FMFinder: A Functional Module Detector for PPI Networks

Manali Modi
Computer Engineering Department
Marwadi Education Foundation
Rajkot, Gujarat
manalim.21@gmail.com

Navjyotsinh Jadeja
Information Technology Department
Marwadi Education Foundation
Rajkot, Gujarat
noon2night88@gmail.com

Kirtirajsinh Zala
Marwadi Education Foundation
Rajkot, Gujarat
kirtirajzala@gmail.com

***Abstract*—Bioinformatics is an integrated area of data mining, statistics and computational biology. Protein-Protein Interaction (PPI) network is the most important biological process in living beings. In this network a protein module interacts with another module and so on, forming a large network of proteins. The same set of proteins which takes part in the organic courses of biological actions is detected through the Function Module Detection method. Clustering process when applied in PPI networks is made of proteins which are part of a larger communication network. As a result of this, we can define the limits for module detection as well as clarify the construction of a PPI network. For understating the bio-mechanism of various living beings, a detailed study of FMFinder detection by clustering process is called for.**

Keywords-functional modules; protein; PPI network; detection methods; inferring PPI network

I. INTRODUCTION

One of the major applications of Biotechnology is in the field of Bioinformatics especially when working with organic data. Analysis of biological processes is the primary objective of bioinformatics. The primary research revolves around hereditary connections, structural alignment of proteins, various protein to protein interaction methods and evaluation methods. Large part of bioinformatics is concerned with various biological processes which are part of Protein-Protein Interaction (PPI). PPI network, is a network of proteins interacting with each other to carry out various biological processes inside an organism. Hence it is very important to study and analyze how these protein modules interact to perform and carry out various metabolic activities. Systematical examination of the properties which are concerned with proteins which give concise depictions of consistent structures in wellbeing and diseases is known as Proteomics [1]. Normally, protein infrequently goes about as a solitary isolated component. Proteins, including those in the indistinguishable cell forms, regularly associate with one another to consolidate into an extensive atom to perform the organic capacities. For example, absorption framework, quality outpouring control, cell spread, cell signal transduction, the ways of action and movements of GSC and cell apoptosis depend on PPI. As needs be, the examination of PPI frameworks frequently serves as the reason to a better comprehension of cell affiliation, strategies, and limits and thusly elucidation of protein correspondence which is a central issue in science [1].

II. PPI MAJOR DATASETS

The postgenomic period is recognized by the accessibility of colossal measure of organic information sets which are truly heterogeneous in nature and hard to examine. Vast scale PPI network throughput such as tandem affinity purification and yeast two hybrid give associations stable as well as transient in nature [4]. PPI networks also have mass spectrometry showing the protein edifices [4]. These datasets, notwithstanding being inadequate, additionally comprise of false positives, and, thusly, the cooperation found in different information sets may not concur with one another. Owing to this difference, it is basic to make utilization of statistic techniques to induce the PPI arranges by discovering solid and reproducible connections and anticipate the associations not discovered yet in the accessible data.

III. INFERRING PPI NETWORK

This segment depicts the statistical techniques that are utilized to discover solid and complete protein-protein association systems. The inference of PPI systems can be done in sundry courses, for example, phylogenetic profiling and ID of basic examples. It is to be noticed that in contrast with quality systems, a great deal of work can be flawed in a protein-protein system surmising utilizing the probabilistic strategies. In a living life form, a few proteins cooperate to do different undertakings framing a protein complex. A large sum of Protein-Protein Interaction information compromises the interaction and it is extremely uncommon the discovering communications amongst large protein numbers. Subsequently, recognizable proof of protein buildings is of prime significance to pick up a superior comprehension of the cell system. Distinguishing protein edifices is a crucial zone of investigation of protein systems, for which different grouping techniques were connected. One of the different methods for recognizing

