ISSN No:-2456-2165

Take on Markov Cluster Detection Algorithm (MCL) over Interactomics Based Protein-Protein Interaction (PPI) Network

Munazza Ahsan Shaikh
Liaquat University of Medical
& Health Sciences Jamshoro,
Sindh, Pakistan

Murtaza Hussain Shaikh
Kyungsung University,
Busan, Korea

Mahreen Jurial Khan
University of Sindh,
Jamshoro, Sindh, Pakistan

Abstract:- The determination of this article is to mature a new protein module detection approach that tries to address the shortcoming of graph-based protein functional module detection algorithms and leverage their biological significance. In our study, we designed a migration strategy that enables proteins to migrate between clusters to finally get grouped with biologically similar proteins. We have strained to progress an enhancement method to help better filter, or precisely reorganize, the outcomes of pre-existing graph-based functional module detection algorithms to be more biologically significant. The Markov Cluster (MCL) detection algorithm technique was adopted as it fits well with the migration principle of the Interactomics based protein network. Also, it is ideal to describe the inherent uncertainty of biological linkages. Besides, spectral clustering was used to get better precision in measuring the distances in the network and to cope with the high-data dimensionality. A study was performed on these techniques to understand their advantages and limitations to define some metrics that take into consideration the biological and topological characteristics of proteins, to adopt the MCL algorithm means and spectral clustering techniques to protein networks context. The statistical tests were positive and with this work, we tried to increase the effects of a widely used graph-based algorithm respectively.

***Keywords**:- Algorithm, Cluster, Interaction, Module, Measurement, Network, Protein, Technique.*

I. INTRODUCTION

Recently, due to the substantial improvement in the field of proteomics, a great interest has been given to Interactomics, that is, the study of protein-protein interactions (PPI), or more generally communications between the cell's macromolecules. Thus, many researchers have tried to improve, manipulate and predict protein-protein interaction model. Indeed, protein complexes are the primary molecular clusters of proteins that work together to accomplish a biological function; we may expect protein complexes to be functionally and physically unified assemblies in the PPI network [1]. The protein complexes can be seen as super-molecular structures that integrate the products of several genes that carry out some interrelated functions. Generally, protein complexes are formed because each participating protein molecule can have multiple binding sites as determined in figure1. Protein complex can perform a range of functions; they may be a multi-enzyme complex that catalyzes a chain of biochemical reactions or a compound of proteins that are participants in a signal transduction pathway [2]. Studies such as [2], [5], and [6] showed the existence of cooperation and interaction between different proteins within a complex. It has also been perceiving that some proteins can be simultaneously elaborate in the foundation of several compounds.

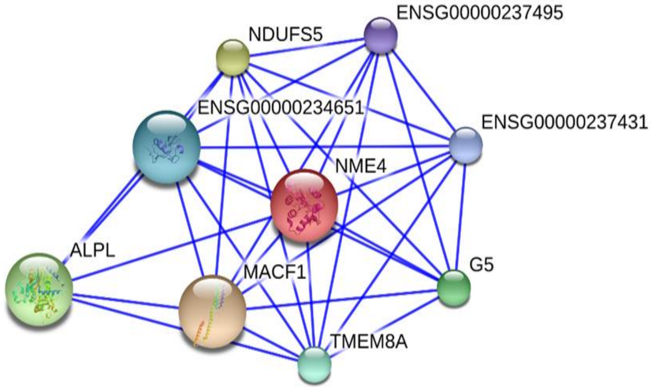

Fig 1:- Macromolecular structure of proteins

Before diving into the methods and models used in detecting protein complexes, we require making a vibrant discrepancy between protein complexes and functional modules. The protein complexes are assortments of proteins that interrelate with each other at equivalent interval and place, establishing a particular multi-molecular mechanism [4]. Entirely the opposite, functional modules comprise of proteins that participate in a specific cellular process while requisite each other at a distinct time and place (unlike conditions or segments of the cell cycle, in single cellular compartment etc.). However, in the literature, many protein complexes and functional module detection techniques do not draw a line between them, because generally obtained interactions do not provide temporal and unique information [3]. Many models have been intended to internment the physiognomies of PPI networks to learn them superior. Typically, the PPI network is exhibited as an undirected graph where vertices are proteins, and edges represent interactions [5].

