Pertanika J. Sci. & Technol. 29 (2): 1323 - 1345 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

SETS: A Seed-Dense-Expanding Model-Based Topological Structure for the Prediction of Overlapping Protein Complexes

Soheir Noori[1,2]*, Nabeel Al-A'araji[3] and Eman Al-Shamery[1]

[1]*Department of Software, University of Babylon, Babylon, Hillah, Iraq*
[2]*Department of Computer Science, University of Kerbala, Karbala, Iraq*
[3]*Ministry of Higher Education, Baghdad, Iraq*

ABSTRACT

Defining protein complexes by analysing the protein–protein interaction (PPI) networks is a crucial task in understanding the principles of a biological cell. In the last few decades, researchers have proposed numerous methods to explore the topological structure of a PPI network to detect dense protein complexes. In this paper, the overlapping protein complexes with different densities are predicted within an acceptable execution time using seed expanding model and topological structure of the PPI network (SETS). SETS depend on the relation between the seed and its neighbours. The algorithm was compared with six algorithms on six datasets: five for yeast and one for human. The results showed that SETS outperformed other algorithms in terms of F-measure, coverage rate and the number of complexes that have high similarity with real complexes.

Keywords: Common neighbours; density; protein complex; protein–protein interaction network; topological structure

ARTICLE INFO

Article history:
Received: 26 October 2020
Accepted: 18 February 2021
Published: 30 April 2021

DOI: https://doi.org/10.47836/pjst.29.2.35

E-mail addresses:
soheir.noori@uokerbala.edu.iq; soheirn.sw.hdr@student.uobabylon.edu.iq (Soheir Noori)
nhkaghed@itnet.uobabylon.edu.iq (Nabeel Al-A'araji)
emanalshamery@itnet.uobabylon.edu.iq (Eman Al-Shamery)
* Corresponding author

INTRODUCTION

The key to exploring cell behaviour is understanding the mechanism of protein complexes. According to Pizzuti and Rombo (2014), protein complexes are a molecular aggregation of two or more proteins assembled by multiple PPIs. Protein-protein interaction (PPI) plays a central role in many biological functions and its network provides a global view of cellular functionality.

ISSN: 0128-7680
e-ISSN: 2231-8526

Advanced experimental techniques have generated a vast amount of data on PPI (Wang et al., 2017). A PPI network is represented as an undirected graph, where the nodes are proteins and the edges are the interactions between proteins. Many algorithms have been proposed for the analysis of PPI networks to discover the protein complex by defining a dense subgraph in the PPI network. Two of the earliest approaches that have been adopted include the Molecular Complex Detection (MCODE) (Bader & Hogue, 2003) and Markov (Van Dongen, 2000) algorithms. Many other algorithms search for cliques as researchers believe that a complete connected graph represents the protein complex such as Cliques Finder (CFinder) (Adamcsek et al., 2006), Clique Percolation- Distance Restriction (CP-DR) (Wang et al., 2010), Maximal Cliques (Liu et al., 2009) and Local Clique Merging Algorithm (LCMA) (Li et al., 2005). All these algorithms detect cliques and then merge them depending on different criteria to identify the protein complex. Most protein complex detection algorithms such as Graph Fragmentation Algorithm (GFA) (Feng et al., 2010), Dynamic Protein Complexes (DPC) (Li et al., 2014), Detect Module from Seed Protein (DMSP) (Maraziotis et al., 2007) use the topological properties of a graph or mix the PPI network with other information like the gene expression. A subgraph is a protein complex having high functional and structural consistency (Hartwell et al., 1999). Since proteins have multiple functions, they can belong to more than one dense subgraph (Palla et al., 2005; Rives & Galitski, 2003). Therefore, a protein complex can have an overlapping structure as is observed in the ClusterOne (Nepusz et al., 2012) and Near-Clique Mining (NCMine) (Tadaka & Kinoshita, 2016) algorithms.

Most of the existing algorithms can detect only highly dense regions as protein complexes and ignore low density complexes (Wang et al., 2018). Further, most of them cannot detect overlapping protein complexes (Zhao & Lei, 2019). In this study, the overlapping protein complexes with different densities are predicted through the seed expanding approach and the topological structure of the PPI networks (SETS) in an acceptable time period, namely less than five minutes for human and about one minute or less for yeast. A pre-processing step needs to be undertaken first to order the proteins according to their degrees. At the start, SETS choose the first node as a seed. Next, the direct neighbours of the seed will be added to the complex, depending on a common neighbour's technique. Notably, the proteins in a complex cannot be chosen as a next seed but can be added to another complex to generate the overlap between complexes. The predicted complex can be accepted if it is greater than the density threshold. Otherwise, all proteins of the complex will return to construct a new complex from a new seed. Finally, the preliminary complexes will be iteratively expanded according to the closeness score. This algorithm outperforms other algorithms.

METHODS

Preliminary Concepts

Generally, PPI is represented as an undirected and unweighted graph G = (V, E), where V is the nodes representative of proteins and E is the edges which represent the interactions between the proteins. The algorithm uses several measurements.

For each vertex *v*, the degree of *v* is the summation of its connected edges (Equation 1).

$$d(v) = \sum_{i} e_i \qquad [1]$$

The density of the set of vertices S ⊂ V is the number of edges among them divided by the number of possible edges between the set nodes (i.e., how close the set to the clique is, ranging between 0 and 1) (Equation 2).

$$density\ (S) = \frac{2\ \times |E|}{|V| \times (|V| - 1)} \qquad [2]$$

The common neighbours (CN) between two proteins (P_i and P_j) are the number of proteins that indices to both divided by the square root of the product of the nodes' degrees (Equation 3).

$$CN = \frac{N_{p_i} \cap N_{p_j}}{\sqrt{d(p_i) * d(p_j)}} \qquad [3]$$

The Algorithm

The results from some experiments (Goldberg & Roth, 2003; Peng et al., 2017) have shown that the methods that used the information of common neighbours are reliable. SETS is a technique that is employed to detect overlapping protein complexes based on common neighbours. Given an undirected and non-weighted graph, the goal of the algorithm is to identify overlapping protein complexes with different densities. The algorithm accomplishes this through the following steps (Appendix 1: Algorithm 1):

1. Those proteins with a degree higher than 1 are set in ascending order, put in a queue 'Q' and its visited label is set to 'False'.
2. The preliminary complex is built starting from the seed. The nodes that share a specific ratio of common neighbours are added iteratively and its visited label is set to true if it satisfies the predefined threshold of shared neighbours in order to avoid selecting it as a seed in the next iteration. The complex will be accepted as

a preliminary one if its density is greater than a predefined threshold. Otherwise, all the nodes visited labels will be set to 'false' and moved to the next node in the Q with a false visited label. After defining the preliminary complex, no nodes will be deleted from the Q, so that we can get overlapping complexes. The preliminary complex will be accepted as a candidate complex if it contains more than three proteins and was not previously defined from another seed to avoid redundancy (Appendix 1: Algorithm 1, steps 1-14).

