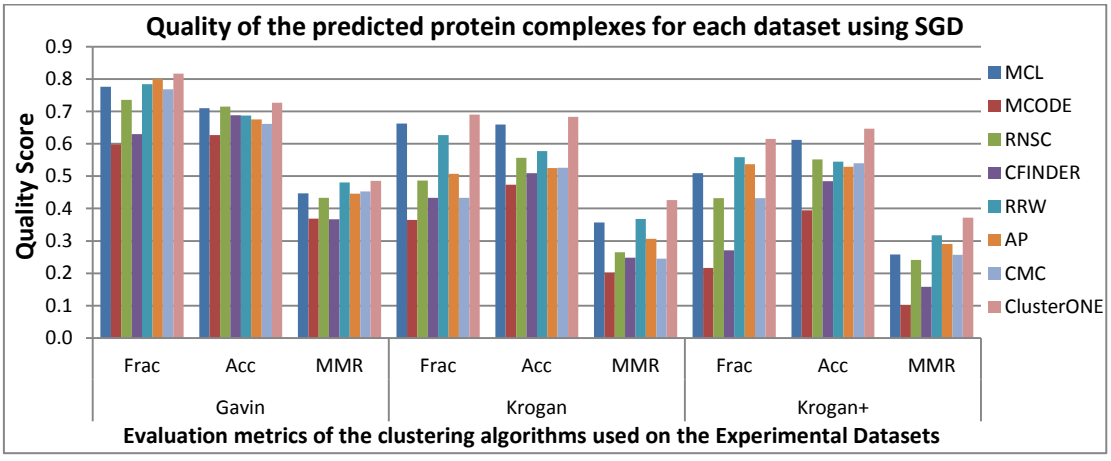


Figure 5.4: Graph showing the comparative analysis of the evaluation metrics of the clustering algorithms for SGD

5.9 Biological Coherence of Predicted Complexes

As mentioned earlier [338], the gold standard protein complex sets are incomplete and as a result, a predicted complex that does not match any of the reference complexes may belong to a valid but previously un-cataloged complex as well. Therefore, relying on the comparison measures outlined in *Section* 5.7, based on a pre-defined gold standard dataset may not give a whole picture. Rather the scores should be supported by measures that assess the biological relevance of predicted complexes, based on the similarity of their functional annotations or the co-localization of the constituent proteins.

5.9.1 Co-Localization Similarity

Biological relevance should play an important role in evaluating the quality of predicted protein complexes. One way this can be done is by using co-localization scores [339]. Since proteins in the same protein complex have a tendency to share common functions, they tend to be located at the same location in a cell. The maximum fraction of proteins in a complex found at the same location is known as the co-localization score for that complex. The average co-localization score is calculated as the weighted average over all complexes and is defined as [339]

$$C = \frac{\sum_j max_i l_{i,j}}{\sum_j |A_j|}. \tag{5.7}$$

Here, $l_{i,j}$ is the number of proteins of complex A_j assigned to the localization group i and $|A_j|$ is the number of proteins in the complex A_j with localization assignments. A high co-localization score usually indicates a high functional similarity between proteins in the same complex.

The localization dataset published by Huh et al. [333] and ProCope [340], a popular tool used to predict and evaluate protein complexes, are used to calculate co-localization scores. *Table* 5.7 shows the co-localization scores of protein complexes detected by the eight algorithms on Gavin and Krogan datasets. The highest co-localization score of 0.746 on the Gavin dataset is achieved by ClusterONE, which is higher than the

next score of 0.735 obtained by MCL. The co-localization score of ClusterONE on the Krogan dataset is 0.723, which is nearly the same as 0.725 obtained by MCL. However, the scale tilts in favor of ClusterONE since MCL cannot handle overlaps. Hence, we judge that the protein complexes detected by ClusterONE have relatively high quality from the biological view point due to high co-localization scores on both Gavin and Krogan datasets.

Table 5.7: Comparison of Co-localization Score

Algorithm	Co-localization Score	
	Gavin	Krogan
MCL	0.735	0.725
MCODE	0.722	0.671
RNSC	0.618	0.584
CFINDER	0.615	0.59
RRW	0.644	0.611
AP	0.721	0.633
CMC	0.628	0.655
ClusterONE	0.746	0.723

Figure 5.5 presents the co-localization score of the algorithms visually.

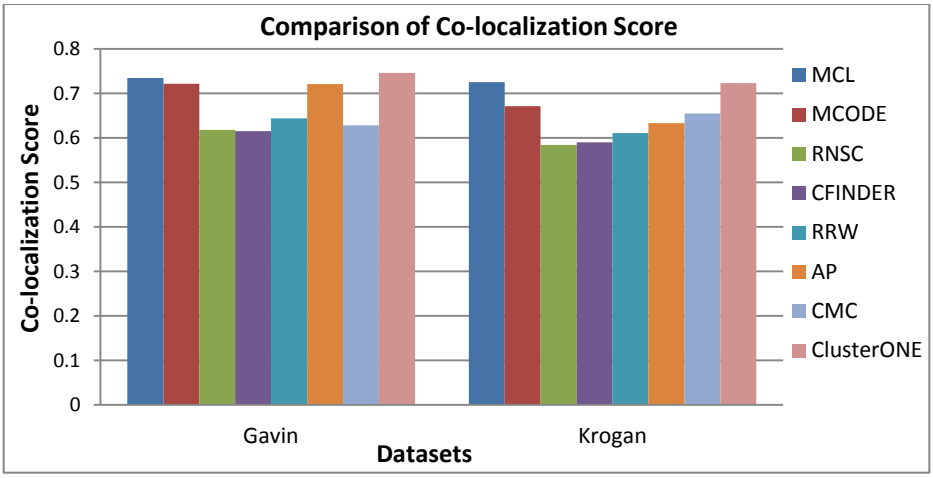


Figure 5.5: Comparison of Co-localization Score

5.9.2 GO Semantic Similarity

Comparing GO terms associated with proteins in a protein complex is another indicator of biological relevance. This can be evaluated in terms of functional similarity [341]. The functional similarity between two proteins is measured in terms of the semantic similarity of GO terms that are used to annotate these proteins. Out of the several variations of semantic similarity that have been proposed [265,275,286,293], the

