ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Graph Theory Analysis of Protein-Protein Interaction Network and Clustering proteins linked with Zika Virus

J. Susymary, R. Lawrance

Research Scholar, Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

Director, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

ABSTRACT: Graph mining is an evolving section particularly to dredge new and insight information from data that is represented as a graph. Graph data such as protein-protein interaction network is ubiquitous in real world so that graph theory approach to network can lead further findings of proteins associated with certain topological characteristic have specific biological function. Different graph mining techniques such as frequent subgraph mining, clustering, classification is available to evaluate the protein-protein interaction networks. One of the technique to find a group of proteins with similar biological function is clustering. Some of the graph based clustering methods include local neighborhood density search method, flow simulation method and population based stochastic search method. Molecular complex detection algorithm based on local neighborhood density search method over protein-protein interaction network of proteins related zika virus has been experimentally evaluated and demonstrated how interesting clusters are found.

KEYWORDS: graph mining; clustering; graph theory; protein-protein interaction; zika virus

I. Introduction

A cell is composed of several biochemical compounds such as DNA, RNA and proteins. Proteins are the most important molecule groups in a living cell. The central dogma of the cell function is that the information from the DNA is transmitted to RNA which is in turn transmitted to proteins. Fig. 1 shows the central dogma of a living cell.

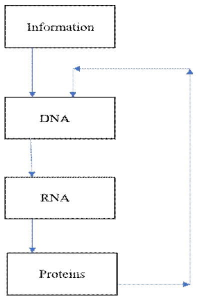


Fig. 1. Central dogma of living cell

Proteins are the information molecules which carries information from one cell to another. Not every protein interacts. Only proteins which possesses signaling properties will interact with each other. Fig.2 shows that the information from the donor cell is carried by the proteins with signaling properties and is passed to the receptor proteins in the target cell which is again passed to the nucleus of the cell which decide the response. Thus, the function of a living cell is performed by the interacting proteins.

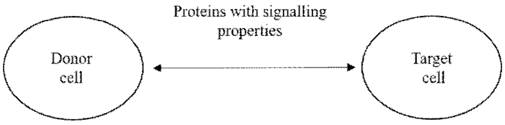


Fig. 2. Information transferring

Protein-protein interaction form large networks. It is the pairwise or complex representation of interacting proteins. Visualization and analysis of protein-protein interaction network helps to pin point the role of interacting proteins and brings a new insight about the function of proteins individually or as in a group.

The main significance of analysis of protein-protein interaction network include: identification of group of proteins that performs a specific biological function, finding function of specific proteins within a group of proteins as well as individually for therapeutic purposes.

Since proteins are responsible for all the biological functions in a cell, most proteins interact to form a group and perform a specific biological activity. Several methods available to analyse protein-protein interaction such as biological methods, vector algebra based methods, statistical methods and more over graph based methods. From the literature review, it has been learned that topological analysis of protein-protein interaction graph can lead to the better understanding of functions of proteins individually and as in a group.

A graph is a collection of vertices or nodes or points which are connected by a set of edges or links or arcs. G= (V, E) is a graph such that, each edge e∈E(G) is a pair of vertices (v_1, v_2) ∈ V(G). A vertex is a single point or a connection point in a graph. An edge in a graph G is an unordered pair of two vertices (v_1, v_2) such that v_1∈V(G) and v_2∈V(G). Fig. 3 shows an undirected unweighted graph with 6 vertices namely A, B, C, D, E, F and G.

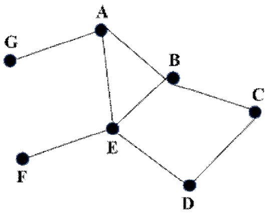


Fig. 3. Graph G

A loop is an edge that joins a vertex to itself. In a graph, G, and edge is a multiple edge if there is another edge in E(G) which joins the same pair of vertices. A simple graph is a graph with no loops or multiple edges.

The most important characteristic of a graph is the degree or connectivity of a vertex. The degree of a vertex is the number of other vertices connected to it. In the Fig. 3 the vertex A has degree 3 and the degrees of the vertices B, C, D, E, F and G are 3, 2,2,4,1,1 respectively.

