A Review on Identification of Protein Complexes in PPI

[1]Greeshma Sara Thomas, [2]Meenu Mathew
[1]PG Student, [2]Assistant Professor
[1]Computer Science and Engineering,
[1]Rajagiri School of Engineering and Technology, Kochi, India

Abstract : Proteins are primary component of cell machinery and life, which are extremely important for cellular function and disease processes. Physical contacts of high specificity is established between the protein molecules due to biochemical process. From the protein protein interaction (PPI) network, various protein complexes are identified which is very important and a challenging task in computational biology for better understanding of cellular engineering. Naturally, PPI data can be depicted in the form of networks, hence it can be represented as an undirected graph of which node and edge are represented as protein and physical interaction between two protein nodes respectively. Review on various algorithms for the protein complex Identification is been discussed.

IndexTerms **- Interaction network, Protein complex, Protein Protein Interaction(PPI) .**

I. INTRODUCTION

Proteins plays an important role in controlling all biological systems in a cell. Several proteins perform their functions differently and the vast majority of proteins interact with others to perform proper biological activity. Protein protein interactions are based on the methods like co-immuno precipitation (co-IP), pull-down assays, cross-linking, label transfer, and farwestern blot analysis. Proteins are the workhorses, responsible for the biological events taking place in a cell. It includes cell growth, proliferation, gene expression, nutrient uptake, motility and inter- cellular communication.

Protein expression is a dynamic process since the cells respond to limitless stimuli and in order to finish specific tasks proteins may not be always expressed or activated. Each cell is different to one another and thus expressed in a type dependent manner. The features of proteins imply a complexity that can be difficult to find, especially when trying to know the protein function in the proper biological process. Protein complex is in the form of quaternary structure made of more than two connected polypeptide chains. Different polypeptide chains constitute different functions. Proteins are attached by non-covalent interactions in protein complexes, and different protein complexes show different degrees of steadiness over time.

PPIs where earlier found out by experimentally through Tandem affinity purification which is also known as tap tagging. It is a technology that is used for the identification of interactions and it involves creating fusion proteins with a designed piece the tap tag on the other end. Through several process like affinity selection and washing a native elution is produced, which consists of new proteins and its interacting partners which can be found out by mass spectrometry.

In this survey, important data processing methods implemented till date for the identification of protein complexes from the networks is been reviewed. Most methods rely on the presumption that protein complexes are embedded as closely connected proteins within the PPI network, and these methods differ in their algorithmic approaches and further biological data is employed to find the complexes. For the survey 7 different algorithms are used and they are MCL, CMC, MCode, ClusterOne, CORE, COACH and WPNCA.

MCL is graph clustering method which is fast and extremely scalable. Initially applied to cluster protein sequences and has proved efficient for clustering large PPI networks due to its scalability. CMC works by continuous merging of maximal cliques evoked from the PPI network. MCODE uses an agglomerative approach for prediction of complexes. ClusterONE works same way as MCODE using seeding and greedy neighborhood expansion. CORE and COACH search for clusters that cling to the core-attachment arrangement specially in yeast complexes. WPNCA is used to find the protein complexes by weighted PageRank-Nibble algorithm and core-attachment structure.

First section is about the introduction of Protein complexes identification and Motivation of the same is explained. Second section gives the background knowledge of PPI and finally, it gives an overview of different algorithms used for the identification of Protein complexes.

II. BACKGROUND KNOWLEDGE

In protein complex, physical interactions are the fundamental elements for the cell functioning and acts as a preprocessing source of salutary targets. Several challenges related to targeting PPIs in particular with small molecules are noticeable and a growing number of functional PPI modulators is being noted and clinically evaluated. Essential proteins are those proteins which are important for the organisms growth.

The deletion of only one of these proteins is sufficient to cause lethality or infertility. Thus, for human pathogens primary proteins treated as potential drug targets. Thus for the identification of conserved essential genes required for the growth of fungal pathogens, it provides a classic strategy for enlightening novel anti-fungal drug targets. Therefore, finding essential proteins is essential not only for the understanding of the minimal needs for cellular life, but also for several other purposes such as drug design. Figure 2.1 represent the PPI network [10].

