Scientific Journal of Impact Factor(SJIF): 3.134
e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2, Issue 3, March -2015

# Clustering Approaches for the Functional Module Detection from PPI Network: A Review

Manali R. Modi[1], Merry K.P[2]

[1]Department of Computer Engineering, Marwadi Education Foundation's Group of Institutions Rajkot, Gujarat
[2]Department of Information Technology, Marwadi Education Foundation's Group of Institutions Rajkot, Gujarat

***Abstract***- *Proteins are the building bit of the living creatures and it assumes transcendent part to fulfill natural procedures of the life forms. Protein Protein Interaction (PPI) network is a sort of system that delineates the fundamental part in organic exercises. PPI is the corner stone of all the biological processes occurring in the organisms. In PPI network, protein interacts with one another to incorporate large molecules. Functional module detection refers to the set of proteins which partake in the same organic courses of biological action. Clustering in the PPI system setting assembles proteins which impart a bigger number of communications. The consequences of clustering can clarify the construction of the PPI arrangement and prescribe possible limits for detecting modules which were at one time uncharacterized. To understand the biological mechanism of the organisms, study of the functional module detection by clustering is required.*

***Keywords***- *PPI network, functional modules, protein, clustering techniques for functional module detection*

## I. INTRODUCTION

Bioinformatics is the interdisciplinary field which involves study and analysis of biological data by using various fields such as mathematics, statistics, science, mining etc. In bioinformatics Protein-Protein Interaction (PPI) contributes a lot for analyzing the biological processes of the organisms. PPI helps a ton for examining the organic procedures of the life forms. Protein-Protein Interaction is a sort of system in which proteins are communicating with each for the natural exercises which are done in the life forms. It is very critical to break down these proteins, collaborate with one another to complete metabolic exercises. For this reason utilitarian module recognition got to be fundamental.

Proteins are the main and single constitute in the organisms that perform different functions. Proteins are involved in biological processes and intermingle with each other. This interaction makes Protein-Protein Interaction network. PPI system can be demonstrated as an undirected graph *G(V,E)*, where vertices are proteins and edges are protein communications. PPI network are composed of various modules. Molecular biology is becoming a highly changing science in which identification of the functional module from the PPI network is crucial. Major Problem in PPI network is detection of functional modules from large network of PPI. It means mining or extracting some interesting cluster from large datasets. Identification of the functional module from the existing network is the decisive problem to understand the function of the organisms [1]. In modern era, high-throughput experimental techniques have made notifiable advances in identifying function from the protein-protein interaction on the large scale. Computational techniques are utilized for catching useful modules for plotting the communication of proteins to see how diverse proteins are schematized to perform various biological functions from higher level substructure of proteins. PPI network can be created from the analysis of the functional module detection. Thus it becomes omnipotent to identify functional module from the pair of proteins and its network.

In the next section, the structure of a review paper is described. Section II contains Importance of functional module detection from PPI network. Section III contains major challenges for identifying modules. Section IV contains associated research study in this area. Section V describes state-of-art methods. Section VI contains comparative study. Section VII contains conclusion of the paper.

## II. IMPORTANCE OF FUNCTIONAL MODULE DETECTION

The twenty first century is the era of natural life science. With the ultimate of sequencing the human genome, proteomic exploration turns into a standout among the most paramount zones of life science. Proteomics is the systematical examination of the distinctive properties of proteins to give clear portrayals of the structure, limit and control of regular structures in wellbeing and disease [1].

Normally, protein rarely go about as solitary segregated element; rather, protein included in the identical cell forms regularly associate with one another to consolidate into an extensive atom to perform the organic capacities. As a case in point, the courses of action and exercises of the hereditary substance duplicate, quality outflow control, cell signal transduction, digestion system, cell spread, and cell apoptosis rely on PPI. PPI are the foundation of the organic

@IJAERD-2015, All rights Reserved 

methodologies occurring in the inner side of the life forms. Accordingly, the examination of PPI systems regularly serves as the premise to a finer understanding of cell association, methods, and capacities and along these lines clarification of protein communication is a focal issue in science [1].