Path between two nodes is the sequence of edges connecting those nodes. There are several paths for specific two nodes. The minimum number of edges required to reach a node from the other node is the shortest path between two nodes. In the Fig. 3 the shortest path between F and G is (F, E, A, G). A path is sometimes called a walk. A path is closed if its first and last vertices are same. Path length is the number of edges in the path. Distance within a network is measured in terms of path.

A cycle of length n, denoted Cn in a graph G is a closed path of length n. Two vertices are connected if and only if there exists a path from one vertex to another. A graph G is a connected graph if, for every vertex v, there is a path to every other vertex in V(G). A graph G is a tree if and only if it is a connected graph with no cycles and has exactly one simple path from one vertex to every other vertex.

A set of vertices C is a clique in the graph Gif, for all pairs of vertices v_1∈C and v_2∈C, there exists an edge (v_1, v_2)∈E(G). A complete graph with n vertices, denoted K_n, is a graph such that V(K_n) is a clique.

A subgraph is a subset of the vertices and edges of a graph. A subgraph S of a graph G is a set of vertices V(S)□V(G) and a set of edges E(S)□E(G). Every edge in E(S) must be an unordered pair of vertices (v_1, v_2) such that v_1∈V(S) and v_2∈V(S).

In a graph, G the subgraph S induced by a set of vertices N□V(G) is composed of V(S)=N and for all pairs of vertices $v_1 \in V(S)$ and $v_2 \in V(S)$, if $(v_1, v_2) \in E(G)$, then $(v_1, v_2) \in E(S)$.

Protein-protein interaction network can be modelled as an undirected, unweighted graph G= (V, E) where V is the set of proteins and E is the set of interactions such that the elements in E is a set of pair of proteins which interact with each other.

Graph theory and graph algorithms are well understood field of computers science. Graph mining is the process of extracting subgraphs from graphs to find a useful information regarding the data which the graph is associated. Several graph mining techniques are there to extract subgraphs. Frequent subgraph mining, clustering, classification etc. are some of the well-known techniques used in graph mining. Graph algorithms suits for one application may not suit for another.

From the publications [], it has been learned that protein-protein interactions have the power law feature of scale free networks. That is, few nodes are of high degree and others are of less degree. Since most proteins participate in only a few interactions and a few proteins participate in huge number of interactions, the protein-protein interaction network follows the power law.

Another characteristic of protein-protein interaction network is that, it possesses “small world effect”. That means, two nodes can be connected through a short path of few edges.

An important characteristic of protein-protein interaction network is disassortativity. That means highly connected nodes rarely directly link to each other.

Among the graph mining techniques, it is learned that graph clustering is very useful in mining group of proteins that performs a specific biological function. There are two types of protein-protein interaction clustering methods.

1. Distance based clustering: will not consider the topological properties of the network.
2. Graph based clustering: based on the topological properties of the network. The graph based clustering methods include:
   a. Local neighborhood density search method
   b. Flow Simulation method
   c. Population based stochastic search method

From the previously published papers it has been learned that analysis of topological properties of protein-protein interaction graph can pave a way to biological inference. Since it is decided to analyse the topological properties of the protein-protein interaction graph to extract the clusters, graph based clustering method has been applied to evaluate the network. A graph clustering algorithm known as Molecular Complex Detection (MCODE) based on the above-mentioned local neighborhood density search method for protein-protein interaction network related to zika virus has been evaluated to find interesting clusters.

II. LITERATURE REVIEW

Literature review helps to analyse the previous work related to the selected research topic. A theoretical base for the research topic can be achieved by exploring the history of the selected topic.

Przulj, N., *et al.* [2] introduce a comprehensive approach using graph properties on large PPI networks to support functional analysis and hypothesis generation. From the results, it has been inferred that by uncovering the network properties of protein interactions, functional annotation for uncharacterized proteins can be computed.

Thomas, A., *et al.* [3] present a simple model for the underlying structure of protein-protein pairwise interaction graph. The frequency of the number of connections per protein under this model does not follow a power law.