3. The candidate complex (CC) will be iteratively expanded according to the closeness score (CS) as obtained through Equation 4 by adding its neighbours 'N_{CC}' that connect with half or more of the proteins of the candidate complex. The expansion will be done in rounds. In each round, the algorithm will search for proteins that satisfy the threshold of closeness score (T_{CS}) to add them to the complex. In the second round, the algorithm will search for proteins, such as those that relate to the updated complex, until no proteins can satisfy the T_{CS}. This step will assist in the identification of complexes with different densities (Appendix 1: Algorithm 1, steps 15-20).

$$CS(C, C\ u) = \frac{N_u\ \cap\ |V_{CC}|}{|V_{CC}|} \qquad [4]$$

4. Redundant complexes will be removed by retaining only one of the exactly matched complexes.

Time Complexity

The execution time has been calculated for each dataset in order to analyse the time complexity of the SETS algorithm. As a pre-processing step, SETS receive a set of ordered nodes Q in increasing order that take $O(n^2)$. SETS process each node in Q that has a visible label set to false and adds its neighbours. This process takes O(n*m) and reduces it to O(N) since not all nodes will be processed. Where n and m represent the number of nodes and their neighbours respectively. The second part of SETS expands each candidate complex c. This takes O(c*m), where m stands for the neighbours of the proteins in a complex. The time complexity of SETS is O(n)+O(c*m). SETS is implemented in python on a 64-bit window system with a 2 GB memory and intel CPU i7 2.40 GHz. Table 1 reports SETS execution time in seconds by using the time package in python.

Table 1
SETS execution time in seconds

Dataset	Time in seconds
Collins	0.255
Gavin	0.104
Krogan	0.191
DIP	0.893
BioGRID	67.479
Human	261.536

Comparison of SETS with Other Algorithms

The performance of the algorithm has been compared to those of six others, namely MCODE (Bader & Hogue, 2003), ClusterONE (Nepusz et al., 2012), NCMine (Tadaka & Kinoshita, 2016), SPICi (Jiang & Singh, 2010), IPCA (Li et al., 2008), and PEWCC (Zaki et al., 2013).

MCODE is one of the seed-extension approaches, which identifies overlapping protein complexes in three steps. Step 1, based on the core clustering coefficient, assigns a weight to every node in the graph. Step 2, extending from seeds that have a high weight, finds a dense region in the weighted graph. Finally, subgraphs that are not dense are filtered.

ClusterOne is another algorithm that detects overlapping protein complexes by starting from the seed protein having the highest degree and then gradually adding and removing proteins to find a cohesive group of proteins that can be overlapped.

On the other hand, NCMine defines a near-complete subgraph as a functional module by using the centrality degree as the weight of the nodes, which then iteratively merges these like cliques to define overlapping modules.

SPICi is a fast heuristic clustering algorithm that selects the seed having the highest weight. The weight represents the degree of the node and then uses a support function to expand the way that the density of the cluster is saved.

IPCA is another algorithm that identifies a dense region in the PPI network as a protein complex. It starts from the seed that has the biggest weight, which is the summation of its weighted edge that represents the number of its common neighbours. IPCA then recursively adds the neighbours of the seed based on two criteria: the shortest path between the seed and the node as well as the probability of its interaction.

Another algorithm that evaluates the protein interactions reliability is PEWCC. It uses the weighted clustering coefficient to detect the protein complex.

All the aforementioned algorithms rely on the topological structure of PPI networks and most of them use the seed-extension approach to detect dense protein complexes.

RESULTS AND DISCUSSION

PPI and Benchmark Datasets

The algorithm has been analysed by concentrating on five PPI networks of *Saccharomyces cerevisiae* (yeast) and one network of *Homo sapiens* (human) (Ma et al., 2017). The latter is a combination of data from two databases: HPRD (Human Protein Reference Database) and BioGRID (version 3.2.109). The PPI datasets of the yeast are Collins and Gavin for ClusterONE (Nepusz et al., 2012), DIP (Xenarios et al., 2002), Krogan (Krogan et al., 2006) and BioGRID from SPICi. Table 2 explains the properties of these datasets. Each dataset contains a different number of proteins having a different number of interactions

that create a variety in the density of network to satisfy the diversity that is required in the PPI networks used with the algorithm. NewMIPS (Mewes et al., 2004) and CYC2008 (Pu et al., 2009) are used as benchmark complexes. All datasets are available online from authors and as Appendix 1 and 2.

Table 2
Number of proteins and intersections, and network density in PPI datasets

Datasets	No. of Proteins	No. of Intersections	Network density
Collins	1622	9074	0.007
Gavin	1855	7669	0.004
Krogan	2675	7084	0.002
DIP	4930	17201	0.001
BioGRID	5361	85866	0.006
Human	15459	144687	0.001

Evaluation Metrics

The quality of a predicted complex was evaluated using various metrics. The definitions of these metrics are introduced as follows:

Recall, Precision, and F-Measure. One of the metrics most commonly used to evaluate any algorithm is recall, precision and F-measure. The overlapping score (OS) is the matching score between the predicted complex (C_1) and benchmark complex (C_2), as expressed in Equation 5. C_1 and C_2 are considered matched if the OS between both is equal to or greater than 0.2 (Altaf-Ul-Amin et al., 2006; Bader & Hogue, 2003).

$$OS(C_1, C_2) = \frac{|C_1 \cap C_2|^2}{|C_1| \times |C_2|} \quad [5]$$

Recall and precision are defined as Equation 6 and 7:

$$Precision = \frac{N(C_1)}{|C_1|} \quad [6]$$

$$Recall = \frac{N(C_2)}{|C_2|} \quad [7]$$

$N(C_1)$ is the number of the predicted complex that satisfies the OS score with at least one complex in the benchmark. $N(C_2)$ is the number of the benchmark complex that satisfies the OS score with at least one predicted complex. The F-measure is a combination of recall and precision (Equation 8).