## III. MAJOR CHALLENGES

There are many experimental and active proposed algorithms among a range of issued approaches, still some structural features of the PPI network avoids the module of protein identification. Following challenges are emerge from the detection of the modules:

### *A. The untrustworthiness of the interaction data of proteins*

The PPI dataset acquired from the biological investigation are unfinished and noisy. There is a reason behind this is the false-positive degree of the output of PPI information is higher as contrasted to the little scale information. Subsequently, to show signs of improvement in mining, the location calculations ought to enhance their robustness. Despite the fact that some successful methodologies focused around different data combinations now are accessible, to learn the negative impacts fetched by the loud information on the discovery quality is a vital issue in PPI useful module detection [2].

### *B. The efficiency of revealing algorithms to detect modules*

A PPI dataset is balanced of many proteins and considerably more connections. It is a vast scale complex system. Subsequently, time complexity is a pragmatic prerequisite for the recognition approaches. Conversely, most algorithm calculations focused around the computational methodologies have higher time complex nature, which will restrict the improvements and handy applications. Thus, investigators need to pay additional considerations towards this challenge [1].

### *C. The overlapping modules of PPI network*

Most prevailing clustering methodologies experience issues for breaking down the PPI information primarily because of the way that a protein can have a few distinctive functions. To be specific, a single protein possibly will be encompassed in one or many modules.

## IV. ASSOCIATED RESEARCH STUDY

Numerous exploration work has been done since last few years by researchers. Qi Yu et al. [4] projected innovative algorithm MOFinder aimed at the overlapping functional module detection from PPI network. MOFinder uses Approximate Minimum Degree Ordering (AMD) algorithm and sliding window protocol. Jianxin Wang et al. [16] introduced Overlapping Hierarchical-Protein Interaction Network (OH-PIN) algorithm for hierarchical and overlapping functional modules identification from PPI network. Bingjing Cai et al. [17] introduced Affinity Purification-Mass Spectrometry (AP-MS) algorithm for uncovering of protein modules from large PPI network. This algorithm devices to recognize the cluster of prey protein that are meaningfully connected with the analogous group of bait protein.

Min Li et al. [18] proposed a framework to extricate complexes of the proteins and functional modules by assimilating genetic factor expression data into PPI datasets. They proposed DFM-CIN algorithm for discovery of functional module based on the identified complexes. Jansen et al. [18] inspected the connection of PPI connections with expression levels of mRNA and scored interpretation action in protein complexes. Tarnow and Mewes [19] recycled the super paramagnetic methodology to assess the multi-information relationships and developed a chart of communicated genes for discovering functional modules. Han et al. [20] examined the PPI dataset of yeast, and wide-open two types of pivot proteins as the modules.

Zhang et al. [21] joins the LGT (Line Graph Transformation) and CPM to find covering framework module of proteins and amasses the covering protein module framework. Wu et al. [22] put forwarded COACH to anticipate buildings by spotting complexes of protein focus and after that including associations. The Local Protein Community Finder [4], (LPCF), uses two close-by bundling computations to discover a gathering near addressed adjacent proteins. Reinhardt and Bornholdt [24] proposed a method to find overlapping gatherings that records the figure to a zero temperature q-pott model with closest adjacent associations. Zhang et al. [24] join the thought of seclusion capacity Q, ghastly unwinding, fluffy cluster means grouping strategy to discern covering group structure. Wang et al. [25] advised Betweenness Commonality Decomposition (BCD) calculation that customs edge shared trait and betweenness of the edge for the recognition of practical modules from extensive PPI system.

## V. STATE-OF-ART METHODS

From the former span, the PPI information have been dissected by high-throughput exploratory systems, for example, two-hybrid frameworks, protein chip innovation, mass spectrometry and also far reaching of the application of content

@IJAERD-2015, All rights Reserved 

mining in PPI systems. Traditional tactic to determine function module is to catch relation between an unannotated proteins and supplementary protein by means of sequence matching algorithm [3]. Numerous incorporated PPI systems have been constructed from these information sources. Through the expansive size of PPI network information, how to productively distinguish functional modules turn into an imperative logical issue and a paramount examination theme in the post genomic time.