Wu, X.R., *et al.* [4] applied graph model theory to analyse the protein-protein interaction networks of seven organisms. Three topological properties were utilized to characterize the process of these protein-protein interaction networks. The experimental results show that degree distributions of the seven protein interaction networks follow the power-law distribution quite well, which means that protein interaction networks are scale-free network with a few nodes having high degree and the rest having low degree. Clustering coefficient obtained for the network indicates high clustering behavior for the seven protein interaction networks. In addition, it can be also found that the shortest-path length and the average shortest-path length calculated is relatively small compared to the large network size. This property is usually referred to as a small-world effect.

Ucar, D., *et al.* [5] propose a novel refinement method based on neighborhoods and the biological importance of hub proteins. It shows that the refinement improves the functional modularity of the PPI (Protein-Protein Interaction) graph and leads to effective clustering into dense components. A detailed comparison of these dense components with the ones obtained from the original PPI graph reveal three major benefits of the refinement: i) Enhancement of existing functional groupings; ii) Isolation of new functional groupings; and iii) Soft clustering of multi-functional hub proteins to multiple functional groupings.

Lawrance, R., *et.al.* [6] has been reviewed most relevant mining association rules as well as main issues when discovering efficient and practical method for microarray gene association analysis.

Lawrance, R., *et.al.* [7] applied association rule to get efficient and important patterns and significant relations among microarray genes.to reveal fatal and crucial reasons for diseases. It provides improving prediction for diseases and treatment.

From the literature review, it has been inferred that topological properties can be used to analyse the protein-protein interaction networks.

III. Dataset Description

In 1947, a virus has been first discovered in a rhesus monkey in Uganda's Zika forest. It is named as Zika virus. Some years later the first human case was reported in Nigeria. At first the disease linked with this virus are fever, malaise, skin rash, conjunctivitis, muscle pain, join pain, headache etc. after that it is found that the virus is linked with virulent form of diseases related to neurological disorders swelling of brain and spinal cord, microcephaly-abnormally small heads and brains in foetuses etc. In India, the zika virus which has no cure or vaccine was found 64 years ago, spread by air travelers. Vigilant attention has been raised against this virus since it can be spread by aedes aegypti mosquitoes which is a carrier of dengue, yellow fever, chikungunya etc.

It has been observed that there are ten proteins linked with zika virus. Three structured proteins C, prM, E and seven non-structured NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5.

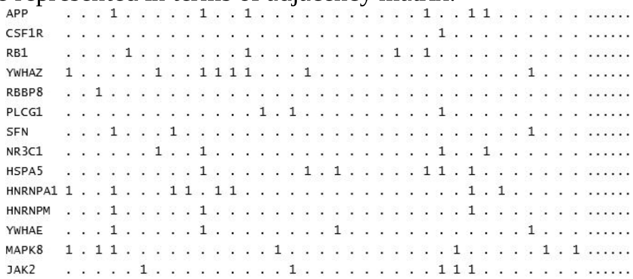
A. DATASET

BioGrid [8] is a database of predicted interactions including protein-protein. The BioGrid database can be accessed directly in R studio and protein-protein interaction network can be obtained by specifying the protein identifiers.

Some proteins namely 'NS1', 'NS3', 'NS4' and 'NS5' linked with zika virus is uploaded to BioGrid and protein-protein interaction network related to these proteins is obtained by using 'igraph' [9] package in conjunction with 'ProNet' [10] package in R studio development environment. The resulting network contain 264 proteins and 2159 interactions.