semantic similarity score proposed by Schlicker et al. [275] has been selected for use. The GO semantic similarity score for a protein complex is defined as the average of semantic similarity scores of all protein pairs within the protein complex [275]. The GO semantic similarity score for a set of complexes is the geometric mean of all complex scores determined separately for the "biological process" and "molecular function" taxonomies. A higher GO semantic similarity score is associated with better quality for a set of protein complexes.

To calculate the GO semantic similarity score, ProCope [340] has been used. *Table* 5.8 shows the results of comparison of GO semantic similarity scores obtained using the eight algorithms on the Gavin and Krogan datasets. These scores are plotted in *Figure* 5.6 for visual comparison. MCODE scores 0.883 followed by ClusterONE with 0.869 for the Gavin dataset. Thus the complexes mined by MCODE have the highest biological significance, but the number of high-quality complexes identified by MCODE is lower compared to that of ClusterONE. In the case of the Krogan dataset, ClusterONE scores better than MCODE, leading to better quality protein complexes.

Table 5.8: Comparison of GO Semantic Similarity Score

Algorithm	GO Semantic Score	
	Gavin	Krogan
MCL	0.8	0.783
MCODE	0.883	0.855
RNSC	0.726	0.735
CFINDER	0.72	0.67
RRW	0.761	0.665
AP	0.752	0.738
CMC	0.79	0.72
ClusterONE	0.869	0.873

The results of the GO semantic similarity comparison scores on Gavin and Krogan dataset have been plotted in *Figure* 5.6 for a visual comparison.

5.10 Statistical Significance

In this section we test whether the difference between the top two performing algorithms, ClusterONE and MCL, is statistically significant or not.

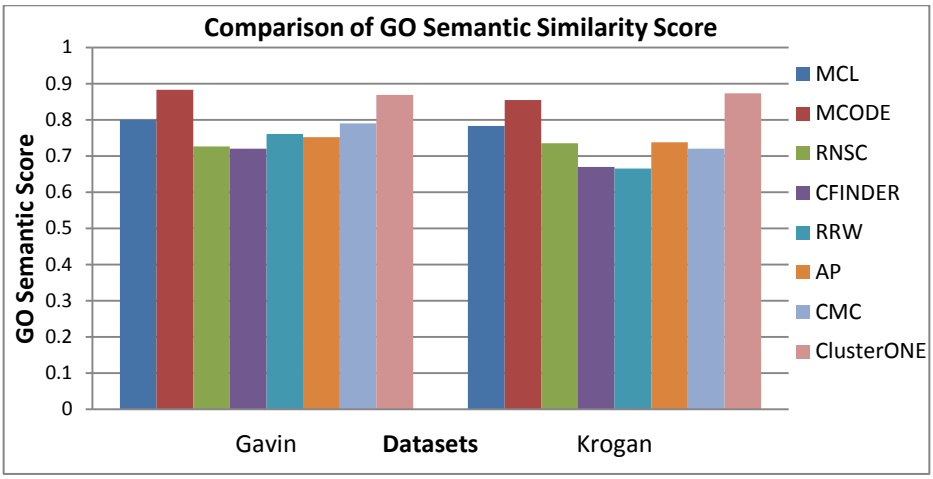


Figure 5.6: Comparison of GO Semantic Similarity Score

5.10.1 Statistical Evaluation of Predicted Complexes

The statistical significance of the occurrence of a protein cluster (predicted protein complex) with respect to a given functional annotation can be computed by the following hyper-geometric distribution in Equation (5.8) in terms of *p-values*.

$$p - value = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i}\binom{|V| - |F|}{|C| - i}}{\binom{|V|}{|C|}} \tag{5.8}$$

where a predicted complex C contains k proteins in the functional group F and the entire PPI network contains $|V|$proteins. The functional homogeneity of a predicted complex is the smallest *p-value* over all the possible functional groups.

A predicted complex with a low *p-value* indicates that it is enriched by proteins from the same function group and it is thus likely to be a true protein complex, so the complex has a high statistical significance. As such, low *p-value* of a predicted complex generally indicates that the collective occurrence of these proteins in the complex does not occur merely by chance and thus the complex has high statistical significance if the *p-value* < 0.0l. [337]

Bihai *et. al* [342] have shown that the proportion of significant complexes over all predicted ones can be used to evaluate the overall performance of various algorithms.

In addition, *P-score* [342] has also been used as an effective evaluation measure for a complex, which is defined in Equation (5.9).

$$P - score = \frac{1}{n}\sum_{i=1}^{n} -\log\,(p - value_i) \tag{5.9}$$

where $p - value_i < Y$and Y is set to 0.01 mentioned above.

5.10.2 Statistical significance of ClusterONE and MCL

To substantiate the statistical significance of the complexes predicted by the two best algorithms ClusterONE and MCL, we calculated the *P-scores* of the top 40 ranked predicted protein complexes based on their low *p-value* that match the true complexes of the benchmark datasets MIPS and SGD. In our experiments, the *p-values* of complexes are calculated with the SGD's GO::TermFinder [343].