There are various experimental test techniques to recognize practical modules in PPI systems. The natural examination demonstrates that a protein unpredictable in PPI network is an atomic arrangement steady in both functional and structure, this implies the nearly associated protein zones in PPI network compare to protein utilitarian modules. By discovering the minimally joined structures from PPI systems, functional modules can be distinguished. Seeing this essential thought, the recognition techniques for protein functional modules focused around machine learning and information mining have become quick and come with valuable supplements in the trial systems [1].

On the basis of the computational approach, for detection of functional module there are six existing approaches which are as mentioned below.

### *A. Graph Theoretic Approach*

This methodology emphases upon the construction topology analysis in the PPI network. It gets deeply partitioned into three subcategories of clustering methods which are vividly described as follows:

*1) Density Based Clustering [1, 6, and 15]*

It searches for the densely connected sub-graphs to identify functional modules from the network. Molecular Complex Detection (MCODE) algorithm is proposed by Bader et al. [6] by assigning weights to each node of the PPI network by acquiring native neighbor density of the node and fetches node with highest weight as the pit node and augments the cluster to form preliminary cluster. MCODE algorithm can handle large PPI network effectively due to its polynomial time complexity. Adamcsek et al. [15] evolve software CFinder which is used to uncover the overlapping cluster from PPI network. Other algorithms are MINE, DPClus which are also used to identify the functional module detection [1].

*2) Hierarchy Based Clustering [7]*

It is applicable to the biological network due to the hierarchical nature of the PPI network. This method merges nodes or divides a graph into sub-graphs to cluster the protein network on the basis of the nonlocal features. UVCluster algorithm is recommended by Arnau et al. [7] for the iterative merges of the proteins by measuring distance between the proteins.

*3) Partition Based Clustering [8]*

It partitions the network and iteratively finds the protein on the edge of the cluster to a neighboring group to pursuit the superior grouping with least expenses. These strategies are not difficult to execute and comprehend clustering network of the protein. King et al. [8] anticipated Restricted Neighborhood Search Clustering (RNSC) that is a neighborhood examine clustering approach to spot the complexes of the proteins that can identify the best partition through cost function.

### *B. Flow Simulation Approach [9]*

It is the approach to scrutinize the degree of the topological and biological features affect by every protein over the extra protein. Signal Transduction Model (STM) is the novel algorithm reflects the signal transduction that selects representative protein for every cluster and modify cluster iteratively based on signal transduction [9].TRIBE-MCL (Markov Clustering), CASCADE, GFA are the other effective algorithm which are simulate a functional and biological flow of the protein.

### *C. Spectral Based Clustering [10, 11]*

This clustering approach converts the problem to quadratic optimization with some constraints through the matrix analysis methodology. These methods are generally useful for large and complex datasets. A Diffusion Model Based Spectral Clustering (ADMSC) steadily solves structure of collection of the PPI network by using random walks [10]. It has advantage of fast partitioning of the PPI network into appropriate biological cluster with approximate equal size. Qin et al. [11] had developed a spectral clustering technique which is useful to perceive complexes of the proteins and functions.

### *D. Supervised Clustering Approach [12]*

Densely connected structure sometimes absent in the PPI network which becomes the problem for the functional module detection. So supervised clustering originates in to the motion. SCI-BIN recommended by Qi et al. [12] for the choice of the fundamental chart topology example and organic properties as the hub and then develops the Bayesian network for respective graph. Described algorithms integrate topologies and biological evidences and learn the functional module from the PPI network. This algorithm has higher precision and recall rate for identifying the functional modules.

@IJAERD-2015, All rights Reserved 

*E. Core Attachment Based Approach [13]*

This approach is based on the three steps: 1. Initial step is forecast of center proteins. 2. Then recognize the connection of protein of each center protein and uproot the immaterial center protein. 3. Last step is to calculate the score and sort the order of important protein complexes [1]. Wu et al. [13] evolve with COACH algorithm which finds the complexes of the proteins core by characterizing center vertices in the adjacent and afterward incorporates connection into center to structure organic structure.