B. NETWORK REPRESENTATION

The obtained network can be represented in terms of adjacency matrix.



```
APP     . . . 1 . . . . . 1 . . 1 . . . . . . . . . . . 1 . . 1 1 . . . . . . ......
CSF1R   . . . . . . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . ......
RB1     . . . . 1 . . . . . . . 1 . . . . . . . . . 1 . 1 . . . . . . . . . . ......
YWHAZ   1 . . . . . 1 . . 1 1 1 1 . . . 1 . . . . . . . . . . . . . . 1 . . . ......
RBBP8   . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ......
PLCG1   . . . . . . . . . . . . . 1 . 1 . . . . . . . . . 1 . . . . . . . . . ......
SFN     . . . 1 . . . 1 . . . . . . . . . . . . . . . . . . . . . . . 1 . . . ......
NR3C1   . . . . . . 1 . . 1 . . . . . . . . . . . . . . . 1 . . 1 . . . . . . ......
HSPA5   . . . . . . . . . 1 . . . . . . 1 . 1 . . . . . 1 1 . 1 . . . . . . . ......
HNRNPA1 1 . . 1 . . . 1 1 . 1 1 . . . . . . . . . . . . . . . 1 . 1 . . . . . ......
HNRNPM  . . . 1 . . . . . 1 . . . . . . . . . . . . . . . . . 1 . . . . . . . ......
YWHAE   . . . 1 . . . . . 1 . . . . . . . . 1 . . . . . . . . . . . . 1 . . . ......
MAPK8   1 . 1 1 . . . . . . . . . . 1 . . . . . . . . . . . 1 . . . . . 1 . 1 ......
JAK2    . . . . . 1 . . . . . . . . . 1 . . . . . . . . . 1 1 1 . . . . . . . ......
```

Fig. 4. Adjacency matrix representation of the network

C. *NETWORK VISUALIZATION*

The obtained network can be visualized using ‘igraph’ object.

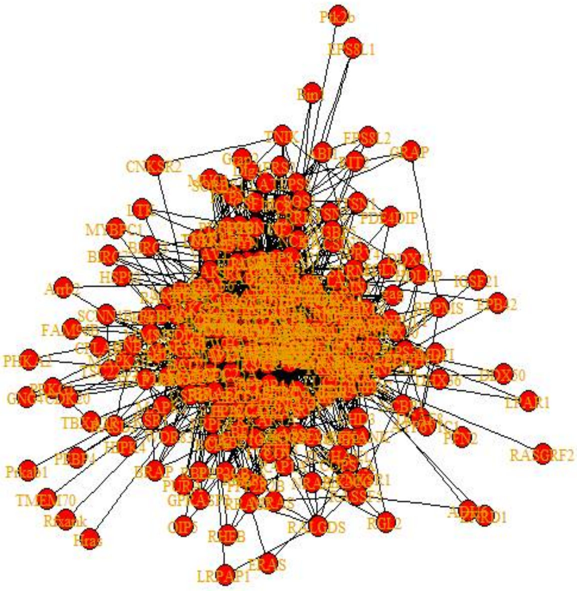
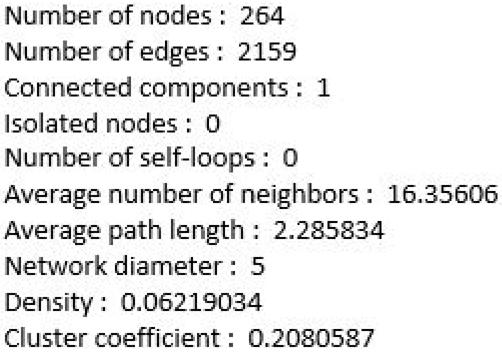


Fig. 5. Protein-protein Interaction network

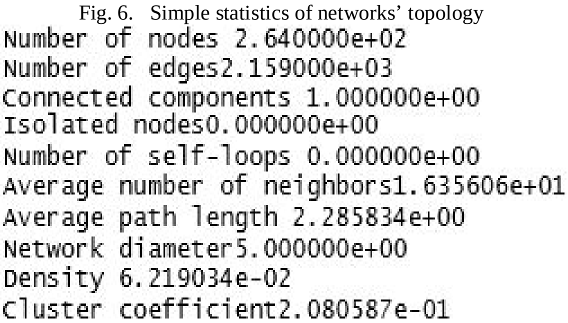
D. *TOPOLOGICAL ANALYSIS OF THE NETWORK*