II. BACKGROUND STUDY

Recent developments in great quantity experimental methods for identification of protein interactions have stemmed in a significant amount of diverse data. However, as valuable as they are, such innovative approaches to studying protein Interactomics have specific boundaries that can be supplemented by the computational techniques [6], [7]. To automatically manipulate this data, researchers have first tried to define some models that can be used to represent protein interaction networks. The simplest and the widely used is the undirected graph, although further distinguished models use labeled edges as shown in figure2.

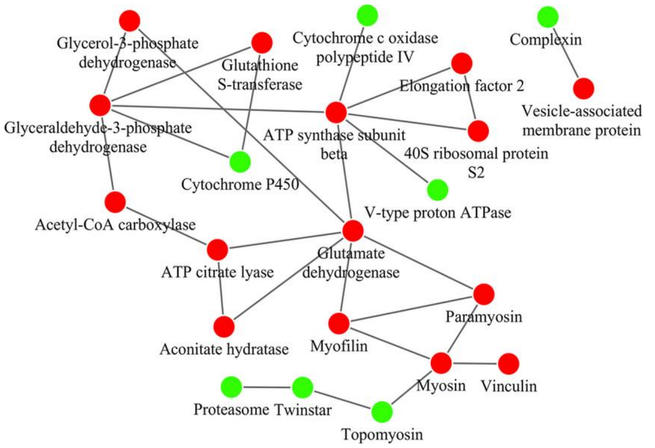

Fig 2:- Labelled Edge Graph of Protein Interactomics

These graph models enable the investigation of biological properties trough graph methodology. Therefore, some research work has been completed, especially on the structure and measurements of protein interactions, in a hope to get some insights into robust and fragile properties of the protein networks and permit scientists to select protein targets for therapeutic intervention [6]. Based on those findings, a range of protein complex and functional module detection algorithms have been developed to help to enhance efforts to reconstruct the genetic basis of the syndrome and determine novel drug targets [9]. Other emerging approaches in protein complex prediction try to determine the formation of complexes by integrating topological information with other protein structural data. Many aspects have been addressed using the protein interaction network, for example in [8] it has proposed a virus classification method based on a virus-host protein-protein interaction network. Various research groups also applied network analysis find gene data sets associated with cancer [10].

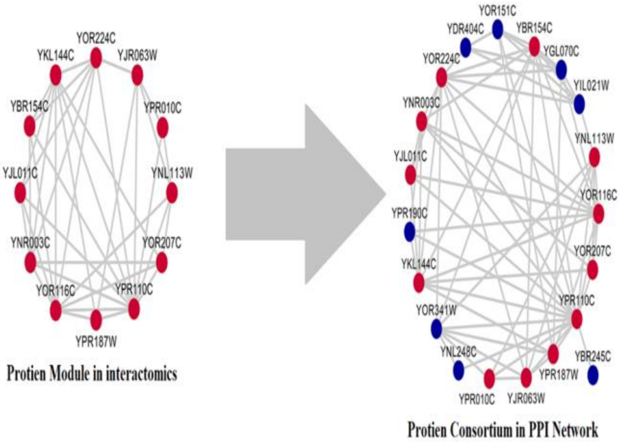

Fig 3:- Protein Ontological Consortium in PP Interaction Network

As presented in figure2.1, efforts have also been made recently by the Protein Ontology (PO) consortium to define a formal ontological structure to describe types of protein complexes and give these types unique permanent identifiers. The understanding of the topological and the architectural doctrines of a biological network can thoroughly provide an insight into various network characteristics [11], [12]. The extraction of global properties aimed to represent PPI networks by using models used in other fields, such as random graphs or scale-free networks. To date, the introduction of a comprehensive PPI network model is still an open research problem, and the assumptions and different type of connectivity structure are yet to be discovered [12]. Furthermore, the incompleteness in PPI networks is also one of the primary limits to designing a perfect model.