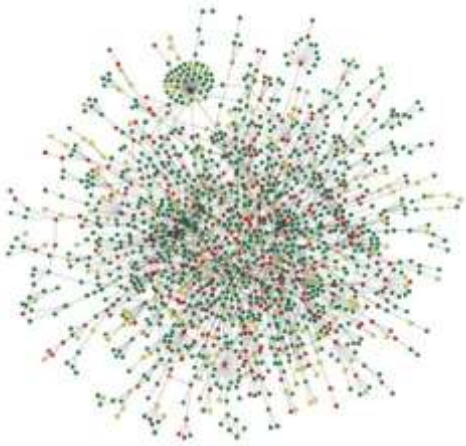

Figure 2.1 Example of PPI network

Different experimental procedures like single gene knock- outs, RNA interference, and conditional knockouts, have been adapted for the prediction and also for the analysis of proteins. However, these experimental techniques are generally laborious and time consuming. Considering the experimental restrain, a highly precise computational approach for identifying essential proteins would be of great value. With the advancement of technology like yeast two hybrid, tandem affinity purification and mass spectrometry, a wealth of protein-protein interaction (PPI) data have been developed, which helped for further learning of genomics and proteomics in network level. Figure 2.2 represents the protein complexes that are identified from PPI Network.

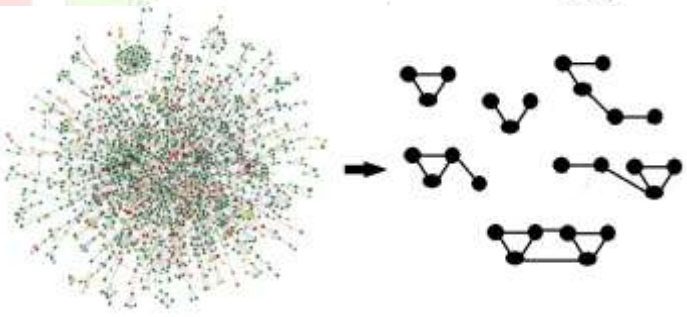

Fig. 2.2 Identification of protein complexes from PPI network

III. ALGORITHMS

3.1 MCL

MCL Girvan and Newman propose G-N algorithm [1], with the maximal edge betweenness breaking up of PPI networks is done by removing the edges iteratively. In graphs, the Markov Clustering (MCL) algorithm finds cluster structure by an arithmetical bootstrapping process. The flow of network is the proximate division of the interaction network and into each node an initial amount

of flow is injected. At each step, a portion of flow goes from one node to its adjacent node through the outgoing edges. Using two operators random walk is calculated using sequence similarity graph. It does using a stochastic matrices called as Markov matrices .These Matrices used to find the mathematical meaning of random walks on a graph. Within a graph the algorithm reproduces random walks using two operators, they are:

1. Expansion
2. Inflation

Expansion is calculated using the normal matrix product by taking the power of a stochastic matrix. Inflation is computed by taking powers entry wise. Then followed by a scaling step that checks if stochastic in the resulting matrix, i.e. the matrix elements (on each column) leads to probability values.

Thus MCL algorithm scales well with increasing graph size, can easily work with both weighted and unweighted graphs, produces good clustering results and robust against noise in graph data. However, it is found that there are many multi- functional proteins which involve different function modules and also the complexes are highly overlapped which makes it difficult for MCL.

