IJCTA, 9(10), 2016, pp. 4269-4276

Exploring Highly Connected Protein Complexes In PPI Networks

***Alannamathew *Reshmiraj *Sreeja Ashok *M. Vjudy**

Abstract : Exploring protein function is one of the most exigent problems of the post genomic era. The goal of clustering in Protein-Protein Interaction networks is to explore highly connected protein complexes. The identification of such highly connected complexes helps in the prediction of uncharacterized protein functions through comparison between the interactions of similar known proteins. The exponential intensification in the need for an accurate and efficient graph clustering methods has resulted in the development of diverse clustering algorithms. In this paper an efficient clustering solution is introduced for automatic detection of clusters which incorporates two methods mainly the redundancy reduction and Monte Carlo optimization in PPI network. This method results in finding highly connected sub graphs rather than fully-connected sub graph with maximum optimization.

Keywords : Clustering, Protein-protein interaction, Monte- Carlo optimization, redundancies in PPI

1. INTRODUCTION

In a cell the main functions are done by the interactions between proteins. A protein is a polypeptide chain of amino acids. The intentional physical contact established between two or more proteins due to some biochemical activities is known as PPI.PPI plays a vital role in the functioning of our body and is considered as the core of the inter-atomics of living cells. Biological processes like immunity, metabolism, gene expression etc are mediated through protein interaction. Any abnormal protein-protein interactions in living beings can result in many human diseases like cancer, Alzheimer's etc. Various studies show that protein rarely acts alone and the proteins taking part in the same cellular process has a tendency to interact with each other. By comparing the interaction of similar known proteins the functions of an uncharacterized protein can be inferred.

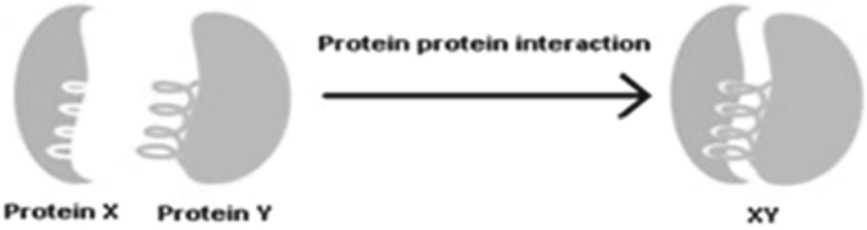

Fig. 1. PPI formation.

Figure 1 shows the formation of a PPI network, where proteins X and protein Y establishes an intentional physical contact with each other and forms a new protein XY. In a network of nodes and edges where each node corresponds to a protein and edges to a PPI interaction, a dense sub graph or a clique corresponds to a protein complex. The main challenge is to identify such complexes from a network which in turn would be helpful in understanding the cell functions. There are different clustering solutions developed for different applications in various domains.

* Department of Computer Science & I.T Amrita School of Arts & Sciences, Kochi Amrita Vishwa Vidyapeetham, AmritaUniversity, India. sreeja.ashok@gmail.com

In Hierarchical approach, clusters are constructed by partitioning the object in either top down or bottom up fashion. This result in the formation of a dendrogram and the clusters are formed by cutting the dendrogram at preferred similarity point[2] . This can be further categorized as Agglomerative and Divisive clustering.[3]Agglomerative clustering begins with each data point in separate cluster and then finding the best pair to merge into new cluster until the desired cluster is formed. In divisive clustering initially, every object belongs to the same cluster which is successively split into smaller sub-clusters. This process continues until the desired cluster is formed.