III. INTERACTOMICS DETECTION TECHNIQUE IN PROTEIN NETWORK

To these days, a significant amount of cooperative effort is made between scientists from different disciplines is being devoted to integrating the available biological data to attain at entire precise and comprehensive networks possible. Nevertheless, even the best-sophisticated exertions are still only beginning to preview protein interactions in detail. Owing to the rapid growth of the Interactomics detection techniques, numerous computational schemes have been established to support biologist get the essence of these data [13]. Hence, in the last decade, we have realized the emergence of a wide variety of protein complex detection methods.

Moreover, one of the primarily used sets of protein complex detection method is graph-based methods. If we rate these set of algorithms regarding space, time and accuracy, we can notice that they perform well regarding time and space but stumble regarding efficiency [12], [13], and [14]. Generally, their unsupervised manner of detecting protein complexes, and they usually tend to ignore the multi-functionality of proteins and the manifestation of false positives and –negatives inside the network [15]. In this article, we established an additional method that takes advantage of the graph-based techniques, in a way that they can express us about the topology of the network, and cracks to leverage the biological accuracy of their results by using a flexible fuzzy spectral (fuzzySpect) technique shared with data from the Gene Ontology (GO).

IV. MCL ALGORITHM INTEGRATION WITH PROTEIN NETWORK

The Markov Cluster algorithm (MCL), finds clusters in the graph by simulating random strides. The algorithm iteratively weakens the flow where it is weak and increases the flow where it is strong [2]. This process will separate the PPI into many segments as protein clusters. Studies showed that MCL outperforms many graph-based multiple detection methods [10]. Several complex validation methods have been proposed to assert the validity of predicted complexes. As different results can be generated by a different method or even with the same method with various parameters, it is of vital importance to improve techniques to scrutinize the accuracy of the predictions. Generally, quality of clusters is estimated the regarding homogeneity and separation based on the definition that element within a cluster are highly connected between them and sparsely combined with outside elements [7]. Usually, clustering results are compared with ground truth derived from various sources such as the famous MIPS (Munich Information Center for Protein Sequences) and SCOP (Structural Classification of Protein) [15]. Some methods check how much their predicted complexes (Pc) match with known complexes (Kc). An overlapping score OS that measures the portion of matched proteins to the size of the Pc and Kc is calculated using the following formula;

$$OS(Pc, Kc) = \frac{|Pc \cap Kc|^2}{|Pc|.|Kc|} \text{ (4-1)}$$

The higher the value of OS is better. Beside the OS value, the amount of true positive (sensitivity) the portion and false negatives (precision) are also calculated using the following formats;

$$Recall = \frac{|Pc \cap Kc|}{|Kc|} \text{ (4-2)} Precision = \frac{|Pc \cap Kc|}{|Pc|} \text{(4-3)}$$

The accuracy of a module can be associated with the harmonic mean (i.e. f-measure) between recall & precision by calculating the following portion;

$$f - measure = \frac{2\,(Recall\,.Precision)}{Recall + Precision} \text{ (4-4)}$$

Another set of methods verify the validity of detected by calculating the functional homogeneity of the proteins of a cluster *C* to an annotated function *F* by expending the *p-value* from the hyper-geometric distribution as follow;

$$Pvalue = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i}\binom{|V|-|F|}{|C|-i}}{\binom{|V|}{|C|}} \text{ (4-5)}$$

V. EFFECTS OF APPLYING MCL TECHNIQUE

When testing with MCL, we witness that overall modules size has increased significantly (Table.5-1). In fact, because MCL generates significant size clusters, the proteins that don’t have a Gene Ontology annotation migrated to clusters with which they interact more and that according to the defined equation (4-2) mentioned above.