The *Table* 5.9 reports for each dataset the comparative measures achieved by the two algorithms on the gold standard MIPS dataset. ClusterONE achieves the largest value of proportion of matched complexes which is approximately 9% to 12% greater than MCL for the three datasets on which the algorithms have been tested. Moreover, the *F-measure* of ClusterONE is 20% to 50% more than that of MCL. The average *P-score* of ClusterONE is seen to be higher than those obtained by MCL, mainly because ClusterONE can detect overlapping clusters properly. A protein pair with a large number of shared neighbours (overlapping) will have a *p-value* very close to zero leading to a higher *P-score*, since the *p-values* are expressed as –log(*p-values*).

Table 5.9: Statistical significance of predicted complexes by ClusterONE and MCL for MIPS Dataset

Dataset	Algorithm	No of Matched Com-plexes	No of Significant Complexes	Proportion of matched complexes (%)	Precision	Recall	F-measure	Average P-Score
Gavin	ClusterONE	**84**	**61**	**72.620**	**0.087**	**0.462**	**0.146**	**3.24**
	MCL	82	52	63.415	0.083	0.115	0.096	2.59
Krogan	ClusterONE	**95**	**73**	**76.842**	**0.089**	**0.433**	**0.148**	**3.49**
	MCL	85	58	68.236	0.077	0.3	0.123	2.97
Krogan+	ClusterONE	**94**	**70**	**74.469**	**0.084**	**0.424**	**0.140**	**3.36**
	MCL	72	47	65.278	0.067	0.273	0.108	2.78

The comparative measures achieved by ClusterONE and MCL algorithms on the gold standard SGD dataset are reported in *Table* 5.10. As was the case with the MIPS dataset, it is noticed that in the case of SGD dataset too ClusterONE follows the same trend and outperforms MCL, showing greater statistical and biological significance of the predicted protein complexes. The average *P-score* of ClusterONE shows an increase of 19%, 15% and 16% over MCL for the Gavin, Krogan and Krogan+ dataset respectively.

Table 5.10: Statistical significance of predicted complexes by ClusterONE and MCL for SGD Dataset

Dataset	Algorithm	No of Matched Com-plexes	No of Significant Complexes	Proportion of matched complexes (%)	Precision	Recall	F-measure	Average P-Score
Gavin	ClusterONE	**102**	**83**	**81.373**	**0.109**	**0.406**	**0.172**	**3.63**
	MCL	96	69	71.875	0.194	0.156	0.173	3.04
Krogan	ClusterONE	**112**	**95**	**84.821**	**0.112**	**0.5**	**0.183**	**3.85**
	MCL	105	81	77.143	0.092	0.313	0.142	3.36
Krogan +	ClusterONE	**110**	**92**	**83.636**	**0.102**	**0.447**	**0.166**	**3.77**
	MCL	92	70	76.087	0.089	0.316	0.139	3.24

5.11 Motivation for an Ensemble Framework

An ensemble has the capacity to achieve consistently well performing results for any dataset by deriving the individual benefits of participating algorithms in terms of noise handling, cluster overlap detection and ability to detect complexes with minimum prior knowledge.

Traditional clustering approaches for protein complex identification are hampered by the following limitations, (i) A PPI network contains a lot of noise and are incomplete, leading to high false-positive rate in identifying interactions; (ii) Classical partitioning schemes dependent on specific initial conditions, seed selection, parameter setting and tend to get different, unstable and unsatisfactory cluster partitions from run to run; (iii) The clusters produced by these methods may lack biological meaning since they are based on a single information source; and (iv) Some proteins are believed to be multifunctional, and effective strategies for the soft clustering of these essential proteins are needed. The challenge, therefore, is how to

effectively integrate multiple biological data sources since a single source may not be adequate.

To tackle these limitations an ensemble method is proposed since ensembles have been able to improve robustness, stability and prediction accuracy of clusterings by combining the output of several algorithms [344]. An ensemble of clustering algorithms may be able to identify larger, denser clusters with improved biological significance. Since the noise in the interaction datasets hampers the detection of accurate protein complexes, one way to compensate for the lack of interaction coverage would be to combine multiple datasets for the purpose of identification of meaningful complexes. Additionally, ensemble-based methods may be preferable for protein complex detection considering that traditional clustering methods search for complexes in "dense" regions and miss complexes of low densities, that is, small complexes of two or three proteins. Moreover, cluster ensembles may also be able to address the problem of local optima and obtain a more globally optimum solution, i.e., a stable clustering solution.

5.12 Protein Complex Detection Ensemble (PCDEN)

A conceptual framework for the proposed Protein Complex Detection ENsemble (PCDEN) is shown in *Figure* 5.7.

The PPI network data is taken as input by n consistently well-performing clustering algorithms BC_1, BC_2, ..., BC_n which generate n individual complexes C_1, C_2, ..., C_n. In this step care must be taken to select the clustering algorithms that are diverse in nature. As a result, the errors made in clustering by one algorithm are likely to be averaged out by the correct clustering of another, so that the overall clustering accuracy is improved and a final unbiased decision can be made.

Next, we take a complex set C_1 generated by the base clustering algorithms BC_1 and identify the node with the highest degree. Similarly, we take the other set of complexes C_2, C_3, ..., C_n generated by the base clustering algorithms BC_2, BC_3, ..., BC_n and identify the corresponding complexes based on the node identified. Now we

identify a common set of nodes (proteins), say P_i, from the corresponding complex sets with the condition that each member element of P_i is present in atleast two complexes. We consider two cases now.