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad [8]$$

Coverage Rate (CR). CR evaluates the number of proteins that have been covered by the predicted complexes (Brohée & van Helden, 2006; Friedel et al., 2008). CR is defined in Equation 9, where C_2 is the set of benchmark complexes, $maxcom_{ij}$ is the maximal common proteins between the i^{th} benchmark and j^{th} predicted complex divided by N_i protein numbers in i^{th} benchmark complex.

$$CR = \frac{\sum_{i=1}^{|C_2|} Max\{maxcom_{ij}\}}{\sum_{i=1}^{|C_2|} N_i} \quad [9]$$

Exact and High Matching with Real Complexes. The quality of predicted complexes was evaluated by reporting the number of real complexes that exactly match with the predicted complexes and that had an OS score greater than or equal to 0.8, excluding the exact match.

Selection of Parameters

The T_{CN}, DT and T_{CS} parameters had been used in SETS. Proteins that were not in the PPI network had been filtered from the benchmark complexes. Only those complexes with more than two proteins were retained and then filtered again to keep only the complexes that had all their proteins in the PPI network. CN was calculated between the proteins in the same filtered benchmark complex. We also calculated the number of complexes that at least two of their proteins satisfied the CN value (Appendix 2). According to the result of benchmark complexes analyses, the T_{CN} is set for each data. Liu et al. (2010) analysed the protein complexes of CYC2008 (Pu et al., 2009), MI PS (Mewes et al., 2004) and Aloy (Aloy et al., 2004). They found that almost 60% of the complexes had a density equal to or more than 0.5. Therefore, DT was set to 0.5 to define complexes that were dense enough to be the preliminary complexes. T_{CS} was set to at least 0.5 to let only the proteins that had a good closeness to the preliminary complex that was to be added. Table 3 explains the threshold of each dataset.

Quality of Predicted Complexes

The performance of SETS was compared with that of six other approaches using five datasets for yeast and one dataset for human. All datasets were unweighted except SPICi, which used weighted networks. Every parameter in all the algorithms was set to default. In addition, complexes with less than three proteins were ignored. All the algorithms were implemented in the Cytoscape software (Shannon et al., 2003) except SPICi, which was

Table 3
Threshold values for each dataset

Datasets	T_{CN}	DT	T_{CS}
Collins	0.3	0.5	0.6
Gavin	0.3	0.5	0.5
Krogan	0.2	0.5	0.5
DIP	0.1	0.5	0.6
BioGRID	0.2	0.5	0.7
Human data	0.1	0.5	0.7

implemented in its web site. The complex is considered matched if the OS with benchmark complex is greater than or equal to 0.2. SETS have the highest F-measure in all cases and competes with other algorithms in recall and precession (Tables 4 & 5). SETS obtain the highest CR in most cases except in Collins and BioGRID, where it obtained the second-highest CR. Besides a few exceptions where its prediction ranks behind that of the PEWCC, the exact and well-predicted complexes by SETS are the best in most cases (Figure 1). All the results are available in the Appendix 1.

The ProCope software tool (Schlicker et al., 2006) was used to evaluate the biological significance of predicted complexes and the data used in the evaluation process was set to 'default'. The evaluation was based on BP and CC. SETS detects more complexes that are

Table 4
Performance analysis for Gavin data with CYC2008 and NewMIPS

	# complex	Recall	Precession	F-measure	CR
Gavin with CYC2008					
SPICi	91	0.36	0.76	0.491	0.504
ClusterONE	258	0.508	0.419	0.459	0.633
NCMine	621	0.513	0.393	0.445	0.64
PEWCC	656	0.517	0.402	0.453	0.596
IPCA	464	0.53	0.457	0.491	0.626
MCODE	101	0.021	0.05	0.03	0.118
SETS	246	0.475	0.602	**0.531**[1st]	**0.656**[1st]
Gavin with NewMIPS					
SPICi	91	0.372	0.736	0.494	0.248
ClusterONE	258	0.53	0.419	0.468	0.417
NCMine	621	0.549	0.39	0.456	0.422
PEWCC	656	0.552	0.433	0.485	0.392
IPCA	464	0.573	0.47	0.516	0.413
MCODE	101	0.021	0.059	0.031	0.045
SETS	246	0.524	0.607	**0.563**[1st]	**0.43**[1st]

significant in BioGRID and human datasets (Figure 2) and ranks second with regard to the rest of the datasets, competing with SPICi, IPCA and ClusterONE algorithms (Appendix 1).

SETS predict overlapping complexes as explained in Table 7 that reports some of these complexes that have high OS scores with benchmark complexes and share some of their

Table 5
Performance analysis for Krogan data with CYC2008 and NewMIPS

	# complex	Recall	Precession	F-measure	CR
Krogan with CYC2008					
SPICi	131	0.458	0.641	0.534	0.583
ClusterONE	240	0.492	0.512	0.502	0.598
NCMine	578	0.458	0.433	0.445	0.593
PEWCC	708	0.525	0.496	0.51	0.593
IPCA	472	0.517	0.595	0.553	0.599
MCODE	60	0.03	0.117	0.047	0.111
SETS	220	0.479	0.764	**0.589**[1st]	**0.68**[1st]
Krogan with NewMIPS					
SPICi	131	0.479	0.618	0.54	0.352
ClusterONE	240	0.442	0.458	0.45	0.323
NCMine	578	0.479	0.427	0.452	0.362
PEWCC	708	0.534	0.476	0.503	0.368
IPCA	472	0.515	0.574	0.543	0.36
MCODE	60	0.021	0.1	0.035	0.038
SETS	220	0.485	0.732	**0.583**[1st]	**0.391**[1st]