*F. Swarm Intelligence Based Approach[14]*

This approach discovers immediate prime arrangement of more troublesome issue by recreating gathering of the social bugs like ants, bees and wasps. Sallim et al. [14] come up with Ant Colony Optimization Protein Interaction Network (ACOPIN) calculation which utilizes ground dwelling insect state enhancement into issue by using the discovery of protein functional modules as the advancement issue of taking care of travelling salesman issue. Functional flow based clustering approach depends on artificial bee colony was developed by Wu et al [14]. NACO-FMD, ACO-MAE, ABC-IFC are some important algorithm for finding functional module detection from the PPI network. NACO-FMD integrates the topological characteristics with the functional modules of the PPI network.

## VI. COMPARATIVE STUDY

Below TABLE I shows analysis of the existing algorithm with its approach brief description and its limitation. And TABLE II describes modules with its functional percentage that are covered by different algorithm. From this review paper it is watched that MOFinder outflanks when applying to yeast and human datasets for overlapping module detection. This relative study will be helpful for the examination of the diverse algorithm which are valuable for the detection of protein modules. Despite the fact that LPCF has most anticipated proteins and covered proteins it has less utilitarian rate.

TABLE I. ANALYSIS OF EXISTING ALGORITHM

| Existing Algorithm | Description | Limitation |
|---|---|---|
| **MCODE [6]** **Cfinder [15]** | Identifies densely connected group of proteins. | It can detect only connected graphs of proteins within the PPI network. |
| **UVCluster [7]** | Based on the agglomerate and divisive algorithm. | Not able to detect overlapping functional modules. |
| **RNSC [8]** | Separate sparsely connected nodes from the PPI network. | Not able to detect overlapping functional modules. |
| **TRIBE-MCL** **STM [9]** | Simulate a biological and functional flow. | It needed sophisticated approach for effectively simulate the stochastic behavior of the network. |
| **ADMSC [10]** | Uses method of matrix analysis for detection of function modules. | It is used for only large and complex organisms. |
| **SCI-BIN [12]** | Uses predefined known features to | It requires deep knowledge about the Bayesian network for |

@IJAERD-2015, All rights Reserved 

| Existing Algorithm | Description | Limitation |
|---|---|---|
| | make clusters. | the complex graph. |
| **CORE COACH [13]** | Employs core-attachment protein relation to cluster. | It has limitation of its time complexity of O(\|v\|3). |

TABLE II. COMPARISION OF EXISTING ALGORITHM

| Existing Algorithm | Predicted Modules | Covered Proteins | Functional Percentage |
|---|---|---|---|
| MCODE | 64 | 64 | 85.70% |
| CFinder | 53 | 184 | 81.10% |
| COACH | 382 | 861 | 46.60% |
| LPCF | 1601 | 4549 | 49.50% |
| MOFinder | 125 | 335 | 90.40% |

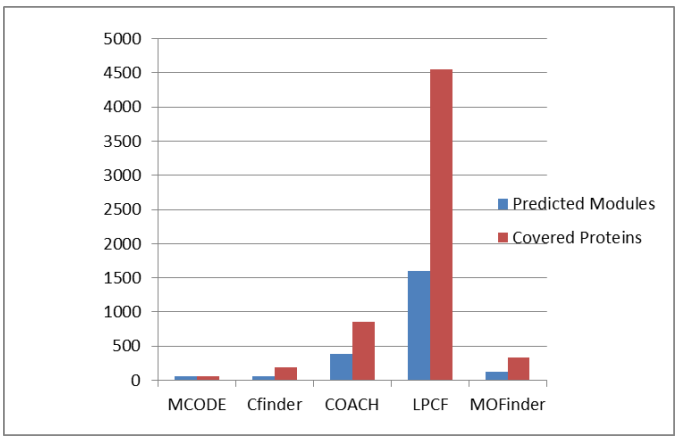


Fig I: Predicted modules and Covered proteins

Fig I shows modules that are predicted and covered proteins by the algorithm. As shown in the fig I LPCF algorithm has highest predicted proteins and covered proteins as compared to other algorithms. For this exploration DAVID tool is used for the mining of modules from the network of huge PPI.

