ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com
Journal DOI:10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Essential Protein Identification from Protein Networks Using Topological and Biological Properties.

***Fathima Shabnam C B[1] and Sminu Izudheen[2].**

1. Department of Computer Science and Information Systems,Mahatma Gandhi University, India.
2. Department of Computer Science and Information Systems,Mahatma Gandhi University, India.

Manuscript Info

Manuscript History:

Received: 12 May 2016
Final Accepted: 23 June 2016
Published Online: July 2016

Key words:
Centrality measures,
Essential proteins,
High through-put techniques,
Protein-protein interaction (PPI) network.

Corresponding Author

Fathima Shabnam C B.

Abstract

Essential proteins play a very important decisive role in the survival of a cellular organisms. Need for identifying essential proteins are increasing for its contribution to the field of drug analysis and synthetic biology is very huge. Centrality methods were the first used to identify essential proteins. Due to its high sensitivity towards network accuracy much more efficient methods which included the biological properties were developed. Cellular localization, biological process, gene expression and domains were some of the biological properties studied along with the network properties to predict essential proteins. This survey focus on studying various methods used to predict essential proteins and compare their performance.

Introduction: -

Human body can be considered as a building made up of bricks where the bricks could be cells, bones, nutrients oreven proteins. Proteins are large bio molecules made up of long chains of aminoacids. Proteins could be classified as essential and non - essential proteins. Here the study focus on identifying essential proteins. Essential proteins have to be included in our diet for the proper metabolic activities to happen inside our body. But as far as synthetic biology is concerned its discovery is also helpful in drug design(Clatworthy et al) and identifying anomalies causing various diseases (Furney et al,2006).

Various experimental approaches for identifying essential proteins that happened to be existing are considered as time consuming and expensive compared to the computational methods. Some of the experiments that biologist opted previously were Single gene knockouts (Giaever et al,2002), RNA interference (Cullen et al ,2005) and Conditional knockouts (Roemer et al,2003). Owing much to the high throughput technologies which generated huge amount of PPI data we are able to analyze essential proteins from its network level. (Jeong et al,2006) was the first to predict the essentiality of a protein with the help of lethality caused by the disruption of link between highly connected proteins. Most of the studies tried to rely on this fact to identify the essential proteins and paved the path for the concept of centralities. Inspired from this, other researchers tried to add biological information along with the topological information.

As there are a lot of works to identify essential proteins it is time to analyze all and tabulate them for future works. A comparison study is essential to predict the best among the methods.

This paper is arranged as following sections:
Section 1gives the introduction and motivation behind this survey.
Section 2 discusses various methods used and a comparison study on them.
Section 3 and 4 concludes the survey and discusses some future works possible.

Literature Survey: -

Development in high throughput techniques have generated large amount of PPI data. Due to the spurious and missing interaction PPI data is not highly reliable. To predict essentiality of proteins both network and biological properties can be used. In the early stages only the network properties were used and the problem was it is highly sensitive to network accuracy. In order to get accurate prediction researchers started to include biological properties. We can classify the prediction methods based on the topological and biological properties used.

Centrality Methods and Essential Proteins: -

A protein interaction network can be represented as graph G (V,E) where V represents the set of proteins and Erepresents the set of edges between pair of proteins (Wang et al,2014). An edge between two vertices u and v can be represented as e (u, v). Figure 1(B) displays an example of a yeast protein–protein interaction network (Gursoy et al,2008), and a small subgraph illustrating a hub protein (node A in Figure 1A).

In graph theory, centrality means the most important vertex in the network. Borrowing the same when we knockout the central node from the network its effect will be lethal. This defines the concept of "Centrality-Lethality rule" (Xionglei et al,2006). These central nodes can act as the trigger for signaling pathways in many diseases (Abedi et al,2015). In (Jeong et al,2001) using the gene essentiality concept evolutionary rate of an organism was studied and they were able to find that central nodes have slower evolutionary rate.

A growing body of research has focused on the prediction of gene essentiality using the network properties and biological features. Consequently, many computational methods have been developed.