The below three figures show the topology of the protein-protein interaction network associated with the specified proteins.



```
Number of nodes :  264
Number of edges :  2159
Connected components :  1
Isolated nodes :  0
Number of self-loops :  0
Average number of neighbors :  16.35606
Average path length :  2.285834
Network diameter :  5
Density :  0.06219034
Cluster coefficient :  0.2080587
```

Fig. 6. Simple statistics of networks’ topology



```
Number of nodes 2.640000e+02
Number of edges2.159000e+03
Connected components 1.000000e+00
Isolated nodes0.000000e+00
Number of self-loops 0.000000e+00
Average number of neighbors1.635606e+01
Average path length 2.285834e+00
Network diameter5.000000e+00
Density 6.219034e-02
Cluster coefficient2.080587e-01
```

Fig. 7. Specific statistics of degree distribution

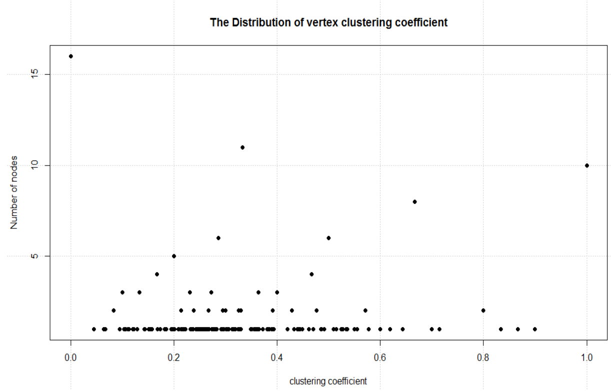
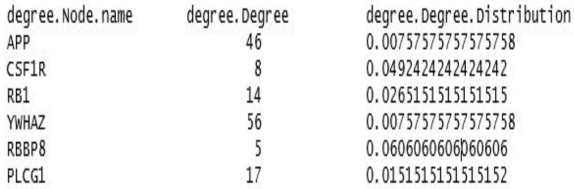


Fig. 8. Clustering coefficient of the vertices in the network

Fig. 9 and Fig. 10 show the degree disstribution of the proteins in the network.



```
degree.Node.name    degree.Degree    degree.Degree.Distribution
APP                            46    0.00757575757575758
CSF1R                           8    0.0492424242424242
RB1                            14    0.0265151515151515
YWHAZ                          56    0.00757575757575758
RBBP8                           5    0.0606060606060606
PLCG1                          17    0.0151515151515152
```

Fig. 9. Clustering coefficient of the network

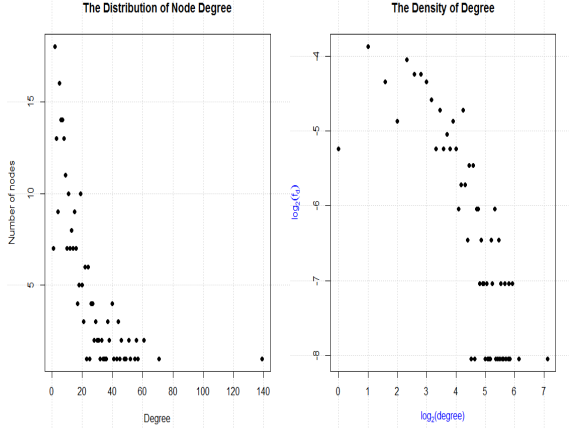
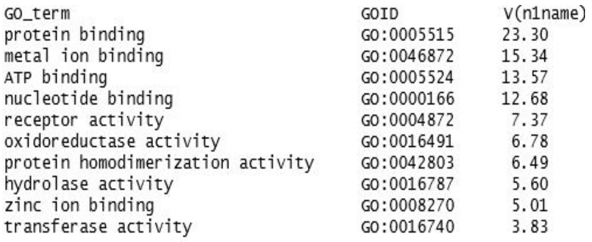