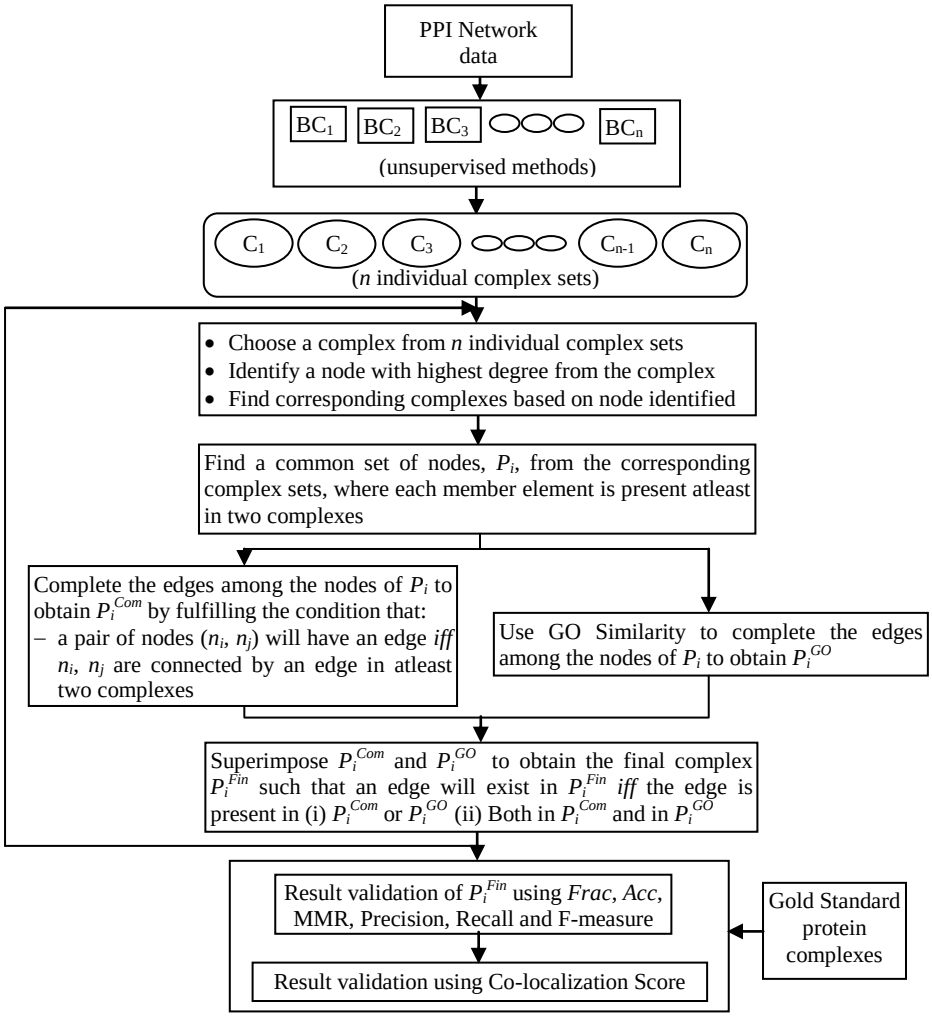


Figure 5.7: Conceptual framework for the proposed protein complex detection ensemble

Case 1: For the common set of nodes P_i, we complete the edges among the nodes of P_i subject to the condition that if n_i and n_j are two nodes present in the common set of nodes P_i, then they will be connected by an edge if and only if (n_i, n_j) are already connected by an edge atleast in two complexes. This will continue for every node and will give rise to a common complex, P_i^{Com}, consisting of nodes that are present in all corresponding complexes and inter-connected by an edge. Hence a complex is generated which is arrived at by consensus among the corresponding complexes.

Case 2: For the common set of nodes P_i, complete the edges to obtain a complex P_i^{GO} based on GO similarity.

Next, we superimpose both the complexes obtained from the common set of nodes P_i, i.e., P_i^{Com} and P_i^{GO} to obtain the final complex P_i^{Fin}. The superimposition is done using the logic given in *Table* 5.11.

Table 5.11: Existence of an edge in the final complex

Edge in P_i^{Com}	**Edge in P_i^{GO}**	**Edge in P_i^{Fin}**
Present	Present	Present
Present	Not Present	Present
Not Present	Present	Present
Not Present	Not Present	Not Present

In other words, if the edge is present for a pair of nodes n_i and n_j in both P_i^{Com} and P_i^{GO} , then the edge is retained in the final predicted complex P_i^{Fin}. If an edge is present for a pair of nodes n_i and n_j in P_i^{Com} and absent in P_i^{GO} or vice-versa, then also that edge is retained even though it may be a case of false alarm. The reason to retain this edge is that in the case of P_i^{Com}, the edge is a result of consensus (where un-supervised approach has been applied) and in the case of P_i^{GO}, the edge is obtained through GO similarity (a case of supervised approach). So in both the cases, the probability of the existence of the edge is justified. The final predicted complex P_i^{Fin} is now stored in a file for the purpose of validation.

The process of identifying the node of the highest degree from the next set of complexes and the process of identifying protein complexes through consensus and through the use of GO similarity is repeated for each corresponding complex. The end result of this phase is that there will be interacting protein complexes with overlapping proteins.

Finally, the validation phase evaluates protein complexes predicted by comparing them to a set of gold standard protein complexes and also establishes the biological relevance of the predicted protein complexes using the Co-localization score.