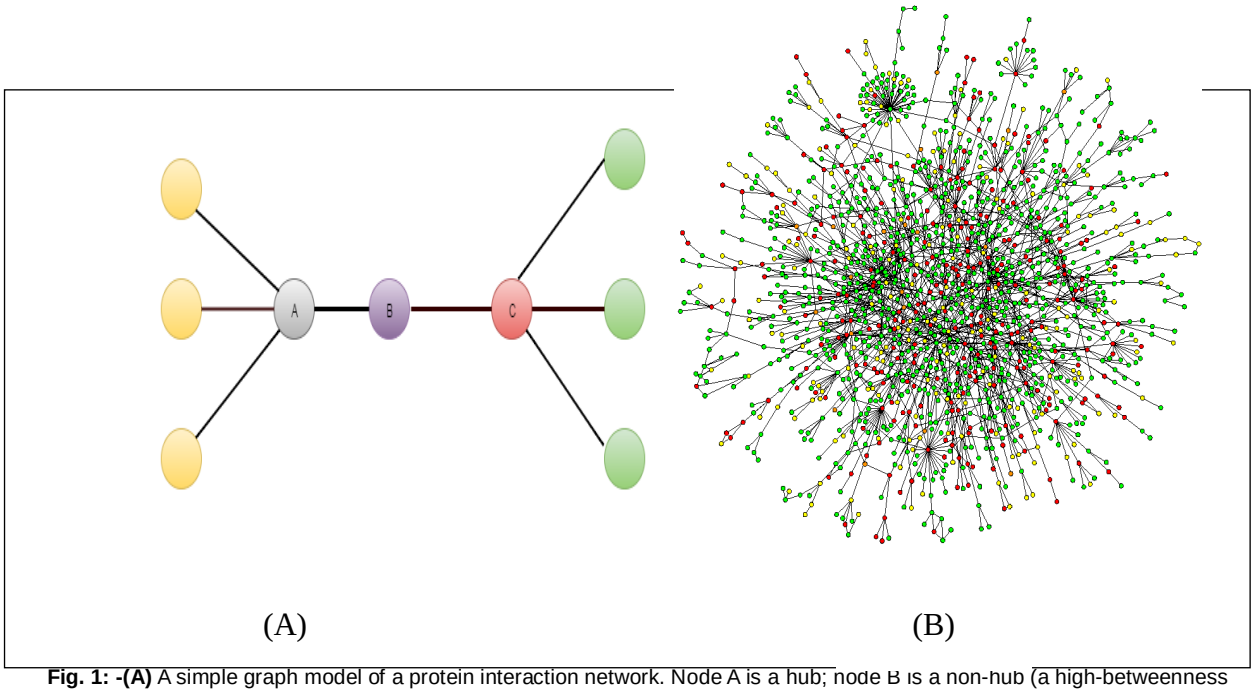

Fig. 1: -(A) A simple graph model of a protein interaction network. Node A is a hub; node B is a non-hub (a high-betweenness node). **(B)** A sample protein interaction network derived from yeast protein interactions.

Some of the centrality methods are degree centrality (DC) (Hahn et al,2005), betweenness centrality (BC) (Joy et al,2005), closeness centrality (CC) (Wuchty et al,2003), subgraph centrality (SC) (Estrada et al,2005), eigenvector centrality (EC) (Bonacich et al,1989), information centrality (IC) (Stephenson et al,1989), edge clustering coefficient centrality (NC) (Wang et al,2012) and so on.

The simplest of all centrality method is Degree centrality. It gives the number of interacting proteins with proteinv_i,$D(v_i)$. Degree centrality uses the basic concept of degree of a node.

To predict the essential proteins, the basic procedure used by all the centrality methods are the same. In the case of degree centrality, for all the proteins first calculate the number of interacting proteins with each protein v_i. Then order the proteins based on the increasing order of degree of protein. Using some sampling methods sample the dataset and predict the results. It is always assumed that whatever be the metric we are using to predict the

essentiality top n percentage is assumed as essential and remaining once as the non-essential. Degree centrality is calculated as

$$DC(u) = \sum_v a_{u,v} \quad (1)$$

where$a_{u,v}$ is 1 if there is a connecting edge between node u and node v, and 0 otherwise.