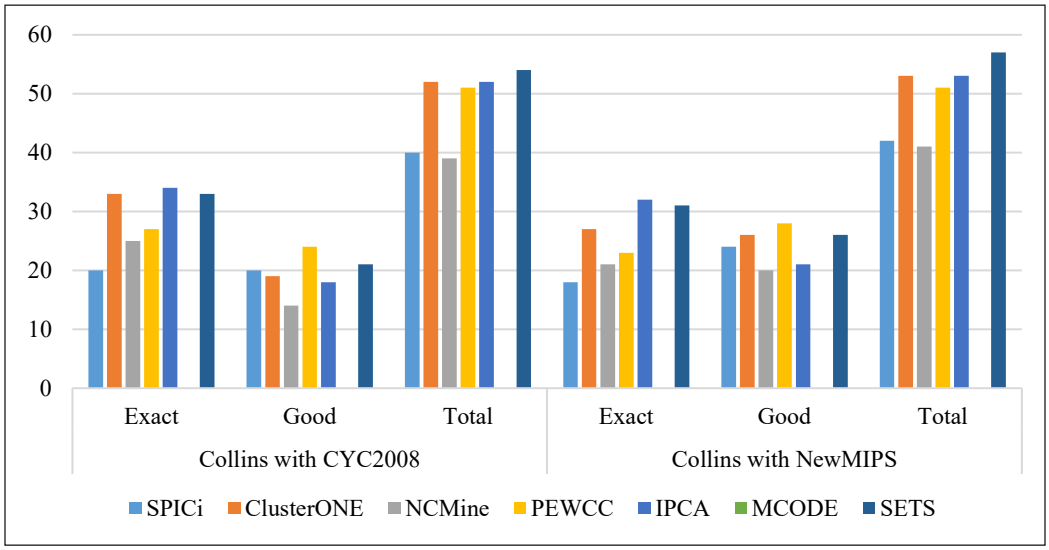

Figure 1. Number of exact and well-predicted complexes in Collins dataset

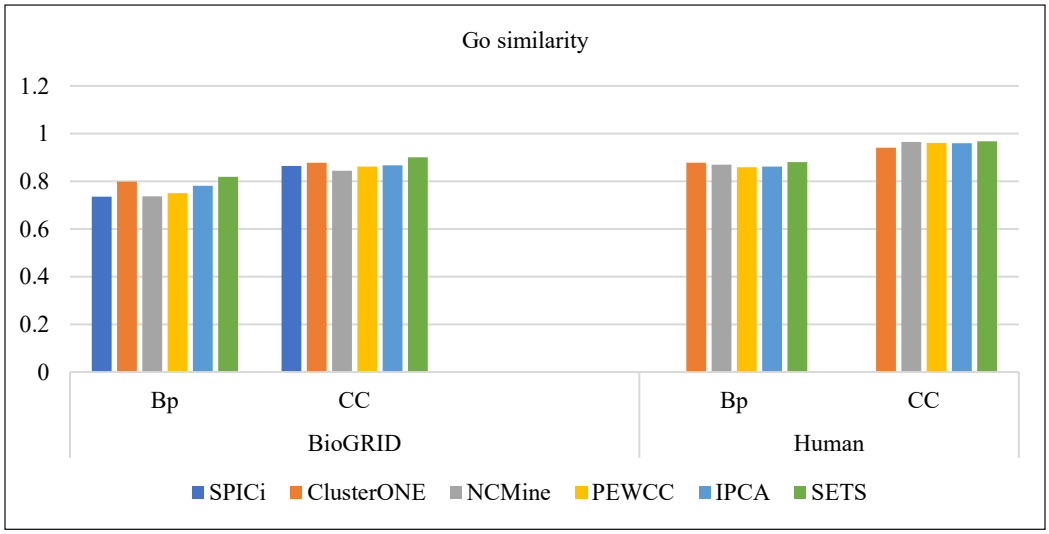

Figure 2. Biological significance of predicted complexes in BioGRID and Human

Table 6
DIP with Newmips reports low density (D.) complexes with a high OS

Real complex	Predicted complex	Inter.	D.	OS
YKR068C YBR254C YDR472W YDR108W YMR218C YDR246W YML077W YOR115C YGR166W YDR407C Length = 10	YLR342W YKR068C YBR254C YDR472W YDR108W YMR218C YDR246W YML077W YOR115C YGR166W YDR407C Length = 11	10	0.36	0.91
YDL005C YNL236W YPR070W YOL135C YNR010W YDR308C YBR193C YBR253W YNL025C YPR168W YMR112C YGL025C YCR081W YGL151W YOL051W YOR174W YLR071C YGR104C YGL127C YHR041C YHR058C YPL042C YER022W YBL093C YDR443C Length = 25	YDL005C YNL236W YPR070W YOL135C YNR010W YDR308C YBR193C YBR253W YMR112C YGL025C YCR081W YGL151W YOL051W YOR174W YLR071C YGR104C YHR041C YHR058C YER022W YBL093C YDR443C Length = 21	21	0.44	0.84
Q0080 YDR322C-A YPL271W YPR020W YBL099W YML081C-A YDR377W YOL077W-A YDL004W YKL016C YDR298C YGR008C YBR039W YLR295C Q0085 Q0130 YDL130W-A YDL181W YPL078C YJR121W Length = 20	Q0080 Q0130 YDR322C-A YBL099W YNL315C YPL271W YBR039W YPL078C YLR295C YML081C-A YPR020W YDR377W YDR298C YDL004W YKL016C Q0085 YDL181W YJR121W Length = 18	17	0.29	0.8

Table 6 *(continue)*

Real complex			Predicted complex			Inter.	D.	OS
YIL084C	YMR075W	YOL004W	YIL084C	YMR075W	YOL004W	10	0.36	0.7
YMR128W	YPL181W	YDL076C	YPL181W	YIL035C	YDL076C YPR023C			
YPR023C	YMR263W	YNL330C	YNL330C	YMR263W	YLR103C			
YPL139C	YNL097C		YPL139C	YNL097C	YBR095C			
Length = 11			Length = 13					
YKR068C	YBR254C	YDR472W	YLR342W	YKR068C	YBR254C	7	0.36	0.64
YDR108W	YDR246W	YML077W	YDR472W	YDR108W	YMR218C			
YOR115C			YDR246W	YML077W	YOR115C			
Length = 7			YGR166W	YDR407C				
			Length = 11					

Note. Inter. is the interaction between predicted and real complexes

Table 7
Predicted overlapping complexes with high OS score from Collins using NewMIPS