Though the method adopted for protein complex detection is computationally expensive, the primary objective is to find complexes with high precision. For this

reason speed of detection is compromised in favour of accuracy and precision. Our main focus is to make use of multiple sources and means to detect and validate the complexes found since a single information source or criterion has its limitations as discussed in *Section* 5.11.

5.13 Experimental Evaluation

Three different clustering methods for protein complex detection, MCODE [215], MCL [218,332] and CFinder [220] are used as the base clustering algorithms. These methods have been selected as they follow different approaches for clustering protein interaction data and MCODE and CFinder can detect overlapping complexes as noted in *Table* 5.1. Moreover, we have also performed an empirical evaluation of these algorithms along with others, i.e., RNSC [210], RRW [217], AP [321], CMC [230] and ClusterONE [200], and have found that the performance of these algorithms fall in the category of good, average and not so good. Hence these algorithms satisfy the diversity criteria of our ensemble. ClusterONE has been found to be the best performing complex finding algorithms and hence the comparison of our ensemble should be with the best performing baseline method.

The PPI data set for the experiments are the ones used for the empirical evaluation and summarized in *Table* 5.3. The GO-slim file was downloaded from http://www.geneontology.org/ and the Biological Process (BP) hierarchy was selected to calculate the GO-driven similarity of proteins. The GO semantic similarity is calculated based on the information content of GO terms and the semantic similarity proposed by Schlicker et al. [275]. The evaluation of the results was done by the SGD and MIPS gold standard datasets, shown in *Table* 5.2.

5.13.1 Results

The results of the comparison of existing approaches with the proposed ensemble PCDEN are reported next.

5.13.1.1 Comparison of PCDEN with Baseline Clustering Algorithms

We compare the performance of the proposed ensemble PCDEN with the approaches described in *Section* 5.6, namely, MCL [218,332], MCODE [215], RNSC [210], CFinder [220], RRW [217], AP [321], CMC [230] and ClusterONE [200] with optimal parameters settings as reported in *Table* 5.4. The algorithms and the proposed ensemble PCDEN are tested on the yeast PPI datasets, namely Gavin [337], Krogan Core [331] and Krogan Extended [331], the details of these datasets are given in *Table* 5.3.

The number of complexes and the matched number of complexes for each dataset as detected by the algorithms are reported in *Table* 5.12. This table also reports the fraction of protein complexes matched by at least one predicted complex, geometric accuracy and maximum matching ratio using the MIPS gold standard.

Although the number of protein complexes identified by PCDEN is lower than several others, 88 of the 152 predicted complexes matched very well with benchmark complexes. It has also been observed that PCDEN obtains the best scores for the quality measures *Frac* and *Acc* across all the three datasets used for comparison. The maximum matching ratio *MMR* score of 0.390 for the Gavin dataset obtained by PCDEN is also the highest indicating better quality of overlapped protein complexes detected. The best scores are highlighted in bold for easy comparison.

Table 5.12: Comparison of the protein complex detection algorithms and proposed ensemble PCDEN on PPI datasets using the MIPS complex dataset

Algorithm	Gavin					Krogan					Krogan+				
	#c	**#m**	**Frac**	**Acc**	**MMR**	**#c**	**#m**	**Frac**	**Acc**	**MMR**	**#c**	**#m**	**Frac**	**Acc**	**MMR**
MCL	251	82	0.711	0.511	0.338	378	85	0.617	0.450	0.276	489	72	0.448	0.421	0.196
MCODE	137	67	0.585	0.457	0.289	77	39	0.289	0.332	0.122	67	26	0.155	0.279	0.102
RNSC	139	73	0.648	0.495	0.323	88	58	0.400	0.394	0.179	96	58	0.362	0.375	0.153
CFINDER	139	68	0.592	0.500	0.286	118	50	0.358	0.380	0.170	124	38	0.225	0.324	0.112
RRW	238	79	0.684	0.457	0.353	323	71	0.518	0.370	0.251	233	77	0.481	0.365	0.224
AP	249	77	0.686	0.454	0.346	225	57	0.411	0.355	0.177	238	55	0.350	0.343	0.169
CMC	345	78	0.675	0.484	0.342	151	53	0.373	0.378	0.174	428	61	0.382	0.346	0.180
ClusterONE	198	84	0.738	0.513	0.390	526	95	0.692	0.451	0.329	531	94	0.593	0.435	0.294
PCDEN	152	**88**	**0.753**	**0.523**	**0.397**	202	**95**	**0.706**	**0.460**	**0.336**	227	**95**	**0.605**	**0.443**	**0.300**

Abbreviations: #c = no. of complexes, #m = no. of matched complexes

The comparison results are presented in *Figure* 5.8 and it is clear that the PCDEN ensemble outperforms the other approaches in all datasets by achieving the highest score in each of the three categories used for comparison. The results obtained by PCDEN are better than those of the single clustering approaches, proving that the

integration of multiple sources has great power to predict protein complexes in PPI networks.