Fig. 10. Degree distribution of vertices in the network

E. GENE ONTOLOGY

Gene Ontology (GO) database provides the molecular function, cellular component and biological process related to the proteins. Fig. 11 and Fig. 12 show Gene Ontology (GO term) of the proteins associated with the network is shown below. It has been observed that 23.30 percent of the total proteins associated with network involves in protein binding, 15.34 percent metal ion binding and so on.



```
GO_term                               GOID          V(n1name)
protein binding                       GO:0005515    23.30
metal ion binding                     GO:0046872    15.34
ATP binding                           GO:0005524    13.57
nucleotide binding                    GO:0000166    12.68
receptor activity                     GO:0004872     7.37
oxidoreductase activity               GO:0016491     6.78
protein homodimerization activity     GO:0042803     6.49
hydrolase activity                    GO:0016787     5.60
zinc ion binding                      GO:0008270     5.01
transferase activity                  GO:0016740     3.83
```

Fig. 11. GO-term of proteins in the network

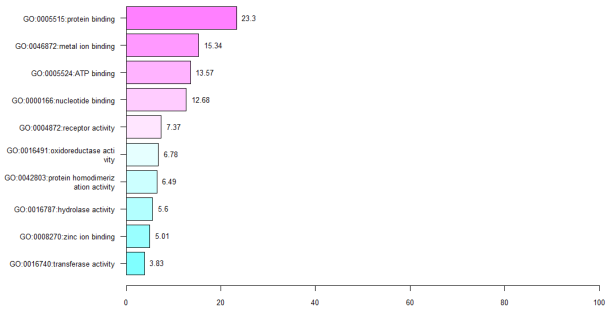


Fig. 12. GO-term plot

IV. METHODOLOGY

Graph based clustering methods in protein-protein interaction network graph use specialized clustering techniques. Since the main goal is to find the densely-connected proteins to perform a specific function, one of the graph based clustering method known as local neighborhood density search has been experimentally discussed.

A. *LOCAL NEIGHBORHOOD DENSITY SEARCH METHOD*

Local neighborhood density search method is based of vertex optimization strategy. That is, for a subgraph, each vertex is connected to so many vertices within the protein-protein interaction network.

Molecular Complex Detection (MCODE), first proposed by Bader and Hogue [11] is one of the algorithm based on local neighborhood density search to detect highly connected vertices. It has three phases:

Phase 1: Vertex weighting: The algorithm assigns a weight to each vertex with respect to its local neighborhood density. That is based on the number of vertices connected to each vertex.

Phase 2: Cluster finding: Takes input as the vertex weighted graph, then staring from the top weighted vertex, it iteratively moves around the top weighted vertex and assign the vertices whose weights are above user defined threshold weight, which is a given percentage away from the weight of the top weighted vertex. This is the vertex weight parameter (vwp). If a vertex is included, its neighbors are iteratively checked in the same manner to see if they are part of the cluster. A vertex is not checked more than once. This process stops once no more vertices can be added to the cluster based on the given threshold and is repeated for the next highest unseen weighted vertex in the network. In this way, the densest regions of the network are identified. The vertex weight parameter decides the density of the resulting cluster. A threshold that is closer to the weight of the top weighted vertex identifies a smaller, denser network region around it.

Phase 3: Optional post processing to filter or add proteins in the resulting clusters. Remove singly connected vertices and expand cluster cores by one neighbor.

MCODE Algorithm:
Input network
Give each vertex a score based on the vertices connected to it
High score = vertex in the dense region

Find complexes
Optionally expand or contract complexes

Resulting clusters are scored and ranked. The cluster score is defined as the product of the cluster subgraph, C= (V, E), density and the number of vertices in the cluster subgraph ($D_C \times |V|$). This ranks larger, more dense clusters higher in the results.

V. EXPERIMENTAL RESULTS

Fig. 13 shows the network visualization of the 5 clusters resulting from the MCODE algorithm. Fig. 14 shows the 5 clusters and associated proteins obtained from the result of MCODE algorithm. Fig. 15 shows the topological comparison of the protein-protein interaction network and the resulting clusters from the algorithm. Fig. 16 shows the GO-term of the proteins associated with each cluster.