Another centrality method is Betweeness Centrality. It is the sum of all pairs shortest paths through which a vertex v pass through.

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (2)$$

where σ (s, t) the number of shortest (s, t)-paths and let σ(s, t|v) be the number of shortest (s, t)-paths passing through some vertex v other than s, t.

Closeness Centrality is much more popular than the other two methods. Because it can predict more essential proteins than the other two. While the Betweeness Centrality tries to measure the influence of a protein has in communicating between protein pairs, Closeness Centrality gives the number of links in the shortest path between the protein pairs. It can be defined as

$$C_v(v) = \frac{N-1}{\sum_j d(i,j)} \quad (3)$$

Where N is the total number of proteins and d(i , j) is the distance between protein i and protein j. Sometime proteins having high betweeness but low connectivity form essential links in the network. HBLC (High Betweeness Low Connectivity) proteins were predicted and they tried to study the effect of betweeness on evolutionary rate. But they couldn't differentiate much between the effect of HBLC and non- HBLC on the evolutionary rate as its number is too small.

Eigenvector centrality is not restricted to any shortest path calculation. The network is represented as adjacencymatrix corresponding to the connected subgraphs and eigenvector values. This will help to portray the effect of each node on its neighbors. Since the matrix could give the effect of a protein on the entire proteins in the protein network it can be considered as extended centrality measure. if R is the adjacency matrix,e_{ij} is the eigen vector and β is the eigen value, then the Eigenvector Centrality can be defined as

$$\beta e_i = \sum_j R_{ij} e_j \quad (4)$$

From the methods so far developed using the network properties, by evaluating the resultset of the works done we can generalize Edge Clustering Coefficient Centrality (NC) as the best one. First the edge weight is calculated as the product of parameters used for evaluating the relationship between two proteins. In NC the parameters used were GO functional similarity (GE) (FastSemSim), co-expression levels among genes (PCC) (Zhang et al,2012), the number of times that a PPI pair involved in PPI triangles (NTE) (Wang et al, 2012), and the protein-protein sequence similarity measured using the Jukes-Cantor likelihood (PP) (Jukes et al,1969). Based on this weight a sorted essential gene candidate list is obtained.

Edge Clustering Coefficient can be defined as

$$ECC(u,v) = \frac{z_{u,v}}{\min(d_u - 1, d_v - 1)} \quad (5)$$

Where $z_{u,v}$ is the number of triangles which actually include the edges in the network and d_u and d_v gives the degree of node u and node v. NC (u) is defined as the sum of ECC of directly interacting nieghbours of node u.

$$NC(u) = \sum_v ECC(u,v) \quad (6)$$

Drawback of centrality methods: -

All the centrality measures take the network property as the input. But the problem with any PPI network is that it is not complete and accurate. However, these data contain missing and spurious interactions (Mering et al,2002). Even the records say that for Y2H and TAP-MS the missing interaction ranges from 43 to 71 percent and 15 to 50 percent

and spurious interaction is 64 and 77 percent respectively (Edwards et al,2002). From this it is quite clear that reliability of PPI network is not adequate. To overcome these problem researchers tried to include biological information along with the topological information and this led to the generation of the second category as mentioned in Table 1.

Essential Protein and Biological Properties: -

The drawbacks of topological properties led to the integration of biological properties into the prediction method. When Hart and his fellows (Hart et al,2007) pointed out the special connection between the protein complexes and essential protein, (Ren et al,2011) used the concept to predict essential protein combined with network topology. It is said that essential proteins are more conserved than non-essential protein and using that concept (Peng et al,2012) developed an iteration method named ION considering the orthology with PPI network. Their prediction results showed high performance over the centrality methods. But they failed to provide any proof to show their performance level with other methods using biological properties.