3.2 CMC

Chua[2], proposes an algorithm named clustering-based on maximal cliques CMC for the detection of dense subgraphs from PPI networks by maximal cliques. The algorithm mainly has three phases:

1. Calculating all the maximal cliques
2. Rank the cliques with the weighted density
3. Combine or removes highly overlapped cliques

Cliques algorithm suggested by Tomita et al. to find maximal cliques is used in CMC. The Cliques algorithm makes use of depth-first search approach to calculate all maximal cliques, and during the enumeration process it can completely reduce non-maximal cliques. Next, CMC assigns and in descending order of their score cliques are ranked. From a PPI network many maximal cliques could be generated and it is found that few of them overlap with one another. If highly overlapped cliques are present, then it should be removed by reducing the result size. It is desirable to form bigger dense subgraphs by adding highly overlapped cliques. Through the inter connectivity of two cliques it checks whether both overlapped cliques that is found should be combined together or not.

It is found that CMC has the highest recall compared to MCode ,MCL CFinder algorithms, but when it reaches its utmost recall the precision is found lower than CFinder and MCL . The average co-localization score of the clusters created by CMC is higher than those created by CFinder and MCL and also there is possibility that few of the unmatched clusters developed are found to be unknown complexes in CMC. The analysis indicates that, it is difficult to discover complexes with low density when compared with MCL, which is expected to be able to discover clusters.

3.3 MCODE

G. Bader and C. Hogue[3],proposes Molecular Complex Detection MCODE algorithm which creates a large networks tractable by taking the dense regions around a protein of interest. The three stages of the algorithm are

1. Vertex weighting
2. Complex prediction
3. Optionally postprocessing to filter or add proteins in the predicted complexes

In vertex weighting it first weights every node based on its local neighborhood densities by the highest k-core of the vertex neighborhood. In the second stage, vertex weighted graph is considered as the input and from the seed vertex it seeds a complex with the largest weighted vertex and repeatedly checks whose weight is above a given threshold it moves outward including vertices in the complex .When a vertex is comprised, its neighbors are repeatedly reviewed in the same manner to see if they are part of the complex. Since complexes cannot overlap in this stage of the algorithm the vertexes are not checked more than once. When no more vertices can be inserted to the complex regarding the given threshold the process stops and is repeated for the next largest not seen weighted vertex in the network. Therefore in the network the densest regions are identified. In the final phase complexes are filtered if they do not contain at least 2-core.

The algorithm may be run with the fluff option .According to a given fluff parameter between 0.0 and 1.0 the size of the complex is increased. For every vertex in the complex, v, its neighbors are added to the complex if they have not yet been seen and if the neighborhood density (including v) is higher than the given fluff parameter.

Vertices that are added by the fluff parameter are not marked as seen, so there can be overlap among predicted complexes with the fluff parameter set. Using the haircut option if the algorithm is executed, complexes identified will be of 2-cored, thereby deleting the vertices that are connected singly to the core complex. If both options are specified, fluff is run first, then haircut.

It is found that this approach to analyzing PPI networks performs using less qualitative information implies that large amounts of available knowledge is buried in huge protein interaction networks and for the better results it considers researching differently, possibly adaptive, vertex scoring functions are taken into account.

3.4 ClusterONE

T. Nepusz, H. Yu, and A. Paccanaro [4], proposes the algorithm to find groups in a protein-protein interaction network that are likely to correspond to protein complexes. It works on a greedy growth process and uses the idea of the cohesiveness score . ClusterONE consists of mainly three major steps.

1. The overlap scores for each pair of cluster is calculated
2. Overlap graph is constructed.
3. Complex candidates less than a threshold are removed

Cluster that are associated to each other are then combined with the candidates of protein complex . And also if it find a group that has no connection to other groups then it is promoted without any additional merging to a complex candidate. In the third and final step of the algorithm, complex candidates that contain less than three proteins or whose density is below a given threshold are discarded. Based on co- localization and over representation scores of ClusterONE and MCL, it is found that ClusterONE complexes displays higher scores on almost all data less than the rest. However it is also found that the inherent architecture of protein is not concerned.

3.5 CORE

H. Leung, Q. Xiang, S. Yiu, and F. Chin[21], proposes the CORE[7] algorithm .The key idea behind this approach consists of three main steps:

1. Predict core components
2. Identify attachments for the cores and eliminate
3. Insignificant cores are computed and rank the significance of predicted complexes

Basically, for each disjoint potential core, a p-score is defined. Then, for each identified core combine all proteins that have interactions with the most of core proteins in the complex. At the same time, some of insignificant core candidates will be removed based on the assumption that only one core is assumed for each complex. Finally, it computes a measure to evaluate the significance of each predicted complex.