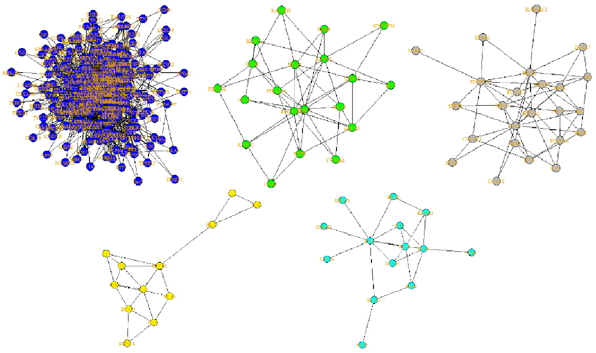
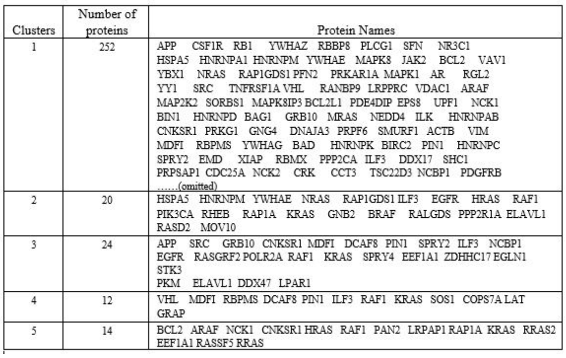
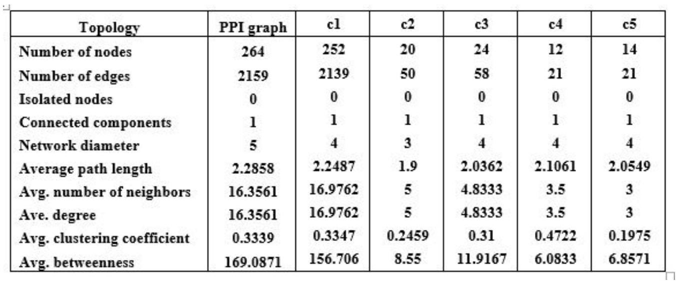


Fig. 13. Protein cluster result from MCODE algorithm



Clusters	Number of proteins	Protein Names
1	252	APP CSF1R RB1 YWHAZ RBBP8 PLCG1 SFN NR3C1 HSPA5 HNRNPA1 HNRNPM YWHAE MAPK8 JAK2 BCL2 VAV1 YBX1 NRAS RAP1GDS1 PFN2 PRKAR1A MAPK1 AR RGL2 YY1 SRC TNFRSF1A VHL RANBP9 LRPPRC VDAC1 ARAF MAP2K2 SORBS1 MAPK8IP3 BCL2L1 PDE4DIP EPS8 UPF1 NCK1 BIN1 HNRNPD BAG1 GRB10 MRAS NEDD4 ILK HNRNPAB CNKSR1 PRKG1 GNG4 DNAJA3 PRPF6 SMURF1 ACTB VIM MDFI RBPMS YWHAG BAD HNRNPK BIRC2 PIN1 HNRNPC SPRY2 EMD XIAP RBMX PPP2CA ILF3 DDX17 SHC1 PRPSAP1 CDC25A NCK2 CRK CCT3 TSC22D3 NCBP1 PDGFRB ……(omitted)
2	20	HSPA5 HNRNPM YWHAE NRAS RAP1GDS1 ILF3 EGFR HRAS RAF1 PIK3CA RHEB RAP1A KRAS GNB2 BRAF RALGDS PPP2R1A ELAVL1 RASD2 MOV10
3	24	APP SRC GRB10 CNKSR1 MDFI DCAF8 PIN1 SPRY2 ILF3 NCBP1 EGFR RASGRF2 POLR2A RAF1 KRAS SPRY4 EEF1A1 ZDHHC17 EGLN1 STK3 PKM ELAVL1 DDX47 LPAR1
4	12	VHL MDFI RBPMS DCAF8 PIN1 ILF3 RAF1 KRAS SOS1 COPS7A LAT GRAP
5	14	BCL2 ARAF NCK1 CNKSR1 HRAS RAF1 PAN2 LRPAP1 RAP1A KRAS RRAS2 EEF1A1 RASSF5 RRAS