Algorithm	Clusters (*Size* ≥ *2*)	Avg. Module size	Perfect Match	Sn	Sp	f-measure
MCL *(cutoff=0.2)*	50/50	24.94	1	0.023	0.184	0.04
FuzzySpect	50/50	71.68	2	0.028	0.228	0.05

Table 1:- Results with the 408 yeast known proteins (MCL)

Nevertheless, proteins with Gene Ontology annotation will try to balance between the biological significance and the topological influence again according to equation (4-2).

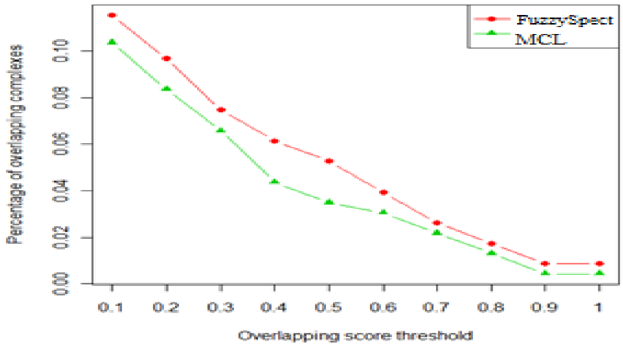

Fig 4:- A Score of FuzzySpect over MCL algorithm

Notice that our method has done a nice tradeoff and see that number of exact match increased and the recall, precision and f-measure have been improved by sternly follow the equation (4-2) and (4-1).Table.5-2 summarizes the Gene Ontology functional enrichment results; the results show that we got a nice biological improvement, with a small decrease in the M.F (f-measure value).

Algorithm	*f-measure*			*Log(p-value)*		
	B.P	*M.F*	*C.C*	*B.P*	*M.F*	*C.C*
MCL (cutoff=0.2)	0.247	0.226	0.34	8.05	6.53	9.6
FuzzySpect	0.79	0.213	0.33	8.88	7.04	10.2

Table 2:- Gene Ontology function analysis (MCL)

Observing the graph plot in figure 5, note that the overlapping ratio has increased in the two extremes (X and Y axis). In one hand, the number of height rate overlapping modules increased, but in the other hand, the ratio of weakly overlapping modules increased because some significant modules increased in their size by "eating" the neighbouring non-enriched proteins.

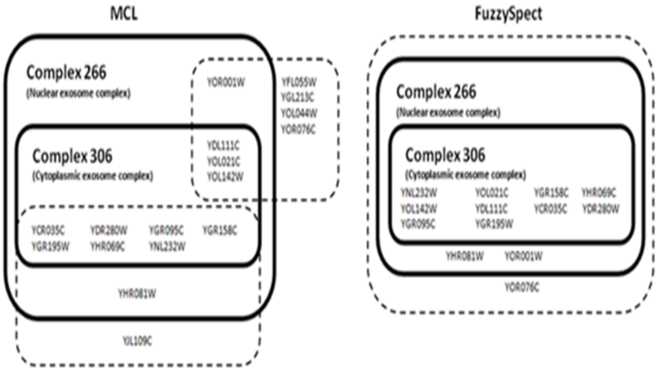

Fig 5:- Illustration of migration of proteins from different cluster

Figure 5 illustrates the process of migration of proteins from a different cluster. As MCL grouped the proteins of the known complexes 266 and 306 in different groups. However, after applying our method proteins migrated from one group to the other to hold both of them.

VI. CONCLUSION

Proteins are essential molecules in our cell, understanding the behaviour of proteins and their interactions help us shed unprecedented nimble on the inner working of the cellular machinery [1], [14]. Henceforth, many biologists in collaboration with different computer system scientists have tried to develop modern methods and techniques to enable accurately conclude the massive amount of biological data for protein interactivity. The work performed in this article is just a minor part of many states of the art researchers that have been conducted to predict protein complex structures and predict the function of proteins. To validate the results, we selected the MCL algorithm based on the size of the clusters they reproduce. The reason to choose MCL is; first, to fix the degree of overlapping for a protein we got the list of known overlapping proteins and checked their overlapping degree.