There are few cluster similarity measures involved in hierarchical clustering. One such measure is the Single linkage clustering, where clustering is done in a bottom-up manner. Clusters with the closest pair of elements are joined to form a new cluster. Here the likeness of two clusters is the similarity of the most similar member, whereas in complete linkage, similarity is the relationship between the most dissimilar members. In Centroid method (average method) the similarity between clusters is given as the distance between their centroids. And the last approach is the Ward's Method, where the sum of squares is calculated for each cluster. The two clusters with the lowest increment in the overall sum of squares are combined. In partitioning method clustering, objects are repositioned from one cluster to another. Here the number of clusters will be pre-set by the user. The existing clustering algorithms have their own drawbacks. The most commonly used Partitioned clustering algorithms like K-Means divides a dataset into a specified number of clusters that is, user have to explicitly specify the number of clusters.[4]K-medoids is another method which tries to minimize the error of sum of squares (SSE).Processing in k-medoids is expensive when compared to k-means method.[5]Density based method group objects according to the density of objects. Density means the number of objects within the neighborhood of particular data objects. DBSCAN is a density based method that can handle large spatial databases but cannot handle duplicate points in datasets.[6]In Grid-Based Method the space is partitioned into fixed number of cells where all the clustering operations are executed. Processing is very fast in this method.

Fuzzy clustering is a soft-computing method where each pattern is associated with every cluster using a function. The most common fuzzy algorithm used is the FCM algorithm. Graph clustering is the process of assembling the vertices of a graph into clusters. It is mainly classified as global and local clustering. Every vertex of the input graph is allocated a group in the output of the method in global clustering, while in local clustering assignments are done only for a definite subset of vertices.[7]

Since by nature the protein-protein interaction forms network structure, graph based clustering is the most appropriate solution for solving PPI networks. Through this paper, a new approach is proposed by integrating redundancy reduction in PPI network with Monte Carlo optimization for automatic detection of meaningful protein clusters.

Necessity of gene/protein expression

With the advancement of technology , there is a huge increase in biological data which resulted in gaining the understanding of human network. Since, it is assumed that proteins sharing common neighbors have similar properties, finding of protein expression can assist in the prediction of characteristics of unknown protein. It is of great importance because it not only offers solid assumptions as to signaling pathways but also provides speculation about the source of the disease.[8]The availability of disease associated clusters has given a better insight about the disease mechanism and also plays a vital role in the advancement of new diagnostics and therapeutics.

Clustering in PPI

Clustering in PPI network is the collection of proteins sharing large number of interactions. Usually the grouping is done on the basis of some similarity measure. As a result of this clustering method, a clear structure of a PPI network is ascertained which make it possible to predict the functions of proteins which were previously not characterized. The two key objectives of clustering are homogeneity and heterogeneity. A homogeneous cluster contains elements of similar type whereas a heterogeneous cluster will contain elements with less similarity.[9] Finding dense sub-graphs is the main purpose of PPI interaction which shows the most significant functional group in protein -protein interactions. Recognizing significant functional unit in PPI network is the first step to realize the

structure and functional dynamic of cell.[10]Group of proteins that highly interact with themselves are known as protein complexes. In the current work, an integrated clustering solution to identify the proteins that function together in the same biological process is introduced.

2. RELATED WORKS

Various methods have been described for finding highly connected protein complexes. In large-scale protein interaction it is difficult to interpret data and it is also challenging to derive function of annotated proteins. Manoj and Liang presented a new algorithm which resolves the problem of interpretation of large data. Two proteins having large number of common neighbors of interaction is considered to have a close functional association, this intuition is used in this paper. Unfortunately this method has a disadvantage that it may not expose all of the functions of the proteins.[11]Sovan Saha and Piyali Chatterjee presented an innovative method that uses sequence similarity to predict protein function.[12]

Later, several other researchers like Bader and Hogue tried to bring in a new method called the MCODE algorithm which provides a very efficient approach for identifying densely-connected sub graphs in large PPI networks. But this algorithm has a weakness in finding protein complexes on modules with a large number of proteins. The paper proposed by Chuan, Young-rae et al[1] provides a concise elucidation about a number of clustering methods which have presented favorable results in application to PPI interaction networks.Ivana Cingovska et al developed a paper that uses two graph clustering techniques for identifying functional groups and predicting protein characteristics from PPI.[13] It also illustrates a general framework for the infinite set of algorithms used for protein function prediction. But this approach has a limitation that it does not consider edge weight into account. Algorithms like PECA automatically find the right number of clusters and final dataset partition without any primal knowledge.[14] Two other researchers Yanjun Qi and William Stafford Noble provided a brief description on the latest attempts made to predict the interactions between proteins and protein domains, they have also tried to point out different methods that used protein interaction data to deduce protein function.[15] Various computational methods for detecting and characterizing protein interactions were analyzed in the paper proposed by sudharamaiah and v. sivasakthi.[16]