Fig. 14. Result from MCODE algorithm



Topology	PPI graph	c1	c2	c3	c4	c5
Number of nodes	264	252	20	24	12	14
Number of edges	2159	2139	50	58	21	21
Isolated nodes	0	0	0	0	0	0
Connected components	1	1	1	1	1	1
Network diameter	5	4	3	4	4	4
Average path length	2.2858	2.2487	1.9	2.0362	2.1061	2.0549
Avg. number of neighbors	16.3561	16.9762	5	4.8333	3.5	3
Ave. degree	16.3561	16.9762	5	4.8333	3.5	3
Avg. clustering coefficient	0.3339	0.3347	0.2459	0.31	0.4722	0.1975
Avg. betweenness	169.0871	156.706	8.55	11.9167	6.0833	6.8571

Fig. 15. Topological comparison of the network and resulting clusters

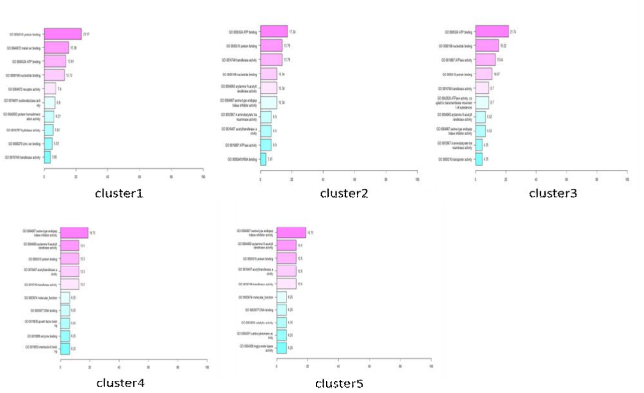


Fig. 16. GO-term for the proteins associated with 5 clusters

VI. CONCLUSION

MCODE algorithm results highly connected, dense clusters. Topological properties of the clusters can be evaluated to find the function of proteins associated within clusters. The biological validation can be done using functional enrichment analysis by incorporating proteins to the gene ontology consortium (GO term) which describes the function of proteins.

REFERENCES

[1] Schaeffer, S,E., “Survey Graph Clustering”, Science Direct, Coputer Science Review 1, 2007, pp: 27-64

[2] Przulj, N., Wigle, D.A., and Jurisica, I., “Functional topology in a network of protein interactions”, Oxford Journals, Vol. 20, No. 3, 2004, pp: 340 - 348

[3] Thomas, A., Cannings, R., Monk, N.A.M., and Cannings, C., “On the structure of protein-protein interaction networks”, Biochemical Society Transactions, Vol. 31, 2003, pp: 1491 - 1496

[4] Wu, X.R., Zhu, Y., and Li, Y., “Analyzing Protein Interaction Networks via Random Graph Model”, International Journal of Information Technology, Vol. 11, No. 8, 2005, pp: 125 - 132

[5] Ucar, D., Asur, S., Catalyurek, U., and Parthasarathy, S., “Improving Functional Modularity in Protein-Protein Interactions Graphs using Hub-Induced Subgraphs”, Proceedings of the 10th European Conference on Principles of Data Mining and Practice of Knowledge Discovery, Springer Berlin Heidelberg, 2006, pp: 371 – 382

[6] Alagukumar, S., and Lawrance, R., "A Selective Analysis of Microarray Data Using Association Rule Mining." Procedia Computer Science vol.47 pp: 3-12, 2015

[7] Alagukumar, S., and Lawrance, R., “Algorithm for Microarray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals”, IJCA Proceedings on National Conference on Research Issues in Image Analysis and Mining Intelligence NCRIIAMI 2015(1),9-14, June 2015

[8] https://thebiogrid.org/download.php

[9] http://igraph.org

[10] Wu, X, Y., and Xia, X, Y., “ProNet Tutorial”, pp: 1-12, 2015

[11] Bader, G, D., and Hogue, C, W, V., “An automated method for finding molecular complexes in large protein interaction networks”, BMC Bioinformatics. 2003; 4: 2.