Second, we notice that the majority of them only overlap with two clusters [15]. Thus we fixed the maximum number of overlap to be two.

VII. FURTHER STUDY

After running our algorithm on the selected algorithms, the results showed that a biological improvement was achieved. Also, it showed that the accuracy of the results contingent upon the initially used algorithm as some algorithms (i.e. DPClus) tends to exclude some proteins which mislead the results obtained after applying the MCL algorithm. Leading this work, we found ideas points that could be useful for further research work, especially in the PPI network. Firstly, to improve an accurate tool to decide about the accuracy, some statistical tests (i.e. p-value & f-measure) even though they have proven their efficiency but still not perfect to have a firm decision about the results. Another validity issue is about Gene Ontology as its database is still incomplete. Secondly, for PPI infer complexity, recently many reliable data sets have been collected which can be used as background to reliably filter and predict new protein complexes and functional modules in data analysis.

ACKNOWLEDGEMENT

The authors would like to thank and acknowledge the RCSB PDB (Protein Databank), National Science Foundation, the National Institutes of Health Bethesda, Maryland USA, for using the protein database to study and conduct statistical analysis on the database.

REFERENCES

[1]. Collins MO, Choudhary JS (2008) Mapping multiprotein complexes by affinity purification and mass spectrometry. Curr Opin Biotechnol 19: 324–330.

[2]. Gully D, Bouveret E (2006) A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in Escherichia coli. Proteomics 6: 282–293.

[3]. Hosp, F., et al. (2015) “Quantitative interaction proteomics of neurodegenerative disease proteins,” Cell Reports, 11(7) (pp.1134–46), doi: 10.1016/j.celrep.2015.04.030.

[4]. Arifuzzaman M, et al. (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. Genome Res 16: 686–691.

[5]. Ewing RM, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 3: 89.

[6]. Suter B, Kittanakom S, Stagljar I (2008) Two-hybrid technologies in proteomics research. Curr Opin Biotechnol 19: 316–323.

[7]. Wang X, Huang L (2008) Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. Mol Cell Proteomics 7: 46–57.

[8]. González-Couto E. Functional and systems biology approaches to Huntington’s disease, Brief Funct Genomics, 2011, vol. 10 (pg. 109-14)

[9]. Bui QC, Katrenko S, Sloot PMA. A hybrid approach to extract protein-protein interactions, Bioinformatics, 2011, vol. 27 (pg. 259-65).

[10]. Tyler AL, Asselbergs FW, Williams SM, et al. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy, BioEssays, 2009, vol. 31 (pg. 220-7)

[11]. Albert R, Jeong H, Barabási AL. Error and attack tolerance of complex networks, Nature, 2000, vol. 406 (pg. 378-81).

[12]. Wand AJ, Englander SW (August 1996). "Protein complexes studied by NMR spectroscopy". Current Opinion in Biotechnology. 7 (4): 403–8. Doi: 10.1016/s0958-1669(96)80115-7. PMC 3442359 Freely accessible. PMID 8768898.

[13]. De Domenico M, Nicosia V, Arenas A, Latora V (April 2015). "Structural reducibility of multilayer networks". Nature Communications. 6: 6864. Doi: 10.1038/ncomms7864. PMID 25904309.

[14]. Jaiswal, Amit (2014). "AtTRB1–3 Mediates Structural Changes in AtPOT1b to hold ssDNA". ISRN Structural Biology. 2014: 1–16. Doi:10.1155/2014/827201.

[15]. Bruce A, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002). Molecular-biology of the cell (4th Ed.). New York: Garland Science. ISBN 0-8153-3218-1, USA.