the protein buildings incorporate diagram division, where the chart is grouped into subgraphs utilizing expense based pursuit calculations. Another methodology is extensively ordered as protection crosswise over species, where arrangement devices are utilized to discover the edifices that are normal in numerous information sets originating from distinctive species. The conclusion of human genome sequencing in turn makes proteomic examination a standout amongst the most critical of life science processes. A systematical investigation procedure in which various properties concerned with are investigated. This investigation helps in representing the structure, capacity and the control of natural frameworks defining health and sickness [1]. At times proteins go about as single disengaged elements. On the other hand, proteins included in the same cell regularly associated with one another to join into an extensive particle. Case in point, the methods and exercises of the hereditary substance duplicate, quality declaration control, cell signal transduction, digestion system, cell proliferation are related to PPI. They are the foundation of organic courses of action occurring in life forms. Consequently, the examination of PPI systems commonly serves as the premise to a superior comprehension of cell association, courses of action, capacities and subsequently to the explanation of protein communication which is a focal issue in science [2]. Last decade saw PPI information dissected by high throughput test systems. Such as two hybrid frameworks, protein chip innovation and mass spectrometry. Established approach to focus protein capacity is to discover homologies between an unannotated protein and other protein utilizing grouping comparability calculation [3]. Numerous coordinated PPI systems have performed identification processes. The huge size of PPI system information becomes a tedious task to effectively distinguish various organic modules as well as essential examination subject in genomic time. There are various organic trial systems to identify functional modules in PPI systems.

A. Module Detection Survey [1]

The authors have dissected existing issue and present metrics for distance. Also they categorized the overall arrangement of practical module discovery as well as execution of numerous calculations by using known values. Final execution in the current scenario is the investigating of the network system. Essential idea driving the research is to distinguish utilitarian modules from existing network. By organizing and utilizing diverse bunching routines with distinctive calculation. In the current research different problems are introduced to assess the discovery of module and its quality and also counting location system's execution. Parameters are Sensitivity, Precision, F-Measure, Recall, Accuracy and Positive Predictive Value, and p-value measure. This research portrays procedures which recognize utilitarian module in comparison with current network.

B. DFM-CIN Algorithm [2]

The authors have presented a novel structure which identifies protein complexes. This paper also proposes practical approach by acclimatizing hereditary component statement information into datasets of PPI. DFM-CIN is the proposed method which calculates revelation of useful module in view of the distinguished edifices. Authors developed Protein Protein Interaction network as a part of static systems of TSNs. The proposed structure not only calculates but can also recognize protein complexes and modules. The research findings of the authors recommend functional modules identification with the help of protein complexes.

C. Protein Function Prediction Using ANN [3]

Authors have built a model by using weighted graph for protein interactions. This acts as a base to put forward and reflect facts related to small world network property. This property filters protein interaction network reliably. In certain situations, individual protein has multiple functions. This makes it an issue of multi labeling in a weighted graph problem. The procedure shows very high reliability amongst the connections of protein in the network. The suggested approach has been tested on MIPS datasets showing high performance in terms of precision and recall while using ANN.

D. MOFinder Algorithm for Overlapping Modules [4]

MOFinder for large PPI networks is proposed by the authors in [4]. PPI data file is primarily converted into sparse matrix in this approach. Then this sparse matrix is processed with global Approximate Minimum Degree Ordering as well as local AMD. Local sparse matrix and local AMD are generated using sliding window protocol along the diagonal. Clustering coefficient for this matrix is required to be calculated and if found higher than the cutoff then we save the sub modules. Else they are discarded. Finally sliding window goes diagonally to fetch remaining modules. Detection of small modules is not possible with this algorithm.

E. Overlapping Modules Mining Using LGT [5]

Authors proposed a method in which they have used Line Graph Transformation for discovering utilitarian modules from large network as well as accumulate ones identifying protein module structure. Resulting modules are identified with projected algorithm which shows high scope among fly, yeast, and worm network of proteins. Investigation on yeast protein networks recommends enormous protein modules which have been found with association of capacity annotation, localization as well as buildings blocks of proteins.

F. COACH Algorithm [6]

In this paper authors promoted COACH method for anticipating constructions with the help of recognizing protein complexes as well as containing protein associations. They measured and investigated protein groups from different viewpoints. From first view, they execute widespread correlation among projected method as well as present approaches with consideration of expected groups in contradiction of targeted protein structures. Second, they admit focal point linking edifices exploiting dissimilar biological evidence as well as learning.

G. Fuzzy Clustering Algorithm [7]

Authors have proposed a method in which they have secluded the capacity of Q fluffy cluster. This means they have

applied grouping strategy in order to discern the covering group structure. The research shows the higher efficiency of the new algorithm in detecting appropriate number of clusters and good clustering.

H. Functional Module Detection Using Metrices [8]

Authors have addressed various problems arising while developing protein protein connections and resultant issues of protein communication information. Hence, they advised for usage of betweenness commonality decomposition for calculating customs edge shared trait and for the recognizing of practical modules from the extensive networks.