Hossein Rahmani et al. perceived a cutting edge method to predict the characteristics of proteins in a PPI network.[17] This approach has been classified into two, inductive and transductive approach and local and global approaches. A new method called a micro-patterning approach is introduced in a paper published on Laboratory Journal which helped in addressing many biological questions.[18] Srinivasa Rao et al. introduced computational methods that will reduce the group of possible interactions to a subset of most expected interactions. Further experiments can be performed considering these interactions as starting point. The gene expression data and protein interaction data will improve the confidence of protein-protein interactions and the corresponding PPI network when used collectively.[19]

3. PROPOSED FLOW

In the proposed approach, highly connected sets of nodes are searched rather than fully connected sets of nodes.Here, two graph clustering methods are incorporated i.e., redundancy reduction and Monte Carlo optimization for automatic detection of meaningful clusters and thus finding highly connected protein complexes in PPI network.The steps involved in the process are as follows

Step 1: Similarity matrix construction taking distance as the parameter from gene expression data

Step 2: Merge the nodes based on the significance of the association between the proteins.

Step 3: Based on the merging, cut the tree into clusters for different cluster size and calculate the maximum optimization value (Q value)

Step 4: Calculate average of sum of all the Q values for each cluster size and the optimum cluster size is the maximum value of the modulus difference between k and $(k + 1)^{th}$ clusters.

Step 1 : Computing similarity matrix

The process starts with the creation of a similarity matrix from the given database. Number of distance measures can be used in clustering to find the closeness of data objects. Some of them are

1. **Minkowski distance :** The Minkowski distance of order q between two points j and k is given by

$$d(j, k) = \sqrt[q]{(|x_{j1} - x_{k1}|^q + |x_{j2} - x_{k2}|^q + ... + |x_{jp} - x_{kp}|q)} \quad (1)$$

where $j = (x_{j1}, x_{j2}...x_{jp})$ and $k = (x_{k1}, x_{k2}...x_{kp})$ are two p-dimensional data objects, and q is some positive integer.

2. **Manhattan distance:** The distance between the points j with coordinates (x_1, y_1) and the point k with coordinates (x_2, y_2) is given by

$$d(j, k) = |x_{j1} - x_{k1}| + x_{j2} - x_{k2}| + ...+| x_{jp} - x_{kp}| \quad (2)$$

3. **Euclidean distance :** Euclidean distance between j and k is given by the formula

$$d(j, k) = \sqrt{(|x_{j1} - x_{k1}|^2 + |x_{j2} - x_{k2}|^2 +...+ |x_{jp} - x_{kp}|^2)} \quad (3)$$

Where 'j' and 'k' are two points in the plane with coordinates (x_1, y_1) and (x_2, y_2)

In the proposed approach, a similarity matrix of protein data is constructed using Euclidean distance formula. Then the similarity matrix is modeled as a graph object, G (V, E) where V corresponds to the nodes or proteins and E represents the similarity value between them.

Step 2: Mergingnodes based on *p*-value

Here p-value is calculated to merge the protein nodes on the basis of the significance of their relationship using the formula

$$p(x, y) = \max_{x \in X, y \in Y} (p(x, y)) \quad (4)$$

Where $p(x, y)$ is the similarity between elements $x \in X$ and $y \in Y$, where X and Y are two clustersthe protein pair with the lowest significance is merged into one.

The resulting matrix is reduced to a matrix of size $k - 1$.The same process is repeated until the final matrix is formed which finally constructs a hierarchical tree like structure called dendogram.