3.6 COACH

M. Wu, X. Li, C. Kwoh, and S. Ng [6], introduce a core- attachment method for the detection of complexes and is implemented using two stages.

1. Detection of cores from the neighbourhood graphs of vertices based on local density
2. Generation of protein complexes by including attachments into the cores

It is pointed out that COACH achieves better prediction performance than other methods which fail to consider intrinsic structure of protein complexes. However, there are no uniform or standard definition for the core and attachments of protein complexes, which arouses our substantial interests in developing method with respect to the core-attachment structure. The advantage of using this algorithms is that it is easy to detect the overlapping protein complexes. However, it fails to consider the complete information of the whole PPI network.

3.7 WPNCA

Weighted PageRank-Nibble algorithm [7] is a novel method to find protein complex from the interaction networks. It works according to edge-clustering coefficients of the edges connecting each protein nodes by assigning adjacent nodes with different probability. It mainly consists of three steps:

1. Dividing the interaction networks into several dense clusters
2. Finding the cores from these clusters
3. Selecting rest of the proteins in the clusters as attachments to form the final predicted complex

PageRank-Nibble algorithm assigns values to the neighbors of a node by dividing the value of the node evenly among its neighbors. In order find good clusters, the node should get higher values with the neighbors that tend to develop clusters. Based on the facts mentioned above, according to their topological features weight is increased on edges and different probability is assigned to the neighbors. This algorithm is found out to produce higher accuracy in terms of the prediction than the existing methods in case of F-measure .

IV.CONCLUSION

Protein complexes are prior molecular entities to perform cellular functions. Protein complexes from PPI networks is been detected by the growth of PPI data. Thus tracing the functional and evolutionary conservation of complexes leads to the further opportunities to study complexes under different contexts and across species. Protein cores are biological hearts of protein complexes so finding these cores is necessary. Multiple algorithms for detection of protein complexes in PPI networks proposed in recent years have own strengths and disadvantages. The review will give brief description of several algorithms which will be helpful for the further advancement of research related to this area.

REFERENCES

[1] S. van Dongen .Graph clustering by flow simulation, PhD dissertation,Univ. Utrecht, Utrecht, The Netherlands, 2000.
[2] G. Liu, L. Wong, and H. Chua ,Complex discovery from weighted PPI networks,Bioinformatics,2009.
[3] G. Bader and C. Hogue . An automated method for finding molecular complexes in large protein interaction networks,BMC Bioinformat., 2003.
[4] T. Nepusz, H. Yu, and A. Paccanaro, Detecting overlapping protein complexes in protein-protein interaction networks,Nature Methods,2012.
[5] H. Leung, Q. Xiang, S. Yiu, and F. Chin, Predicting protein complexes from PPI data: A core-attachment approach,J. Comput.Biol.,2009.
[6] M. Wu, X. Li, C. Kwoh, and S. Ng , A core-attachment based method to detect protein complexes in PPI networks,BMC Bioinformat.,2009.
[7] Wei Peng, Jianxin Wang, Bihai Zhao, and Lusheng Wang, Identification of Protein Complexes Using Weighted PageRank-Nibble Algorithm and Core -Attachment Structure,Ieee/Acm Transactions on Computational Biology And Bioinformatics,2015.
[8] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney ,Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters,Internet Math.,2009.
[9] X. Tang, J. Wang, J. Zhong, and Y. Pan, Predicting essential proteins based on weighted degree centrality,IEEE/ACM Trans.Comput. Biol. Bioinformat.,2014.
[10] Albert-Laszlo Barabasi and Zoltn N Oltvai ,Network Biology: Under- standing The Cell's Functional Organization, Nature Reviews Genet- ics,2004