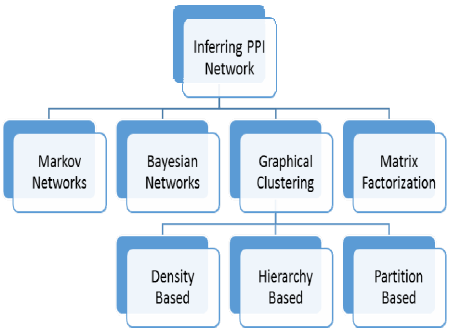

Fig. 1. Intferring PPI Network

IV. PROPOSED FRAMEWORK

All the above procedures devised some shortcomings with the reference of quantity of practical module recognition. Hence, FMFinder approach is designed for amending enactment of prevailing procedure. The steps for the proposed flow of the algorithm are shown in Figure 2. For the proposed FMFinder, PPI file in terms of protein network and clustering coefficient 0.45 are the inputs and in turn functional modules from the Human and Yeast database are the production of the protein networks. Figure 3 shows the proposed algorithm FMFinder for functional module detection.

V. IMPLEMENTATION AND COMPARISON WITH EXISTING TECHNIQUES

Human PPI Dataset and Yeast PPI Dataset were used to perform FMFinder algorithm based implementation. Interacting protein data is stored in these datasets. Human dataset was collected from Human Protein Release Dataset (HPRD). Yeast dataset was collected from Dataset of Interacting Proteins (DIP). 1800 protein interactions are included in the Yeast dataset and 39200 interactions of proteins are found in human PPI network. FMFinder algorithm processes on these datasets to detect functional modules. Analysis and comparison with earlier algorithms and approaches along with their limitations are described in TABLE I. Table II depicts modules identified with its major modules for human dataset. Table III depicts modules identified with its major modules for yeast dataset. From this review it is obvious that FMFinder 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 is valuable for the detection of protein modules. Despite the fact that LPCF has most anticipated proteins and covered proteins it has less utilitarian rate. Figure 5 shows modules that are predicted and covered proteins by the algorithm in Human Database. As shown in the Figure 5 and Table II FMfinder algorithm has highest predicted proteins and covered proteins as compared to other algorithms for human database. Figure 6 shows functional percentage of identifying modules from the large network by different algorithms in Yeast Database. It can easily seen from the Figure 6 and Table III that FMFinder shows highest functional percentage for detecting functional modules for Yeast database. Thus FMFinder has highest accuracy for discovering overlapping modules from PPI complex network.

1. First stage is to get PPI dataset as a data PPI record.

2. Then check whether data is suitable to continue further for the utilitarian module location.

3. Convert PPI organize into meager framework.

4. Apply Multiple Minimum Degree (MMD) calculation for the recognition of the utilitarian modules.

5. Finally, functional module identified by taking after every above step.

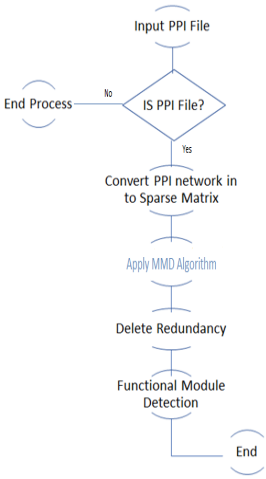

Fig. 2. Proposed Flow of FMFinder Algorithm

```
Input: 1. PPI file in form of Protein1, Protein2
       2. Clustering Coefficient (CC)
Output: Functional Module
Mat= zeros (P, P);
for each p ϵ P do
        Mat (p1, p2) = 1;
end for
FM = MMD (Mat);
if CC (FM) ≥ CC do
        Add FM to functional modules
End if
for each modules M1 in functional modules do
        If module M1 ϵ module M2
        discard module M1 from functional modules
    end if
end for
```

Fig. 3. Proposed Algorithm

MMD Algorithm:

1. Calculate degree of protein nodes.
2. Determine new minimum degree.
3. Mass elimination.
4. Determine new elimination graph after mass elimination.
5. Stop

Fig. 4. Flow of MMD Algorithm

TABLE I. ALGORITHMIC ANALYSIS

Algorithm	Description	Limitation
MCODE	Identifies densely connected group of proteins	Detects only connected graphs of proteins within the PPI network
DPClus	Based on the agglomerate and divisive algorithm	Unable to detect overlapping functional modules
MOFinder	Identifies functional modules, specially overlapping modules	Detects only small size of modules, less than 12, from human and yeast database
FMFinder	Identifies functional modules	Detects 256 modules from human database and 109 modules from yeast database. Major size module is 4.