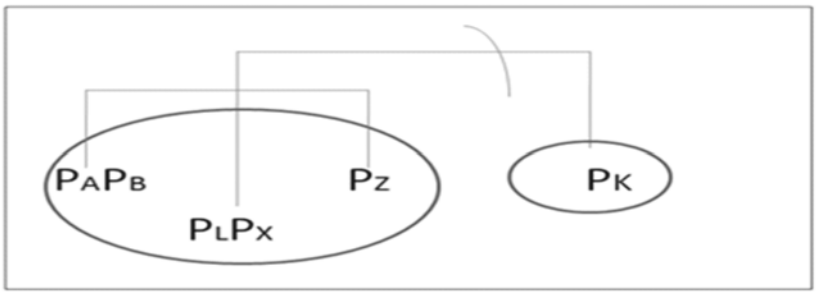

Fig. 2. Hierarchical structure forming protein complexes.

Figure 2 depicts the cluster formation in PPI using p-value. During the first iteration the rows and columns of the protein pair with the smallest p-value is merged, *i.e.* P_A and P_B are merged forming a new cluster ($P_A P_B$), similarly other clusters like ($P_L P_X$), P_Z and P_K are formed. These clusters are again merged to form larger clusters like ($P_A P_B$, $P_L P_X$,P_z)

Step 3: Exploring highly connected network*:*

In this step the tree is cut into K clusters. For each cluster, the number of nodes and edges are identified and the Q value is calculated using equation 5. The objective function used here is to maximize the difference of Q value of two succeeding cluster solutions *i.e.* $Q(k + 1) - Q(k)$ should be maximum.

$$Q(p) = \frac{2m}{n(n-1)} \quad (5)$$

Where '*m*' defines the no of edges and '*n*', the no of nodes in the sub-graph. The object function 'Q' exemplifies the cluster density. A sub graph with object function 1 is considered to be fully connected whereas a sub-graph with no internal edge will have a Q value of0. The sub graph with the highest Q value is considered to be highly connected protein complex. Figure 3 represents a flow chart illustrating the proposed approach