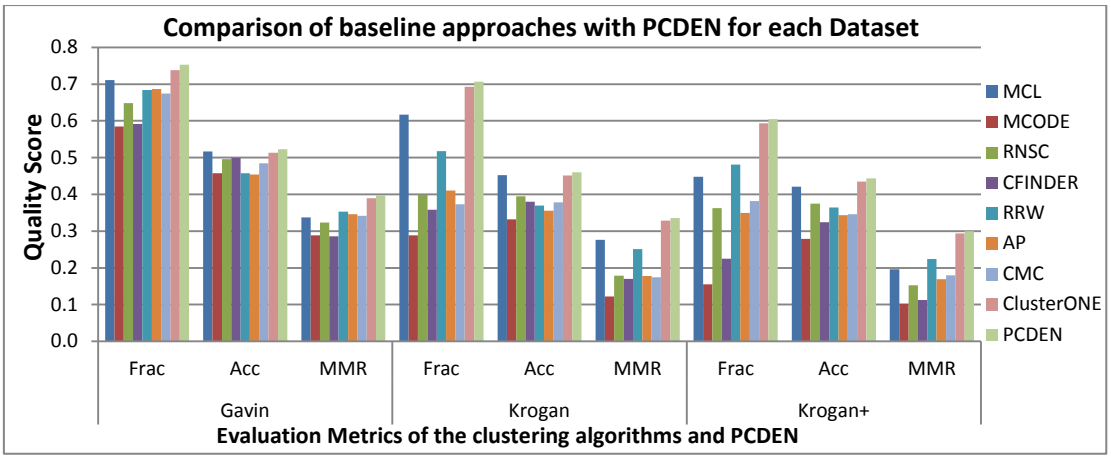


Figure 5.8: Comparing the evaluation metrics of the clustering algorithms and PCDEN for the MIPS dataset

Similarly, *Table* 5.13 shows the scores for the number of complexes detected, number of matched complexes, *Frac*, *Acc* and *MMR* for the three datasets achieved by the algorithms and PCDEN using the SGD gold standard.

Table 5.13: Comparison of the protein complex detection algorithms and proposed ensemble on PPI datasets using the SGD complex dataset

Algorithm	Gavin					Krogan					Krogan+				
	#c	#m	Frac	Acc	MMR	#c	#m	Frac	Acc	MMR	#c	#m	Frac	Acc	MMR
MCL	251	96	0.776	0.710	0.447	375	105	0.662	0.659	0.357	487	92	0.509	0.612	0.258
MCODE	134	74	0.598	0.627	0.367	76	58	0.364	0.474	0.202	69	39	0.216	0.394	0.102
RNSC	136	91	0.736	0.715	0.434	88	78	0.486	0.556	0.265	93	78	0.432	0.552	0.241
CFINDER	132	78	0.630	0.688	0.367	114	69	0.433	0.509	0.248	124	49	0.271	0.484	0.158
RRW	233	97	0.785	0.687	0.480	324	100	0.627	0.578	0.368	235	101	0.559	0.545	0.317
AP	248	99	0.800	0.676	0.446	221	81	0.507	0.525	0.306	235	97	0.537	0.529	0.291
CMC	348	95	0.768	0.661	0.453	156	69	0.433	0.526	0.245	423	78	0.432	0.540	0.257
ClusterONE	198	102	0.817	0.727	0.486	526	112	0.690	0.683	0.426	531	110	0.615	0.647	0.371
PCDEN	152	**105**	**0.833**	**0.742**	**0.495**	202	**115**	**0.704**	**0.697**	**0.435**	227	**115**	**0.627**	**0.660**	**0.379**

Abbreviations: #c = no. of complexes, #m = no. of matched complexes

It is observed that PCDEN achieves the best results in protein complex detection with respect to the three quality measures, for all the three datasets and also identifies the maximum number of matched complexes.

Figure 5.9 shows the comparison of the clustering algorithms and PCDEN for SGD dataset and it is noticed that PCDEN outperforms the other approaches in all datasets.

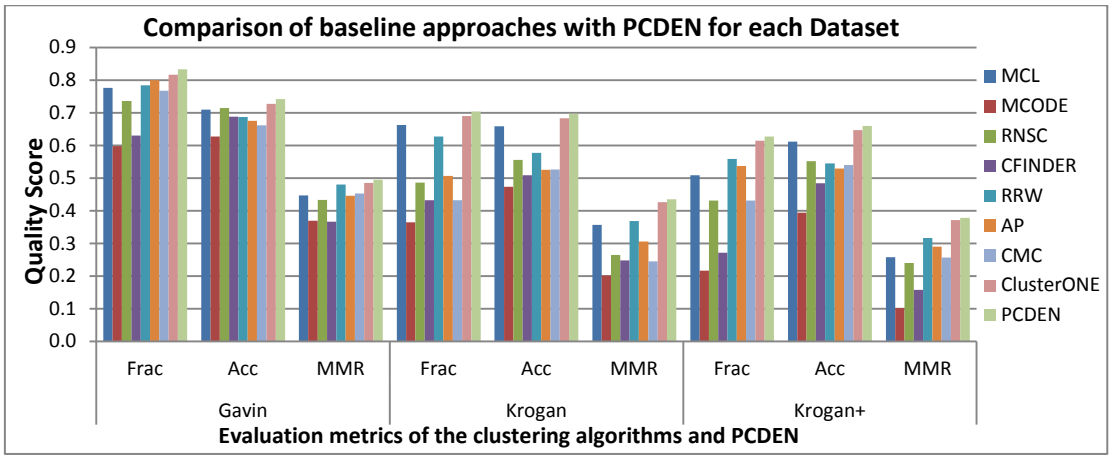


Figure 5.9: Comparing the evaluation metrics of the clustering algorithms and PCDEN for SGD dataset

The co-localization scores of protein complexes of the eight algorithms and PCDEN on the Gavin and Krogan datasets is shown in *Table* 5.14.