@IJAERD-2015, All rights Reserved 

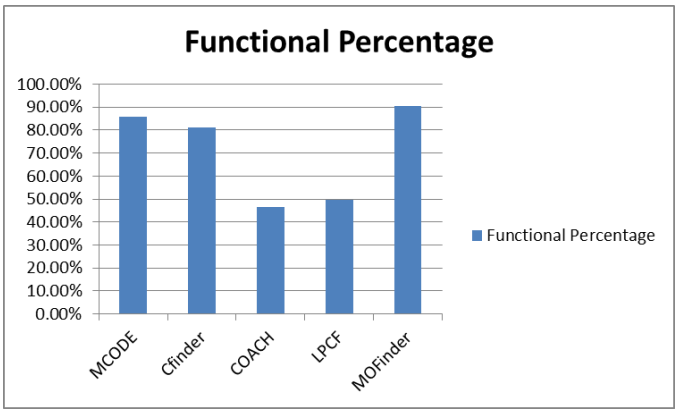


Fig II: Functional percentage of identifying modules

Fig II shows functional percentage of identifying modules from the large network by different algorithms. It can easily seen from the fig II that MOFinder shows highest functional percentage for detecting functional modules. Thus MOFinder has highest accuracy for discovering overlapping modules from PPI complex network.

## VII. CONCLUSION

Presented review paper portrays point by point approaches which are utilized as a part of recognizing useful modules from the PPI system. Likewise we had examined each algorithm with its approach, limitations, predicted proteins, covered proteins and functional percentage of modules which are actually identified from the existing algorithm. Our future arrangement is to enhance our work by introducing algorithm which improves some functionality of identifying functional modules.

## REFERENCES

[1] Junzhong Ji, Aidong Zhang, Chunnian Liu, Xiaomei Quan and Zhijun Liu, "Survey: Functional Module Detection from Protein-Protein Interaction Networks", *IEEE Transaction on Knowledge and Data Engineering, vol. 26, no. 2, Feb 2014, pp.261-273.*

[2] Min Li, Xuehong Wu, Jianxin Wang and Yi Pan, "Towards the Identification of Protein Complexes and Functional Modules by Integrating PPI Network and Gene Expression Data" *BCM Bioinformatics, 2012, pp.1-12.*

[3] L. Shi, Y. R. Cho and A. Zhang, "Prediction of Protein Function from Connectivity of Protein Interaction Network", *International Journal of Computational Bioscience, vol. 1, 2010, pp. 1-5.*

[4] Qi Yu, Gong-Hua Li and Jing-Fei Huang, "MOfinder: A Novel Algorithm for Detecting Overlapping Modules from Protein-Protein Interaction Network ", *Hindawi Publishing Corporation, Journal of Biomedicine and Biotechnology, 2012, pp. 1-10.*

[5] Kire Trivodaliev, Aleksandra Bogojeska and LjupcoKocarev, "Exploring Function Prediction in Protein Interaction Networks via Clustering Methods, "PLOS ONE, *June 2014, pp. 1-13.*

[6] G. D. Bader and C. W, Hogue,"An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks," *BMC Bioinformatics, vol. 4, no. 2, 2003, pp. 1-8.*

[7] V. Arnau, S. Mars and I. Marin,"Iterative Cluster Analysis of Protein Interaction Data, "Bioinformatics, *vol. 21, no. 3, 2005, pp. 364-370.*

[8] A.D. King, N. Przulj, and I. Jurisica, "Protein Complex Prediction via Cost-Based Clustering,"*Bioinformatics, vol. 20, no. 17, 2004, pp. 3013- 3020.*

[9] W. Hwang, Y.R. Cho, A. Zhang, and M. Ramanathan,"A Novel Functional Module Detection Algorithm for Protein-Protein Interaction Networks," *Algorithms for Molecular Biology, vol. 1, 2006, pp. 24.*

[10] K. Inoue, W. Li, and H. Kurata, "Diffusion Model Based Spectral Clustering for Protein-Protein Interaction Networks," *PLoS ONE, vol. 5, no. 9, 2010, pp. e12623.*

[11] G. Qin and L. Gao, "Spectral Clustering for Detecting Protein Complexes in Protein-Protein Interaction (PPI) Networks," Mathematics and Computer Modeling, vol. 52, 2010, pp. 2066-2074.