All these methods when tried to consider only one property a machine learning based computational approach relying on network topological features, cellular localization and biological process information was developed (Gustafson et al,2006) for predicting essential genes. They used j48 algorithm to generate a decision tree to rove the importance of their parameters in predicting essential genes. More importantly they could use this decision tree to generate cellular rules governing essentiality.

Among the methods that uses the biological properties so far better results were obtained for two algorithms: PeC (Zhang et al,2012) and UDoNC(Peng et al,2015).

In Pec to predict essentiality of a protein gene expression profiles are used along with the edge clustering coefficient. So here the biological term is gene expression profiles and topological property is edge clustering coefficient. To measure the performance, they only considered the proteins from DIP database and showed better performance when compared with other centrality methods. To measure the gene expression profiles values they used the Poisson correlation coefficient.

$$PCC(X,Y) = \frac{1}{s-1}\sum_{i=1}^{s}\left(\frac{g(X,i)-mean\ (g(X)}{StandardDeviation\quad (g(X)}\right) * \left(\frac{g(Y,i)-mean\ (g(Y)}{Standard\ \ Deviation\ \ (g(Y)}\right) \qquad (7)$$

The new centrality measure PeC(u) is defined as the sum of product if ECC and PCC.

$$PeC(u) = \sum_{v} ECC(u,v) * PCC(u,v) \qquad (8)$$

UDoNC is the most recent method developed to predict essential genes by combining topological properties and the protein domain.

Protein domain is the basic building block of protein structure. Domain confines to a particular function of a protein or it can contribute to its evolution. Sometimes similar domains tend to perform different function in different proteins. That means one protein domain type could be present in more than one protein. Based on this fact the algorithm UDoNC predicted the essential proteins from the PPI data.

An example of a protein that contains multiple SH3 domains is the cytoplasmic protein Nck. Nck belongs to the adaptor family of proteins and it is involved in transducing signals from growth factor receptor tyrosine kinases to downstreamsignal recipients. The domain composition of Nck is illustrated in Figure 2.1 below.

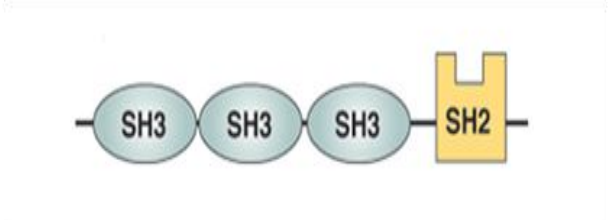

Figure 2.1: -Domain composition of Nck

According to UDoNC a protein is said to be essential if it consists of rarely occurring domains in other protein and as non-essential if it consists of frequently occurring domains.
Essentiality of a protein was defined in term of number of protein domain and its frequency. Probability of protein u was defined as

$$P(u) = NDT_{norm} * SFD_{norm} \quad (9)$$

Where NDT and SFD are number of domain types and sum of frequency of domains respectively. Finally,UDoNC was calculated as sum of product of ECC and weight of each edge. From the results of UDoNC they made it quite clear that their method is efficient than all other predicting methods. However, there is still room for improvement.

Database of Essential gene: -

Yeast PPI data are downloaded from DIP (Xenarioset al,2002) database. It consists of more than 5000 proteins and >25,000 interactions for yeast.

The essential gene can be integrated from MIPS (Mewes et al,2006), SGD (Cherry et al,1998), DEG (Zhang et al,2009)and SGDP(http://www.sequence.stanford.edu/group/), which contains 1,285 essential proteins. OGEE (Chen et al,2012) is an another database that gives the complete collection of essential and non-essential genes for yeast.