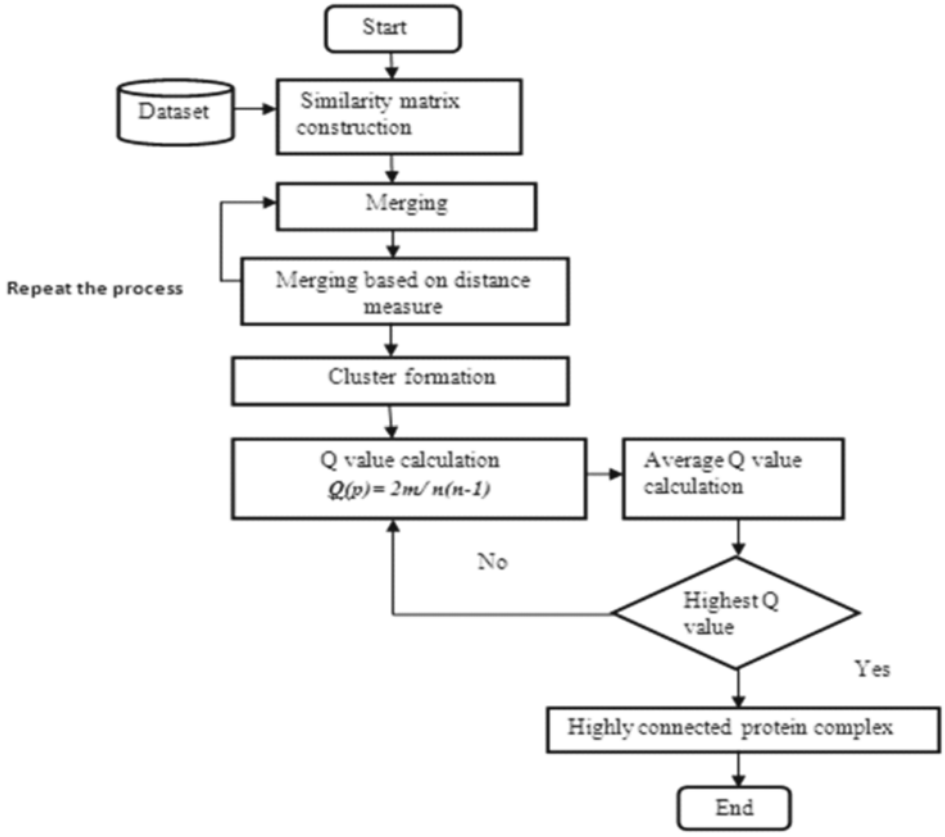

Fig. 3. Process flow of the proposed system.

4. EXPERIMENT AND RESULT

In order to implement the proposed method yeast dataset from UCI machine learning repository is imported, which is a center for machine learning and intelligent systems. The Dataset consist of 1484 instances and 9 attributes with 1 class label. Table 1 depicts the class distribution of the yeast dataset.

Table 1 : Class distribution of Yeast Dataset

Class labels	*Actual distribution*
CYT	463
NUC	429
MIT	244
ME3	163
ME2	51
ME1	44
EXC	37
VAC	30
POX	20
ERL	5

Initially a similarity matrix is derived where each row and column represents a protein .The distance measure computed is taken as p-value that illustrates the similarity between two proteins. The technique is tried with different cluster sizes varying from 2 to 13 and Q value for each cluster distribution is calculated. The average of the sum of the Q values of different clusters is also computed. The computed Q-value for each cluster (6 – 13) is represented in the Table2.

Table 2: Result of *q*-value calculation for different clusters

Clusters	*Cluster distribution*	*Q value*	*Average Q value*
cluster 6	815, 533, 15, 64, 54, 3	0.002453988, 0.003752345, 0.1333333, 0.03125, 0.03703704, 0.6666667	0.1457489
cluster 7	815, 526, 15, 64, 54, 7, 3	0.002453988, 0.003802281, 0.1333333, 0.03125, 0.03703704, 0.285714, 0.6666667	0.165751
cluster 8	641, 526, 15, 64, 174, 54, 7, 3	0.003120125, 0.003802281, 0.1333333, 0.01325, 0.1149425, 0.03703704, 0.2857143, 0.6666667	0.1594832
cluster 9	641, 526, 15, 64, 167, 54, 7, 7, 3	0.003120125, 0.003802281, 0.1333333, 0.03125, 0.1197605, 0.03703704, 0.2857143, 0.2857143, 0.6666667	0.1620682
cluster 10	641, 526, 15, 53, 167, 54, 7, 11, 7, 3	0.003120125,0.003802281,0.1333333,0.03773585, 0.01197605,0.03703704,0.2857143,0.1818182, 0.2857143,0.6666667	0.1986541
cluster 11	641, 524, 15, 53, 167, 54, 7, 11, 7, 2, 3	0.003120125,0.003816794,0.1333333,0.03773585, 0.01197605,0.03703704,0.2857143,0.1818182, 0.2857143,1,0.6666667	0.2406303
cluster 12	641, 524, 15, 53, 162, 54, 7, 11, 7, 2, 5, 3	0.003120125,0.003816794,0.13333333,0.03773585, 0.01234568,0.03703704,0.2857143,0.1818182, 0.2857143,1,0.4,0.6666667	0.2539418
cluster 13	476, 524, 15, 53, 165, 162, 54, 7, 11, 7, 2, 5, 3	0.004201681,0.003816794,0.1333333,0.3773585, 0.01234518,0.03703704,0.2857143,0.1818182, 0.2857143, 0.1818182,1,0.4,0.6666667	0.2433484

Difference of Q value of succeeding clusters *i.e.* (k + 1) and Q (k) is computed and the optimum value is taken as the best cluster size.