TABLE II. FUNCTIONAL MODULES IDENTIFICATION ON HUMAN DATABASE

Algorithm	Modules Identified	Major Module Size
MCODE	21	3
DPClus	102	8
MOFinder	221	12
FMFinder	265	4

TABLE III. FUNCTIONAL MODULES IDENTIFICATION ON YEAST DATABASE

Algorithm	Modules Identified	Major Module Size
MCODE	15	3
DPClus	54	8
MOFinder	90	3
FMFinder	109	3

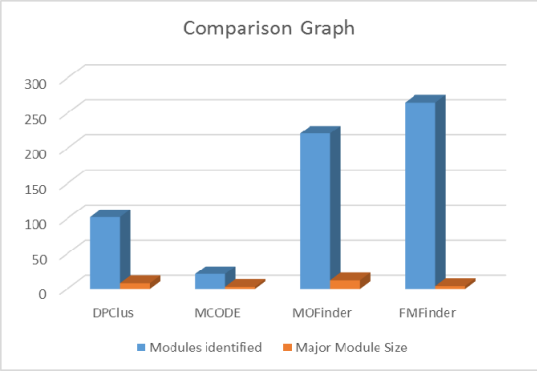

Fig. 5. Comparision Graph for Human Database

VI. CONCLUSION

The current paper portrays different methods which are exploited as a part of reviling functional modules from a large database. We examined each dataset, inferred PPI network, analysis of existing algorithms and its covered proteins and functional percentage of modules which are actually identified from the existing algorithms. Results show that the FMFinder proposed algorithm outperforms previous algorithms (MCODE, DPClus, MOFinder) by detecting more modules of proteins in both human and yeast databases.

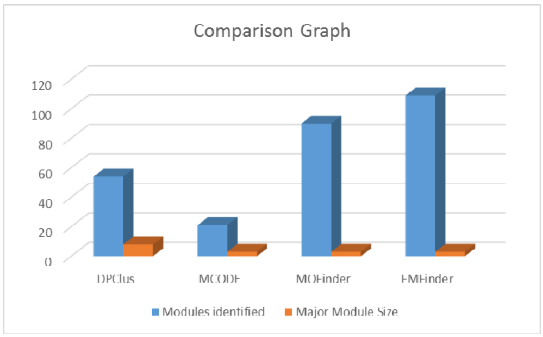

Fig. 6. Comparision Graph for Yeast Database

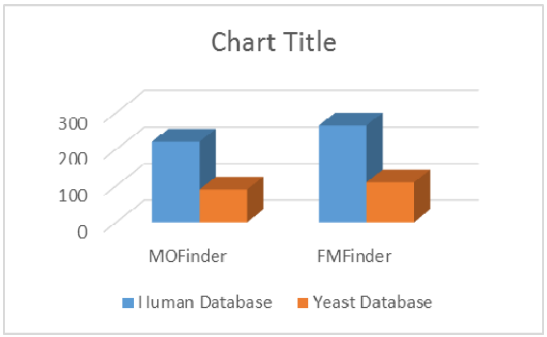

Fig. 7. Graph Comparison between FMFinder and MOFinder

REFERENCES

[1] J. Ji, A. Zhang, C. Liu, X. Quan, Z. Liu, "Survey: Functional Module Detection from Protein-Protein Interaction Networks", IEEE Transaction on Knowledge and Data Engineering, Vol. 26, No. 2, pp. 261-273, 2014

[2] M. Li, X. Wu, J. Wang, Y. Pan, "Towards the Identification of Protein Complexes and Functional Modules by Integrating PPI Network and Gene Expression Data", BCM Bioinformatics, pp. 1-12, 2012

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

[4] Q. Yu, G. H. Li, J. F. Huang, "MOfinder: A Novel Algorithm for Detecting Overlapping Modules from Protein-Protein Interaction Network", Journal of Biomedicine and Biotechnology, Vol. 2012, pp. 1-10, 2012

[5]] S. Zhang, H. W. Liu, X. M. Ning, 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, pp. 68–84, 2009

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

[7] S. Zhang, R. S. Wang, X. S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c-means clustering", Physica A, Vol. 374, No. 1, pp. 483– 4490, 2007

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