Predicted complex	Real complex	OS	Predicted complex	Real complex	OS	Overlapping proteins
YFL039C	**YFL039C**	1	**YFL039C**	**YFL039C**	0.92	**YNL107W**
YJL081C	**YJL081C**		YDR334W	YDR334W		**YFL039C**
YOR244W	YOR244W		**YJL081C**	**YJL081C**		**YJL081C**
YHR090C	YHR090C		YML041C	YLR385C		**YGR002C**
YNL107W	**YNL107W**		**YNL107W**	YML041C		
YNL136W	YNL136W		YBR231C	**YNL107W**		
YEL018W	YEL018W		YLR085C	YBR231C		
YHR099W	YHR099W		YDR485C	YDR485C		
YPR023C	YPR023C		YDR190C	YLR085C		
YFL024C	YFL024C		YAL011W	YDR190C		
YJR082C	YJR082C		YPL235W	YAL011W		
YDR359C	YDR359C		**YGR002C**	YPL235W		
YGR002C	**YGR002C**			**YGR002C**		
YKL144C	YHR143W-A	0.94	YHR143W-A	YHR143W-A	0.93	**YOR210W**
YOR210W	YKL144C		YOR341W	YOR341W		**YPR110C**
YOR116C	**YOR210W**		**YOR210W**	**YOR210W**		**YNL113W**
YPR190C	YOR116C		**YPR110C**	**YPR110C**		**YBR154C**
YPR110C	YPR190C		**YNL113W**	**YNL113W**		**YPR187W**
YNL113W	**YPR110C**		YOR340C	YOR340C		**YOR224C**
YDL150W	**YNL113W**		YJR063W	YJR063W		
YBR154C	YDL150W		YOR151C	**YBR154C**		
YPR187W	**YBR154C**		**YBR154C**	YNL248C		
YKR025W	**YPR187W**		YNL248C	YDR156W		
YDR045C	YKR025W		YDR156W	**YPR187W**		
YNR003C	YDR045C		**YPR187W**	YPR010C		
YNL151C	YNR003C		YPR010C	**YOR224C**		
YOR207C	YNL151C		**YOR224C**	YJL148W		
YJL011C	YOR207C		YJL148W			
YOR224C	YJL011C					
	YOR224C					

proteins. The complexes predicted by SETS are of various densities and not restricted to dense ones as is the case with other algorithms that use the topological structure of PPI. SETS can achieve higher F-measure with different densities of PPI network in contrast with other algorithms whose F-measure decreases when the PPI network density does. Table 6 reports some of the low-density complexes that have high OS scores with benchmark complexes.

ECC vs. CN with SETS

The CN calculation in Algorithm 1 (Appendix 1) is replaced with an edge clustering coefficient (ECC) to compare the F-measure in both cases (Figure 3). SETS with ECC is high only with Collins, which has the highest network density. It, on the other hand, achieved a lower F-measure than SETS with CN in other datasets that have different densities. Radicchi et al. (2004) realized that ECC might not suite for PPI networks as it was disassortative. This was proven with SETS that performed better using CN with different network densities.

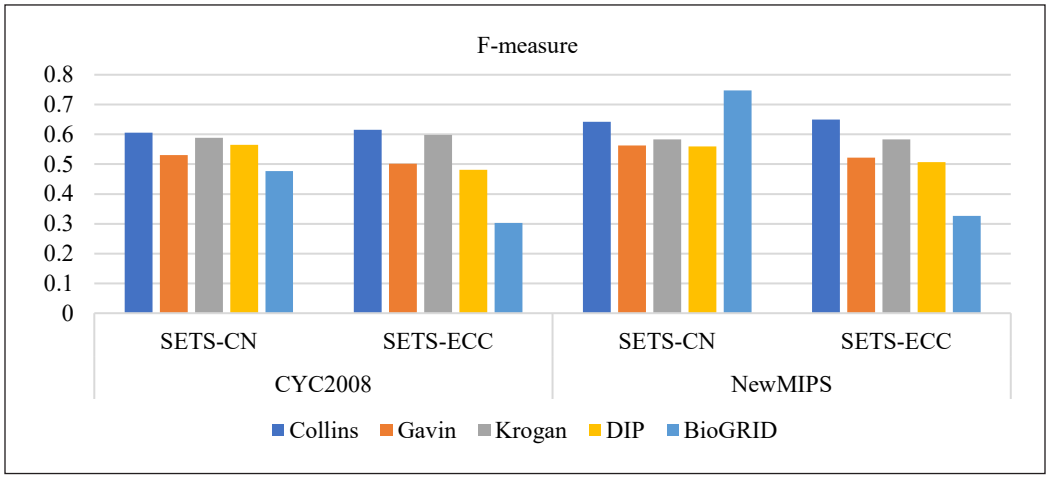

Figure 3. F-measure of ECC vs. CN with SETS

CONCLUSION

In this paper, the seed-expansion model has been proposed based on the topological structure of PPI networks to predict overlapping protein complexes with various densities. The main idea behind this algorithm is (i) choosing the first node in Q that is not visited before as a seed, (ii) adding the seed's neighbour that shares a specific percentage of common neighbours and accepting the complex if its density is more than or equal to the density threshold (DT) and (iii) expanding each complex by adding the proteins that are close to the complex's proteins. SETS could achieve high accuracy in all datasets that have

different densities with good biological significance of predicted complexes compared to other methods. SETS can be further improved by using biological information as gene expression or gene ontology.

ACKNOWLEDGEMENT

We would like to thank the University of Kerbala and the University of Babylon, Iraq for partially sponsoring this research.

REFERENCES

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., & Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, *22*(8), 1021-1023. https://doi.org/10.1093/bioinformatics/btl039

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.-C., Bork, P., Superti-Furga, G., & Serrano, L. (2004). Structure-based assembly of protein complexes in yeast. *Science*, *303*(5666), 2026-2029. https://doi.org/10.1126/science.1092645

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., & Kanaya, S. (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, *7*, Article 207. https://doi.org/10.1186/1471-2105-7-207

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*, Article 2. https://doi.org/10.1186/1471-2105-4-2

Brohée, S., & van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, *7*, Article 488. https://doi.org/10.1186/1471-2105-7-488

Feng, J., Jiang, R., & Jiang, T. (2010). A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(3), 621-634. https://doi.org/10.1109/TCBB.2010.78

Friedel, C. C., Krumsiek, J., & Zimmer, R. (2008). Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. In M. Vingron & L. Wong (Eds.), *Lecture notes in computer science: Research in computational molecular biology* (Vol. 4955, pp. 3-16). Springer. https://doi.org/10.1007/978-3-540-78839-3_2.