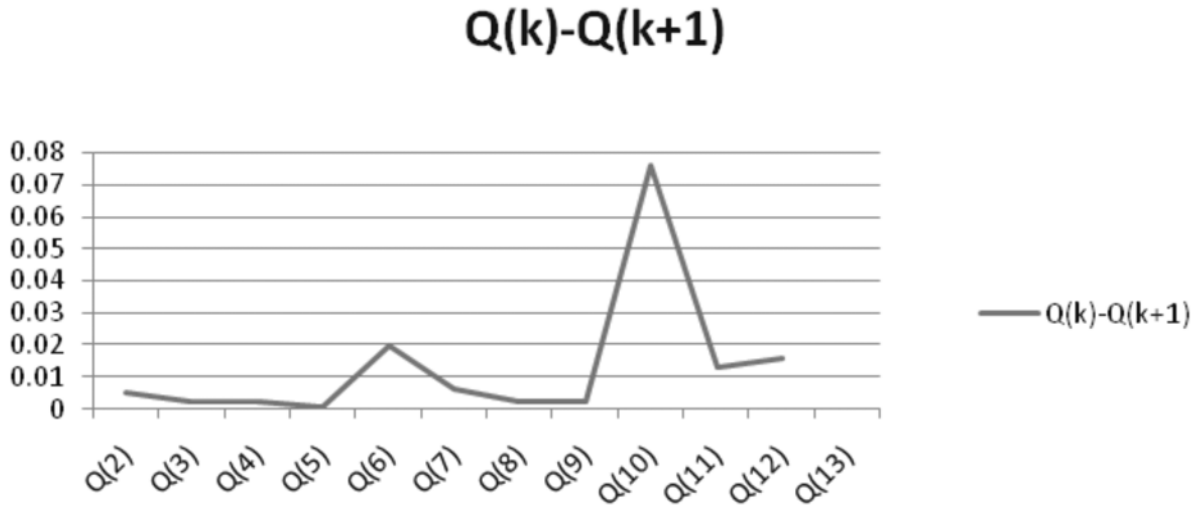

Fig. 4. Diff (Q(k) and Q(k + 1)) plotted against each clusters.

Figure 4 plots the variation of Q value of k^{th} cluster with $(k + 1)^{th}$ cluster. The difference in the Q values between cluster 10 and cluster 11 is hefty and the observed cluster distribution inferred after clustering the dataset with size 10 is close to the actual cluster distribution of the yeast dataset. Thus it can be inferred that cluster 10 cuts the most highly connected protein complexes.

The proposed approach is compared with standard benchmark algorithms like *k*-means and hierarchical clustering. Here we have to explicitly specify the cluster size whereas the proposed approach automatically forms the clusters. Even after providing the clusters, the distribution of data objects in each cluster for both k-means and hierarchical is found contrasting with the actual dataset. The results are depicted in Table 3.

Table 3: Comparison between k-means, single linkage and proposed system

Method	***Cluster distribution***
k-means ($k = 10$)	112, 316,155,134,79,181,3,15,116,373
Hierarchical clustering (Single linkage) $k = 10$	1451,4,11,4,7,1,1,1,1,3
Proposed method (automatic detection)	641,526,15,53,167,54,7,11,7,3

Since biological data are usually not categorized, automatic detection of meaningful groups are more relevant and needed in this domain. Hence the new approach will be more suitable for exploring the protein complexes and for further analysis of each specific group.

5. CONCLUSION

In this paper two different approaches are used to explore highly connected protein complexes: the merging of p-values and Monte-Carlo optimization method. Nearly no protein achieves its function in isolation, thus it is very important to discover most of the existing interactions between them. The objective of the proposed approach is to investigate the most highly connected protein complexes or clusters so as to determine the function of uncharacterized protein complexes, which belong to the same cluster. The proposed method is proven to be successful in finding highly connected and optimized protein complexes by evaluating the object function or the Q-value. The major highlights of the proposed approach when compared with standard clustering solution are the automatic detection of protein complexes.