Sl. No	Essential Protein Prediction Methods	
	Topological Properties	**Biological Properties**
1	Degree Centrality	Gene Expression(PeC)
2	Betweeness Centrality	Protein Domain(UDoNC)
3	Closeness Centrality	Integration of cellular localization and biological process information
4	Eigen Vector Centrality	
5	Subgraph Centrality	
6	Edge Clustering Coefficient Centrality	

Table 1: -Classification of Essential Protein Prediction Methods

Performance evaluation measures: -

Since the system for essential gene identification is a prediction system evaluation measures are necessary. Usually the results are classified as True positive(TP), True negative (TN), False positive (FP), False negative(FN).
Some of the evaluation measures are as follows:

Sensitivity or true positive rate (TPR) : $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$

Specificity (SPC) or true negative rate (TNR): $SPC = \frac{TN}{N} = \frac{TN}{FP+TN}$

Precision or positive predictive value (PPV): $PPV = \frac{TP}{TP+FP}$

Negative predictive value (NPV): $NPV = \frac{TN}{TN+FN}$

Accuracy (ACC): $ACC = \frac{TP+TN}{P+N}$

F1 score, is the harmonic mean of precision and sensitivity: $F1 = \frac{2TP}{2TP+FP+FN}$

Matthews correlation coefficient (MCC): $MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Conclusion: -

Through this survey we made a study on essential gene prediction methods. Works done show that centrality methods which used only the topological properties showed less performance when compared to the methods that included biological properties along with the topological properties. So many works on this field indicate its importance in the field of human disease analysis and drug design.

Future Work: -

As the application areas are disease diagnosis, drug analysis and cosmetics, we could predict the importance of essential gene in these areas. Very active areas is disease analysis with the help of this algorithm we could establish the correlation between essential genes and human disease gene.

References: -

1. **E. Clatworthy, E. Pierson, and D. T. Hung,** "Targeting virulence: A new paradigm for antimicrobial therapy," Nature Chem.
2. **S. J. Furney, M. M. Alb_a, and N. L_opez-Bigas(2006),** "Differences in theevolutionary history of disease genes affected by dominant or recessive mutations," BMC genomics, vol. 7, no. 1, p. 165.
3. **G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai- Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston(2002),** "Functional profiling of the Saccharomyces cerevisiae genome," Nature, vol. 418, pp. 387–391**.**
4. **L. M. Cullen and G. M. Arndt (2005), "**Genome-wide screening for gene function using rnai in mammalian cells," Immunol. Cell Biol., vol. 83, no. 3, pp. 217–223.
5. **T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, F. Tandia, A. Linteau, S. Sillaots, C. Marta, N. Martel, S. Veronneau, S. Lemieux, S. Kauffman, J. Becker, R. Storms, C. Boone, and H. Bussey(2003),** "Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery," Molecular Microbiology, vol. 50, pp. 167–181.
6. **H. Jeong, S. P. Mason, A.-L. Barab_asi, and Z. N. Oltvai (2001),** "Lethality and centrality in protein networks," Nature, vol. 411, no. 6833, pp. 41–42.
7. **J. Wang, X. Peng, W. Peng, and F.-X. Wu (2014),** "Dynamic protein interaction network construction and applications," Proteomics, vol. 14, no. 4/5, pp. 338–352.
8. **Attila Gursoy, OzlemKeskin and Ruth Nussinov(2008),**" Topological properties of protein interaction networks from a structural perspective", Biochem. Soc. Trans. 36, 1398–1403; doi:10.1042/BST0361398.
9. **Xionglei He, and JianzhiZhang(2006),** "Why Do Hubs Tend to Be Essential in Protein Networks?," PLoS Genetics , www.plosgenetics.org, June , Volume 2 , Issue 6 , e88.
10. **Abedi and Gheisari (2015),** "Nodes with high centrality in protein interaction networks are responsible for driving signaling pathways in diabetic nephropathy", PeerJ3: e1284; DOI 10.7717/peerj.1284
11. **M. W. Hahn and A. D. Kern (2005),** "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks," Molecular Biol. Evol., vol. 22, no. 4, pp. 803–806.
12. **M. P. Joy, A. Brock, D. E. Ingber, and S. Huang (2005),** "Highbetweenness proteins in the yeast protein interaction network," BioMed Res. Int., vol. 2005, no. 2, pp. 96–103.
13. **S. Wuchty and P. F. Stadler(2003**), "Centers of complex networks," J. Theoretical Biol., vol. 223, no. 1, pp. 45–53.
14. **E. Estrada and J. A. Rodr_ıguez-Vel_azquez(2005),** "Subgraph centrality in complex networks," Phys. Rev. E, vol. 71, no. 5, p. 056103.
15. **P. Bonacich(1989),** "Power and centrality: A family of measures," Amer. J. Sociol., vol. 92, pp. 1170–1182, 1987.
16. K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," Soc. Networks, vol. 11, no. 1, pp. 1–37.
17. **J. Wang, M. Li, H. Wang, and Y. Pan (2012**), "Identification of essential proteins based on edge clustering coefficient," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 9, no. 4, pp. 1070–1080, Jul.
18. **M MFastSemSim.** Available: http://sourceforgenet/p/fastsemsim/home/ Home/unpublished.
19. **Li M, Zhang H, Wang Jx, Pan Y (2012**)," A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data". BMC SystBiol 6: 15.