Goldberg, D. S., & Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences, USA*, *100*(8), 4372-4376. https://doi.org/10.1073/pnas.0735871100

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, *402*(6761), C47-C52. https://doi.org/10.1038/35011540

Jiang, P., & Singh, M. (2010). SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics*, *26*(8), 1105-1111. https://doi.org/10.1093/bioinformatics/btq078

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., & Tikuisis, A. P. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, *440*(7084), 637-643. https://doi.org/10.1038/nature04670

Li, M., Chen, J. E., Wang, J. X., Hu, B., & Chen, G. (2008). Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics*, *9*, Article 398. https://doi.org/10.1186/1471-2105-9-398

Li, M., Chen, W., Wang, J., Wu, F. X., & Pan, Y. (2014). Identifying dynamic protein complexes based on gene expression profiles and PPI networks. *BioMed Research International*, *2014*, Article 375262. https://doi.org/10.1155/2014/375262

Li, X. L., Foo, C. S., Tan, S. H., & Ng, S. K. (2005). Interaction graph mining for protein complexes using local clique merging. *Genome Informatics*, *16*(2), 260-269. https://doi.org/10.11234/gi1990.16.2_260

Liu, G., Wong, L., & Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics*, *25*(15), 1891-1897. https://doi.org/10.1093/bioinformatics/btp311

Liu, G., Yong, C. H., Wong, L., & Chua, H. N. (2010, December 18-21). *Decomposing PPI networks for complex discovery* [Paper presentation]. 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Hong Kong, China. https://doi.org/10.1109/BIBM.2010.5706577.

Ma, C. Y., Chen, Y. P. P., Berger, B., & Liao, C. S. (2017). Identification of protein complexes by integrating multiple alignment of protein interaction networks. *Bioinformatics*, *33*(11), 1681-1688. https://doi.org/10.1093/bioinformatics/btx043

Maraziotis, I. A., Dimitrakopoulou, K., & Bezerianos, A. (2007). Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinformatics*, *8*, Article 408. https://doi.org/10.1186/1471-2105-8-408

Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., & Stümpflen, V. (2004). MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, *32*(suppl_1), D41-D44. https://doi.org/10.1093/nar/gkh092

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, *9*(5), 471-472. https://doi.org/10.1038/nmeth.1938

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814-818. https://doi.org/10.1038/nature03607

Peng, X., Wang, J., Peng, W., Wu, F. X., & Pan, Y. (2017). Protein–protein interactions: Detection, reliability assessment and applications. *Briefings in Bioinformatics*, *18*(5), 798-819. https://doi.org/10.1093/bib/bbw066

Pizzuti, C., & Rombo, S. E. (2014). Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, *30*(10), 1343-1352. https://doi.org/10.1093/bioinformatics/btu034

Pu, S., Wong, J., Turner, B., Cho, E., & Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, *37*(3), 825-831. https://doi.org/10.1093/nar/gkn1005

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences, USA*, *101*(9), 2658-2663. https://doi.org/10.1073/pnas.0400054101

Rives, A. W., & Galitski, T. (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences, USA*, *100*(3), 1128-1133. https://doi.org/10.1073/pnas.0237338100

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, *7*, Article 302. https://doi.org/10.1186/1471-2105-7-302

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498-2504. https://doi.org/10.1101/gr.1239303

Tadaka, S., & Kinoshita, K. (2016). NCMine: Core-peripheral based functional module detection using near-clique mining. *Bioinformatics*, *32*(22), 3454-3460. https://doi.org/10.1093/bioinformatics/btw488

Van Dongen, S. M. (2000). *Graph clustering by flow simulation* [Doctoral dissertation, Utrecht University]. Utrecht University Publication. https://dspace.library.uu.nl/bitstream/handle/1874/848/full.pdf?sequence=1&isAllowed=y.

Wang, J., Liu, B., Li, M., & Pan, Y. (2010). Identifying protein complexes from interaction networks based on clique percolation and distance restriction. *BMC Genomics*, *11*, Article S10. https://doi.org/10.1186/1471-2164-11-S2-S10

Wang, R., Liu, G., Wang, C., Su, L., & Sun, L. (2018). Predicting overlapping protein complexes based on core-attachment and a local modularity structure. *BMC Bioinformatics*, *19*, Article 305. https://doi.org/10.1186/s12859-018-2309-9

Wang, Y., You, Z., Li, X., Chen, X., Jiang, T., & Zhang, J. (2017). PCVMZM: Using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein–protein interactions from protein sequences. *International Journal of Molecular Sciences*, *18*(5), Article 1029. https://doi.org/10.3390/ijms18051029

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, *30*(1), 303-305. https://doi.org/10.1093/nar/30.1.303

Zaki, N., Efimov, D., & Berengueres, J. (2013). Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinformatics*, *14*, Article 163. https://doi.org/10.1186/1471-2105-14-163

Zhao, J., & Lei, X. (2019). Detecting overlapping protein complexes in weighted PPI network based on overlay network chain in quotient space. *BMC Bioinformatics*, *20*, Article 682. https://doi.org/10.1186/s12859-019-3256-9

APPENDIX 1

Algorithm 1

Inputs: Q that contains ordered proteins
Output: The sets of predicted protein complexes (COMPLEXES).

```
1.  For each protein in Q
2.    IF visited_label == False
3.      Add protein to complex set (COMP)
4.      Set visited_label of protein to True
5.      For each neighbour of protein
6.          Find the common neighbours (CN) between protein and neighbours
7.          IF CN >= T_CN
8.              Add neighbour to COMP
9.              Set visited_label of neighbour to True
10.     IF density(COMP) >= DT and COMP IS not in COMPLEXES
11.       Add COMP to COMPLEXES
12.     ELSE
13.       For each protein in COMP
14.         Set visited_label of protein to False
15. For each complex in COMPLEXES
16.   For each round
17.       Find neighbours N_CC of complex's proteins
18.       For each protein in N_CC
19.           IF CS(CC, protein) >= T_CS
20.               Add protein to complex
21. Return complexes
```

Collins

	R	P	F	CR	# Complexes	# matched complexes	Max	Exact	Good	Total
CYC2008										
SPICi	0.419	0.736	0.534	0.69	106	78	70	20	20	40
ClusterONE	0.559	0.547	0.553	**0.797**	203	111	103	33	19	52
NCMine	0.517	0.475	0.495	0.763	377	179	71	25	14	39
PEWCC	0.53	0.521	0.525	0.738	426	222	89	27	24	51
IPCA	0.542	0.64	0.587	0.751	342	219	68	34	18	52
MCODE	0.051	0.107	0.069	0.121	103	11	79	0	0	0
SETS	0.521	0.725	**0.606**	0.767	218	158	70	33	21	**54**
NewMips										
SPICi	0.473	0.726	0.573	0.443	106	77	70	18	24	42
ClusterONE	0.588	0.542	0.564	**0.519**	203	110	103	27	26	53
NCMine	0.537	0.501	0.518	0.493	377	189	71	21	20	41
PEWCC	0.546	0.533	0.539	0.479	426	227	89	23	28	51
IPCA	0.567	0.705	0.628	0.486	342	241	68	32	21	53
MCODE	0.03	0.087	0.045	0.055	103	9	79	0	0	0
SETS	0.555	0.761	**0.642**	0.488	218	166	70	31	26	**57**

Note. R: Recall, P: Precession, F: F-measure, CR: Coverage Rate, Max: Maximum size of the complex.