Table 5.14: Comparison of Co-localization Score

Algorithm	Co-localization Score	
	Gavin	Krogan
MCL	0.735	0.725
MCODE	0.722	0.671
RNSC	0.618	0.584
CFINDER	0.615	0.59
RRW	0.644	0.611
AP	0.721	0.633
CMC	0.628	0.655
ClusterONE	0.746	0.723
PCDEN	**0.764**	**0.736**

As expected, the highest co-localization score of 0.764 on the Gavin dataset is achieved by PCDEN, followed by ClusterONE with a score of 7.46. Also, the co-localization score on the Krogan dataset of PCDEN is 0.736 which is higher than 0.725, the score of obtained by MCL and 0.723 achieved by ClusterONE is, which is nearly the same as that of. Hence the protein complexes detected by PCDEN ensemble have relatively high quality from the biological view due to the high co-localization score on both the Gavin and Krogan dataset.

The results of comparisons of Co-localization score of PCDEN and the baseline algorithms on the Gavin and Krogan dataset are illustrated in *Figure* 5.10.

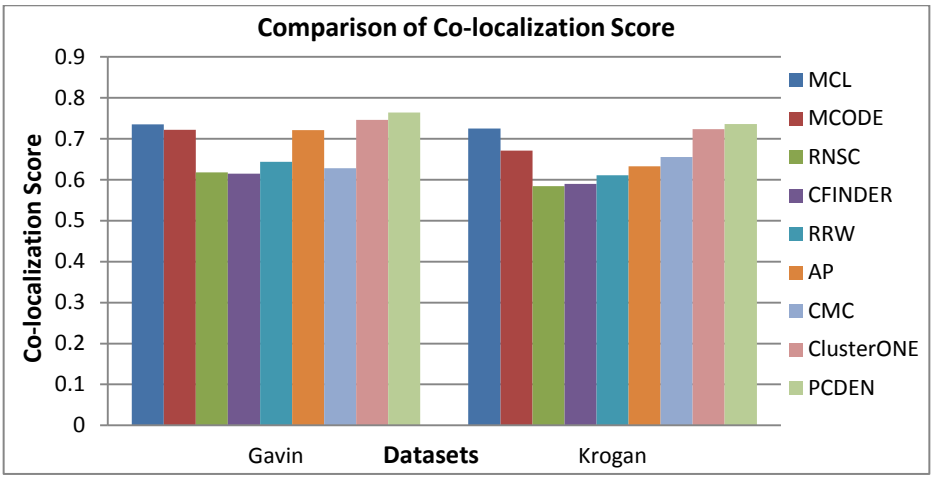


Figure 5.10: Comparison of Co-localization Score

5.13.1.2 Comparison of PCDEN with Ensemble Approaches

In addition to comparison with individual clustering methods, we compare our proposed ensemble PCDEN with ensemble approaches proposed in recent years, namely Ensemble Non-negative Matrix Factorization (NMF) [345], Ensemble Clustering Bayesian Nonnegative Matrix Factorization (EC-BNMF) [346], Bipartite Graph Ensemble (BGENS) [347] and Full Graph Ensemble (FGENS) [347].

The ensemble framework for detecting protein complexes NMF [345] proposed by Greene et al [345] first generates a collection of non-negative matrix factorizations with different number of dimensions. Then a hierarchical meta-clustering algorithm is employed to aggregate these factorizations and produce a disjoint hierarchy of meta-clusters. Finally, the results are transformed into a soft hierarchical clustering of the original dataset.

The next ensemble approach EC-BNMF [346] consists of two phases. In the generation phase, useful information in the form of "features" is extracted from several base clustering results. Each base clustering result is regarded as a "feature" of the original PPI network that indicates two proteins are connected if they have occurred in the same cluster at least once. In the complex detection phase, a Bayesian NMF-based ensemble clustering is employed which utilizes the group information provided by the edges between interacting proteins to detect protein complexes from the PPI network.

BGENS [347] is a bipartite graph that is constructed based on the affiliation of proteins and clusters. The original PPI data is initially clustered into a set of *C* partitions using the base clustering algorithms. Each clustering partition consists of a set of clusters from a certain clustering method. BGENS constructs a bipartite graph between protein nodes in the dataset and the cluster nodes obtained from the base clustering algorithms.

The cluster ensemble, FGENS [347], is constructed from the bipartite graph BGENS, by omitting the interactions between the proteins and also the relationships between the clusters. The authors consider the original protein interactions as the basis of cluster-belonging preference for those proteins that appear in different clusters. The proteins are partitioned into different final clusters by taking into consideration their original relationships with other proteins. The method proposed in BGENS is followed to partition the cluster nodes generated by the base clustering algorithms and a graph of proteins and clusters is constructed. The graph is now partitioned using the spectral clustering method, which partitions the graph into *K* parts with the objective of minimizing the cut.

Analysis of Ensemble Results

To evaluate the performance of the proposed cluster ensemble method PCDEN, the validity measures, Precision, Recall and F-measure are used as discussed in *Section* 2.8.1.5. The results of the ensemble PCDEN are compared with the results of the ensemble approaches discussed earlier. Based on the comparison, the proposed ensemble method outperforms other approaches by getting higher F-measure and Recall rates for the MIPS gold standard, as shown in *Table* 5.15. The F-measure and Recall rates for SGD gold standard are 0.635584 and 0.586210 for PCDEN, which is almost at par with FGENS, which has scored 0.636667 for F-measure and 0.588107 for Recall rate. However, PCDEN finds more matched modules from the PPI network than others and this demonstrates its effectiveness.

NMF and BC-NMF methods use only the PPI network topology whereas PCDEN integrates information from GO similarity sources and so achieves higher scores than these methods. It also shows better results than BGENS and FGENS for the MIPS

gold standard. Its F-measure and Recall rates for SGD gold standard are almost at par with FGENS but are better than BGENS, the reason being that FGENS combines gene expression data and GO similarity information, which is also the case with PCDEN. This indicates that the integration of multiple information sources greatly benefits the protein complex detection.