6. ACKNOWLEDGEMENT

This work is supported by the DST Funded Project, (SR/CSI/81/2011) under Cognitive Science Research Initiative in the Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham University, Kochi.

7. REFERENCES

1. Lin, Chuan, Young-rae Cho, Woo-chang Hwang, Pengjun Pei, and Aidong Zhang. "Clustering methods in protein-protein interaction network." *Knowledge Discovery in Bioinformatics: techniques, methods and application* (2007): 1-35.
2. Rokach, Lior. "Decomposition methodology for classification tasks: a meta decomposer framework." *Pattern Analysis and Applications* 9, no. 2-3 (2006): 257-271.
3. Ali, Raza, Usman Ghani, and Aasim Saeed. "Data clustering and its applications." *Available at the web address: http://members. tripod. com/asim_saeed/paper. htm* (1998).
4. Nazeer, KA Abdul, and M. P. Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." In *Proceedings of the World Congress on Engineering*, vol. 1, pp. 1-3. 2009..
5. Singh, Shalini S., and N. C. Chauhan. "K-means v/s K-medoids: A Comparative Study." In *National Conference on Recent Trends in Engineering & Technology*, vol. 13. 2011.
6. Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, vol. 96, no. 34, pp. 226-231. 1996.

7. Schaeffer, Satu Elisa. "Graph clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.
8. Rahmani, Hossein, Hendrik Blockeel, and Andreas Bender. "Predicting the functions of proteins in protein-protein interaction networks from global information." In *JMLR: Workshop and Conference Proceedings*, vol. 8, pp. 82-97. 2010.
9. Pizzuti, Clara, Simona E. Rombo, and Elena Marchiori. "Complex detection in protein-protein interaction networks: a compact overview for researchers and practitioners." In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pp. 211-223. Springer Berlin Heidelberg, 2012.
10. Spirin, Victor, and Leonid A. Mirny. "Protein complexes and functional modules in molecular networks." *Proceedings of the National Academy of Sciences* 100, no. 21 (2003): 12123-12128.
11. Pratim Samanta, Manoj, and Shoudan Liang. "Redundancies in Large-scale Protein Interaction Networks." *arXiv preprint physics/0303027* (2003).
12. Saha, Sovan, and Piyali Chatterjee. "PROTEIN FUNCTION PREDICTION FROM PROTEIN INTERACTION NETWORK USING PHYSICO-CHEMICAL PROPERTIES OF AMINO ACIDS."
13. Cingovska, Ivana, Aleksandra Bogojeska, Kire Trivodaliev, and Slobodan Kalajdziski. "Protein Function Prediction by Clustering of Protein-Protein Interaction Network." In *ICT Innovations 2011*, pp. 39-49. Springer Berlin Heidelberg, 2012.
14. Ashok, Sreeja, and M. V. Judy. "A novel iterative partitioning approach for building prime clusters." *International Journal of Advanced Intelligence Paradigms* 7, no. 3-4 (2015): 313-325.
15. Qi, Yanjun, and William Stafford Noble. "Protein interaction networks: protein domain interaction and protein function prediction." In *Handbook of Statistical Bioinformatics*, pp. 427-459. Springer Berlin Heidelberg, 2011.
16. Ramaiah, sudha. "experimental and computational techniques in protein interactions v. Sivasakthi and sudha ramaiah."
17. Hartuv, Erez, Armin Schmitt, Jörg Lange, Sebastian Meier-Ewert, Hans Lehrach, and Ron Shamir. "An algorithm for clustering cDNAs for gene expression analysis." In *Proceedings of the third annual international conference on Computational molecular biology*, pp. 188-197. ACM, 1999.
18. DE, GIT-LABOR. "STAY UPDATED GIT LAB PORTALS." (2015).
19. Rao, V. Srinivasa, K. Srinivas, G. N. Sujini, and G. N. Kumar. "Protein-protein interaction detection: methods and analysis." *International journal of proteomics* 2014 (2014).