Gavin

	R	P	F	CR	# Complexes	# matched complexes	Max	Exact	Good	Total
CYC2008										
SPICi	0.36	0.76	0.491	0.504	91	70	13	14	10	24
ClusterONE	0.508	0.419	0.459	0.633	258	108	40	11	22	33
NCMine	0.513	0.393	0.445	0.64	621	244	43	9	14	23
PEWCC	0.517	0.402	0.453	0.596	656	264	36	11	20	31
IPCA	0.53	0.457	0.491	0.626	464	212	37	15	19	34
MCODE	0.021	0.05	0.03	0.118	101	5	137	0	0	0
SETS	0.475	0.602	**0.531**	**0.656**	246	148	37	12	25	**37**
NewMips										
SPICi	0.372	0.736	0.494	0.248	91	67	13	11	15	26
ClusterONE	0.53	0.419	0.468	0.417	258	108	40	11	19	30
NCMine	0.549	0.39	0.456	0.422	621	242	43	10	16	26
PEWCC	0.552	0.433	0.485	0.392	656	284	36	13	21	34
IPCA	0.573	0.47	0.516	0.413	464	218	37	17	25	42
MCODE	0.021	0.059	0.031	0.045	101	6	137	0	0	0
SETS	0.524	0.607	**0.563**	**0.43**	246	159	37	13	31	**44**

Note. R: Recall, P: Precession, F: F-measure, CR: Coverage Rate, Max: Maximum size of the complex.

Krogan

	R	P	F	CR	# Complexes	# matched complexes	Max	Exact	Good	Total
CYC2008										
SPICi	0.458	0.641	0.534	0.583	131	84	20	17	15	32
ClusterONE	0.492	0.512	0.502	0.598	240	123	23	12	15	27
NCMine	0.458	0.433	0.445	0.593	578	250	25	5	17	22
PEWCC	0.525	0.496	0.51	0.593	708	351	31	15	24	**39**
IPCA	0.517	0.595	0.553	0.599	472	281	22	19	15	34
MCODE	0.03	0.117	0.047	0.111	60	7	73	0	0	0
SETS	0.479	0.764	**0.589**	**0.68**	220	168	62	19	20	**39**
NewMips										
SPICi	0.479	0.618	0.54	0.352	131	81	20	16	17	33
ClusterONE	0.442	0.458	0.45	0.323	240	110	23	9	12	21
NCMine	0.479	0.427	0.452	0.362	578	247	25	7	13	20
PEWCC	0.534	0.476	0.503	0.368	708	337	31	10	22	32
IPCA	0.515	0.574	0.543	0.36	472	271	22	13	18	31
MCODE	0.021	0.1	0.035	0.038	60	6	73	0	0	0
SETS	0.485	0.732	**0.583**	**0.391**	220	161	62	14	21	**35**

Note. R: Recall, P: Precession, F: F-measure, CR: Coverage Rate, Max: Maximum size of the complex.

DIP

	R	P	F	CR	# Complexes	# matched complexes	Max	Exact	Good	Total
CYC2008										
SPICi	0.555	0.507	0.53	0.541	219	111	22	13	8	21
ClusterONE	0.436	0.336	0.38	0.466	342	115	23	7	7	14
NCMine	0.542	0.291	0.378	0.497	1074	312	28	8	10	18
PEWCC	0.678	0.317	0.432	0.582	1544	490	42	22	18	**40**
IPCA	0.589	0.318	0.413	0.516	826	263	32	17	10	27
MCODE	0.008	0.04	0.014	0.116	50	2	180	0	0	0
SETS	0.653	0.498	**0.565**	**0.593**	540	269	41	18	14	32
NewMips										
SPICi	0.573	0.479	0.522	0.334	219	105	22	11	8	19
ClusterONE	0.412	0.304	0.35	0.265	342	104	23	5	5	10
NCMine	0.546	0.287	0.376	0.32	1047	308	28	5	11	16
PEWCC	0.683	0.318	0.434	**0.39**	1544	491	42	16	17	**33**
IPCA	0.579	0.311	0.405	0.323	826	257	32	18	8	26
MCODE	0.006	0.04	0.011	0.046	50	2	180	0	0	0
SETS	0.64	0.496	**0.559**	**0.39**	540	268	41	16	17	**33**

Note. R: Recall, P: Precession, F: F-measure, CR: Coverage Rate, Max: Maximum size of the complex.

BioGRID

	R	P	F	CR	# Complexes	# matched complexes	Max	Exact	Good	Total
CYC2008										
SPICi	0.432	0.186	0.26	0.613	440	82	141	4	2	6
ClusterONE	0.487	0.265	0.343	0.697	476	126	83	1	7	8
NCMine	0.737	0.123	0.211	0.807	3671	451	95	4	10	14
PEWCC	0.873	0.196	0.32	**0.872**	4048	792	750	11	21	**33**
IPCA	0.576	0.14	0.226	0.758	2718	381	80	2	14	16
MCODE	0.008	0.036	0.014	0.073	56	2	192	0	0	0
SETS	0.644	0.379	**0.477**	0.816	633	240	93	6	23	29
NewMips										
SPICi	0.436	0.18	0.254	0.442	440	79	141	2	5	7
ClusterONE	0.488	0.25	0.331	0.496	476	119	83	1	6	7
NCMine	0.695	0.13	0.219	0.551	3671	478	95	5	8	13
PEWCC	0.826	0.21	0.335	**0.627**	4048	850	750	10	20	**30**
IPCA	0.591	0.138	0.223	0.538	2718	374	80	2	13	15
MCODE	0.006	0.036	0.01	0.029	56	2	192	0	0	0
SETS	0.622	0.382	**0.474**	0.561	633	242	93	6	21	27

Note. R: Recall, P: Precession, F: F-measure, CR: Coverage Rate, Max: Maximum size of the complex.