Table 5.15: Comparison of the ensemble approaches with the proposed ensemble PCDEN

Ensemble Technique	MIPS			SGD		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
BGENS	0.597300	0.506186	0.547982	0.627466	0.531751	0.575658
FGENS	0.643034	0.544943	0.589939	0.655767	0.588107	0.636667
NMF	0.512575	0.432938	0.469402	0.538997	0.455255	0.493598
EC-BNMF	0.553980	0.467910	0.507320	0.603345	0.509605	0.552527
PCDEN	0.654574	0.552875	0.599442	0.674211	0.586210	0.635584

A graphical comparison of the clustering approaches using Precision, Recall and F-measure is given in *Figure* 5.11.

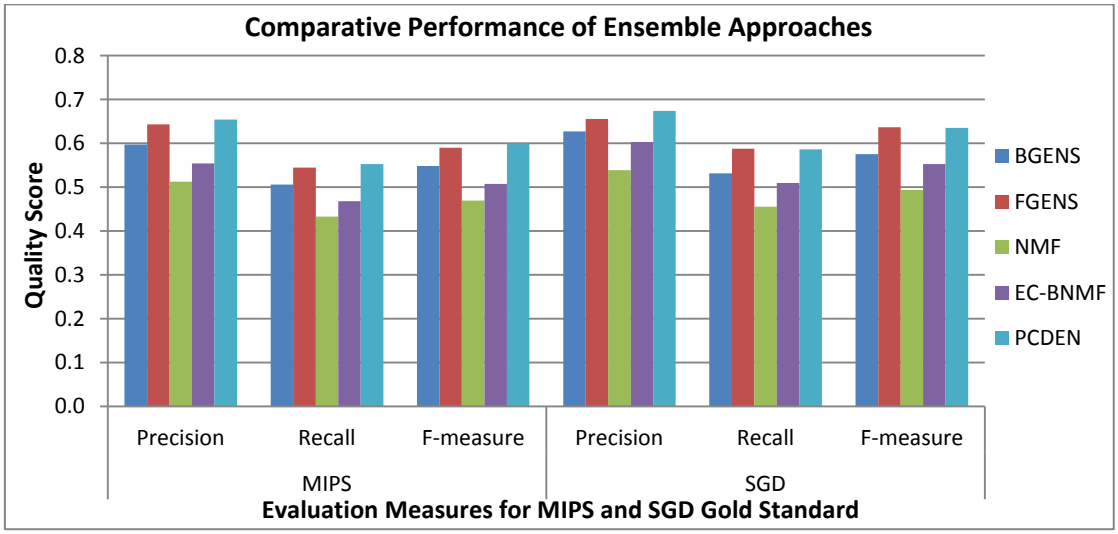


Figure 5.11: Comparison of the ensemble approaches with the proposed ensemble PCDEN

The performance of PCDEN is better than the other cluster ensemble methods, i.e., NMF and BC-NMF, but is at par for one result with FGENS. This indicates the effectiveness of a graph based cluster ensemble method such as FGENS and PCDEN in detecting protein complexes, especially so, when additional information about the proteins is integrated into the network.

5.14 Discussion

Protein complexes are important for understanding principles of cellular organization and function. A better comprehension of their roles allows us to realize how a protein complex disorder can affect the biological processes in which it is involved. Proteins are related to diseases and diseases are usually caused by an erroneous production of some protein complex. Therefore, much work has focussed on the prediction of protein complexes from the PPI networks. However, the PPI datasets from high-throughput techniques are fraught with noise. In response, some research groups propose a number of data integration and affinity scoring schemes and construct weighted networks. In this chapter, eight algorithms have been compared and the results validated against two gold standard datasets independently and the findings are reported. The biological relevance of the protein complexes detected is measured using a co-localization score and GO semantic similarity. It has been observed that the complexes obtained by ClusterONE displayed comparable accuracies when matched against known gold standard complexes. MCL was the closest in performance to ClusterONE, with the exception that MCL produced only non-overlapping clusters - a distinct advantage of ClusterONE.

It has also been observed that traditional clustering approaches for protein complex identification are hampered by a number of factors, such as PPI networks being noisy and incomplete gives rise to high false-positive rate of interactions. The partitioning schemes themselves are dependent on specific initial conditions and parameter settings; the end result is unstable and unsatisfactory cluster partitions which lack biological meaning. Keeping these challenges in mind, a framework has been developed for an ensemble method and established with satisfactory results, for the purpose of identifying protein complexes from interaction datasets.

An ensemble approach is generally able to improve the robustness and stability of the clusterings by combining the output of several algorithms, thus improving the overall prediction accuracy. Such a method can identify larger, denser clusters with improved biological significance. Ensemble-based methods are preferable for protein complex detection also considering that traditional clustering methods search for complexes in

"dense" regions and miss complexes of low densities, that is, small complexes of two or three proteins. Moreover, cluster ensembles can also address the problem of local optima and obtain a globally optimum solution, i.e., a stable clustering solution.

The proposed ensemble called Protein Complex Detection Ensemble (PCDEN) is compared with eight algorithms and the results are validated independently using two gold standard datasets. The biological relevance of the protein complexes detected is measured using a co-localization score and GO semantic similarity. PCDEN is also compared with four ensemble approaches. The PCDEN ensemble outperforms the other clustering approaches in all datasets by achieving the highest score in each of the categories used for comparison. It can be concluded that the complexes obtained by PCDEN achieve accuracies comparable to gold standard complexes. The results prove that the integration of multiple sources increases the precision in detecting protein complexes in PPI networks.