[12] Y.J. Qi, F. Balem, C. Faloutsos, J. Klein Seetharaman, and Z. Bar- Joseph, "Protein Complex Identification by Supervised Graph Local Clustering," *Bioinformatics, vol. 24, no. 13, 2008, pp. 250-260.*

[13] M. Wu, X.L. Li, C.K. Kwoh, and S.K. Ng, "A Core-Attachment Based Method to Detect Protein Complexes in PPI Networks,"*BMC Bioinformatics, vol. 10, article 169, 2009.*

[14] S. Wu, X.J. Lei, and J.F. Tian, "Clustering PPI Network Based on Functional Flow Model through Artificial Bee Colony Algorithm,"*Proc. Seventh Int'l Conf. Natural Computation, 2011, pp. 92-96.*

[15] B. Adamcsek, G. Palla, I.J. Farkas, I. Derenyi, and T. Vicsek, "CFinder: Locating Cliques and Overlapping Modules in Biological Networks," *Bioinformatics, vol. 22, no. 8, 2006, pp. 1022-1023.*

[16] Jianxin Wang, Jun Ren, Min Li and Fang-Xiang ," Identification of Heirarchical an Overlapping Functional Modules in PPI Networks," *IEEE Transaction on Nanobioscience, vol. 11, no. 4, Dec 2012, pp.386-395.*

@IJAERD-2015, All rights Reserved 

[17] Binging Cai, Haiying Wang, Huiru Zheng and Hui Wang, Xiang ," Detection of Protein Complexes from Affinity Purification Mass Spectrometry Data,"*BMC System Biology, vol. 6, no. 4, Apr 2012, pp.1-10.*

[18] Min Li , Xuehong Wu, Jianxin Wang and Yi Pan ," Towards the Identification of Protein Complexes and Functional Modules by Integrating PPI Network and Gene Expression Data,"*BMC Bioinformatics, vol.13, no. 4, 2012, pp.1-15.*

[19] Tornow S and Mewes H W,"Functional Modules By Relating Protein Interaction Networks And Gene Expression", *Nucleic AcidsRes,vol. 31, 2003, pp. 6283–6280..*

[20] *HanD,BertinN,HaoT,etal.,"*Evidence For Dynamically Organized Modularity In The Yeast Protein-Protein Interaction Network*.", Nature 2004, pp. 430:88–93.*

[21] ] S. Zhang, H. W. Liu, X. M. Ning, and X. S. Zhang, "A hybrid graph-theoretic method for mining overlapping functional modules in large sparse protein interaction networks," *International Journal of Data Mining and Bioinformatics, vol. 3, no. 1, 2099, pp. 68–80.*

[22] M. Wu, X. Li, C. K. Kwoh, and S. K. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinformatics, vol. 10, article 169, 2009, pp. 1-5.*

[23] J. Reichardt and S. Bornholdt, "Detecting fuzzy community structures in complex networks with a potts model," *Physical Review Letters, vol. 93, no. 21, 2004, pp. 1-5.*

[24] S. Zhang, R. S. Wang and X. S. Zhang,"Identification of overlapping community structure in complex networks using fuzzy c-means clustering," *Physica A, vol. 374, no. 1, 2007, pp. 483– 485..*

[25] C. Wang, C. Ding, Q. Yang and S. R. Holbrook, "Consistent dissection of the protein interaction network by combining global and local metrics,"*Genome Biology,vol.8,no.12,, 2007, pp. 1-10.*

@IJAERD-2015, All rights Reserved 