Human

	R	P	F	CR	# Complexes	# matched complexes	Max	Exact	Good	Total
SPICi	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
ClusterONE	0.223	0.235	0.229	0.33	1037	252	96	11	9	20
NCMine	0.552	0.221	0.315	0.459	7776	1716	111	7	12	19
PEWCC	0.68	0.276	0.393	**0.559**	9036	2495	394	21	34	**55**
IPCA	0.463	0.266	0.338	0.455	6533	1736	93	5	9	14
MCODE	0.001	0.014	0.002	0.04	74	1	377	0	0	0
SETS	0.498	0.405	**0.447**	0.484	2026	822	223	18	26	44

Note. R: Recall, P: Precession, F: F-measure, CR: Coverage Rate, Max: Maximum size of the complex.

Biological significance

	Collins		Gavin		Krogan		DIP		BioGRID		Human	
	Bp	CC	Bp	CC	Bp	CC	Bp	CC	Bp	CC	Bp	CC
SPICi	0.953	0.976	0.954	0.977	0.957	0.982	0.947	0.962	0.736	0.865	NA	NA
ClusterONE	0.902	0.958	0.782	0.893	0.849	0.91	0.809	0.876	0.799	0.878	0.879	0.941
NCMine	0.924	0.961	0.858	0.917	0.838	0.917	0.788	0.926	0.737	0.845	0.871	0.966
PEWCC	0.942	0.968	0.88	0.933	0.861	0.928	0.804	0.935	0.751	0.862	0.86	0.961
IPCA	0.967	0.981	0.882	0.937	0.883	0.938	0.802	0.946	0.782	0.868	0.862	0.96
SETS	0.962	0.976	0.897	0.944	0.897	0.947	0.833	0.951	0.82	0.901	0.881	0.968

ECC vs. CN in SETS

	CYC2008		NewMIPS		Human - CORUN	
	SETS-CN F-measure	SETS-ECC F-measure	SETS-CN F-measure	SETS-ECC F-measure	SETS-CN F-measure	SETS-ECC F-measure
Collins	0.606	**0.615**	0.642	**0.65**	**0.447**	0.333
Gavin	**0.53**	0.502	**0.563**	0.522		
Krogan	0.589	**0.598**	**0.583**	**0.583**		
DIP	**0.565**	0.481	**0.559**	0.507		
BioGRID	**0.477**	0.303	**0.474**	0.327		

APPENDIX 2

Analysis of Benchmark Complexes

The benchmark dataset is analysed using PPI networks. Tables contain the number of proteins in each PPI as well as the number of proteins that are in benchmark complexes but are not in PPIs. The number of complexes in the benchmark dataset is reported, the benchmark complexes from proteins that are not in PPI are filtered out and only the complexes that have a length of more than two proteins are retained. The benchmark complexes are filtered again and only those complexes that have all its proteins in PPI are retained. The CN is calculated between the proteins of the same complex for different thresholds. The number of complexes where at least two of its proteins are satisfied at the threshold is reported and according to the number of complexes that satisfied T_{CN} to the number of complexes from the second filter, almost 25% of complexes from the second filter, the threshold T_{CN} is set to each PPI. The F-measure of each dataset with a different threshold proved the accuracy of the selected threshold.

Collins

	# proteins In Collins	# proteins in benchmark but not in PPI	# benchmark complexes	First filter	Second filter
CYC2008	1662	382	236	145	102
NewMIPS	1662	695	328	221	106

T_{CN}	CYC2008	NewMIPS
0.1	1	2
0.2	14	10
0.3	24	24
0.4	36	38
0.5	47	48

T_{CN}	F-measure (CYC2008)	F-measure (NewMIPS)
0.1	0.602	0.638
0.2	0.606	0.638
0.3	0.606	0.642
0.4	0.588	0.623
0.5	0.571	0.604

Gavin

	# proteins In Collins	# proteins in benchmark but not in PPI	# benchmark complexes	First filter	Second filter
CYC2008	1855	439	236	143	86
NewMIPS	1855	724	328	218	90

T_{CN}	CYC2008	NewMIPS
0.1	5	6
0.2	17	17
0.3	25	29
0.4	33	36
0.5	47	51

TCN	F-measure (CYC2008)	F-measure (NewMIPS)
0.1	0.483	0.51
0.2	0.499	0.526
0.3	0.53	0.563
0.4	0.537	0.566
0.5	0.544	0.565

Krogan

	# proteins In Krogan	# proteins in benchmark but not in PPI	# benchmark complexes	First filter	Second filter
CYC2008	2675	389	236	169	119
NewMIPS	2675	604	328	249	123

T_{CN}	CYC2008	NewMIPS
0.1	12	17
0.2	40	43
0.3	70	68
0.4	88	90
0.5	99	103

TCN	F-measure (CYC2008)	F-measure (NewMIPS)
0.1	0.6	0.575
0.2	0.589	0.583
0.3	0.531	0.548
0.4	0.468	0.494
0.5	0.421	0.442

DIP

	# proteins In DIP	# proteins in benchmark but not in PPI	# benchmark complexes	First filter	Second filter
CYC2008	4930	138	236	226	191
NewMIPS	4930	194	328	313	231

T_{CN}	CYC2008	NewMIPS
0.1	61	100
0.2	133	173
0.3	158	201
0.4	168	215
0.5	170	218

TCN	F-measure (CYC2008)	F-measure (NewMIPS)
0.1	0.565	0.559
0.2	0.516	0.518
0.3	0.464	0.507
0.4	0.35	0.407
0.5	0.194	0.407

BioGRID

	# proteins In BioGRID	# proteins in benchmark but not in PPI	# benchmark complexes	First filter	Second filter
CYC2008	5361	6	236	236	231
NewMIPS	5361	31	328	322	301

T_{CN}	CYC2008	NewMIPS
0.1	75	165
0.2	159	242
0.3	199	277
0.4	217	292
0.5	222	296

TCN	F-measure (CYC2008)	F-measure (NewMIPS)
0.1	0.406	0.416
0.2	0.477	0.474
0.3	0.45	0.485
0.4	0.358	0.408
0.5	0.247	0.297

Human

	# proteins In Human dataset	# proteins in benchmark but not in PPI	# benchmark complexes	First filter	Second filter
CORUM	15459	157	2351	2340	2196

T_{CN}	CORUM
0.1	1483
0.2	1948
0.3	2071
0.4	2123
0.5	2145

TCN	F-measure (CORUM)
0.1	0.447
0.2	0.417
0.3	0.325
0.4	0.242
0.5	0.137