20. **Wang J, Li M, Wang H, Pan Y (2012)," Identification** of essential proteins based on edge clustering coefficient". IEEE/ACM Trans ComputBiolBioinform 9: 1070–1080.
21. **Jukes TH, Cantor CR (1969)**," Evolution of protein molecules**".**
22. **C.vonMering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork (2002),** "Comparative Assessment of Large-Scale Data Sets of Protein- Protein Interactions", Nature, vol. 417, no. 6887, pp. 399-404.
23. **A.M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein (2002),** "Bridging Structural Biology and Genomics: Assessing Protein Interaction Data with Known Complexes", Trends in Genetics, vol. 18, no. 10, pp. 529-536.
24. **G. T. Hart, I. Lee, and E. M. Marcotte(2007),** "A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality", BMC Bioinformat., vol. 8, no. 1, p. 236.
25. **J. Ren, J. Wang, M. Li, H. Wang, and B. Liu(2011),** "Prediction of essential proteins by integration of ppi network topology and protein complexes information", in Proc. 7th Int. Conf. Bioinformat. Res.Appl., pp.1224.
26. **W. Peng, J. Wang, W. Wang, Q. Liu, F.-X. Wu, and Y. Pan (2012),** "Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks", BMC Syst.Biol., vol. 6, no. 1, p. 87.
27. **A.Gustafson, E. Snitkin, S. Parker, C. DeLisi, and S. Kasif(2006),** "Towards the identification of essential genes using targeted genome sequencing and comparative analysis", BmcGenomics,vol. 7, no. 1, p. 265.
28. **M. Li, H. Zhang, J.-X. Wang, and Y. Pan (2012),** "A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data", BMC Syst. Biol., vol. 6, no. 1, p. 15.
29. **Wei Peng, Jianxin Wang, Yingjiao Cheng, Yu Lu, Fangxiang Wu, and Yi Pan (2015)**, "UDoNC: An Algorithm for Identifying Essential Proteins Based on Protein Domains and Protein-Protein Interaction Networks ", IEEE/ACM Transactions on computational biology and bioinformatics, vol.12, No.2, March/April.
30. **Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, et al. (2002)," DIP**, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions". Nucleic Acids Res 30: 303–305.
31. **H.-W. Mewes, D. Frishman, K. F. Mayer, M. M€unsterk€otter, O.Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V.St€umpflen(2006),** "Mips: Analysis and annotation of proteins from whole genomes in 2005," Nucleic Acids Res., vol. 34, no. suppl 1, pp. D169–D172.
32. **J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, and M. Schroeder (1998),** "SGD: Saccharomyces genome database," Nucleic Acids Res., vol. 26, pp. 73–79.
33. **R. Zhang and Y. Lin (2009),** "Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes," Nucleic Acids Res., vol. 37, no. suppl 1, pp. D455–D458, 2009.
34. **Chen WH, Minguez P, Lercher MJ, Bork P (2012**)," OGEE: an online gene essentiality database". Nucleic Acids Res 40: D901–D906.