Computer Engineering and Applications Vol. 8, No. 1, February 2019

Identification of Significant Proteins Associated with Diabetes Mellitus Using Network Analysis of Protein-Protein Interactions

M. Syafiuddin Usman[1], Wisnu Ananta Kusuma[1,4*], Farit Mochamad Afendi[2,4], Rudi Heryanto[3,4]

[1]*Department of Computer Science, Bogor Agricultural University*
[2]*Department of Statistics, Bogor Agricultural University*
[3]*Department of Chemistry, Bogor Agricultural University*
[4]*Tropical Biopharmaca Research Center, Bogor Agricultural University*
**ananta@apps.ipb.ac.id*

ABSTRACT

Computation approach to identify significance of proteins related with disease was proposed as one of the solutions from the problem of experimental method application which is generally high cost and time consuming. The case of study was conducted on Diabetes Melitus (DM) type 2 diseases. Identification of significant proteins was conducted by constructing protein-protein interactions network and then analyzing the network topology. Dataset was obtained from Online Mendelian Inheritance in Man (OMIM) and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) which provided protein data related with a disease and Protein-Protein Interaction (PPI), respectively. The results of topology analysis towards Protein-Protein Interaction (PPI) showed that there were 21 significant protein associated with DM where INS as a network center protein and AKTI, TCF7L2, KCNJ11, PPARG, GCG, INSR, IAPP, SOCS3 were proteins that had close relation directly with INS.

Keywords: Diabetes Melitus, Protein-Protein Interaction (PPI) Network, OMIM, Protein-Protein Significance, STRING

1. INTRODUCTION

The existence of significant proteins in our body is very essential for growth because proteins are the place where all nutritions are needed[1]. Germs infection (Wolfe et al. 2007) or unhealthy lifestyle [2] can cause some diseases in fungtional disorders of those significant proteins[3] and the worst result of them is the malfunction of certain vital organs in our body[4]. One of the diseases caused by functional disorder of significant proteins is DM. Therapy by using proper medication needs to be done to restore those proteins in their normal functions[3].

The initial step in the discovery process of proper medication is identification of the significant proteins associated with the disease[5]. The application of experimental methode to identify the significant proteins such as single gen knockouts[6], interferenceof RNA [7] and conditional knockouts [8] can be conducted. However, the application of these methods is generally constrained by great cost and long time[9].

Computation approach for identifying significant proteins associated with disease is proposed as solutions against experimental constrain. One of them is by constructing and analyzing network topology of protein-protein interaction [9][10].

This can be done because of the availability of protein-protein interactiondata in large numbers and the technology advances of high throughput[9].

The purpose of this research is to identify significant proteins associated with DM through computational approach by analyzing subnetwork topology from protein-protein interaction network. The identified significant proteins are used in the next step in developing of medicine with network pharmacology method.

2. RESEARCH METHOD

2.1 DATA SET

Dataset used in this research are (1) data candidate of significant proteins associated with DM extracted from OMIM database[3]. This data could be obtained by defining the query type of "Gene Map" with query key of "diabetes mellitus". (2) data of protein-protein interaction (PPI). This data was obtained from STRING Database by involving 2,5 million of protein from 630 different organisms[11].

2.2 METHODOLOGY

2.2.1 NETWORK CONSTRUCTION OF PROTEIN-PROTEIN INTERACTION AND SUBNETWORK

Protein-protein interaction is a physical contact formed between two or more proteins studied from some perspectives such as biochemistry, quantum chemistry, molecular dynamics, signal transduction and electrostatic force[12][13][14] which can help in understanding how biological network operates[15].

Subnetworkis PPI network in the amount of minimal node from all node (protein). Nodes which are not traversed the shortest path of interaction binary pairs between candidates proteins will be removed from the network so it can be obtained the network which only involves essential nodes with large degree value. Subnetwork is constructed from all the shortest paths of binary pairs from all candidate proteins collected from PPI network. These network and the subnetwork are then constructed by using programming language of Python and Cytoscape application [16] for network visualization.

2.2.2 TOPOLOGY ANALYSIS OF SUBNETWORK FROM PPI NETWORK

Connectivity degree (k), betweenness centrality (BC) and closeness centrality (CC) are three important parameters in graf theory used for subnetwork analysis proses from PPI network and determination of significant proteins associated with a disease[17]. Degree (k) is defined as the amount of interaction related with a protein and another protein adjacent directly. BC is defined as the size of node centrality in the graphic which calculating the number of times the node acts as a bridge along the shortest path between two other nodes [18] and formulated as follows:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (1)$$

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

where σ_{st} is the number of the shortest path from 's'to't', andσ_{st}(v) is the number of the shortest path from 's'to't' through 'v' knot. CC is defined as the average length of the shortest path to access all other proteins in the network. The greater the value of CC, the more central the protein[19]. CC is formulated as follows:

$$C_C(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)} \quad (2)$$

where d_G(s,t) is the length of the shortest path between two nodes of 's'and't'in graphic G, where d_G(s, s) = 0, d_G(s, t) = d_G(t, s) inundirected graph.

2.2.3 SEARCHING FOR PROTEINS WITH THE HIGHEST VALUE OF BC FROM THE SUBNETWORK FOR CONSTRUCTING NETWORK BUFFER

Subnetwork describes linkage between candidate protein and strong non-candidate protein (Figure2). Thus, determining the major proteins is based on those which are widely used as a crossing (high BC). Critical point of high BC is set in 5% of total network node set [20][21]. All high BC proteins obtained are used to construct the buffer network.Information of significant proteins, proteins (node) of network center and proteins that are directly adjacent to the central protein obtained from formed network buffer. The protein of network center is determined from protein (node) that has the highest CC value from the subnetwork[19].

2.3 EVALUATION OF NETWORK BUFFER ENDURANCE

Evaluation of network buffer endurance was conducted by building test network which used only a portion of combination from 63 protein candidates that had PPI data. The amount of protein candidates will be eliminated from 1 to 6 (10% of 63) protein [10]. For the amount of proteins eliminated from 2 to 6 proteins, the center protein would always be inlucuded to be eliminated to test its robustness in the network. The numbers of proteins candidate combination would be 63 test combinations if 1 protein candidate was removed. If 2 proteins candidate (including the center protein) eliminated, there would be 62 test combinations. For the numbers of protein removed from 3 to 6 proteins, the numbers of proteins candidate combinations would be taken 30 test combinations respectively chosen in random order [10]. As an illustration, if there were 3 proteins candidates eliminated, then the numbers of proteins candidate combinations were 234.260 (63 x 62 x 60) combinations. Thus, the total numbers of test combinations conducted were 265 (58 + 57 + (30 x 5)) tests.

3. RESULTS AND ANALYSIS

3.1 DATA OF SIGNIFICANT PROTEINS CANDIDATES AND PPI DATA

OMIM databasere stored 83 significant proteins candidates relevant associated with DM. While STRING database restored 63 out of 83 candidates protein which had 2,053 data of PPI and the other 20 did not have the PPI data. From all PPI data obtained, there were 418 seed proteins other than candidates protein involved.

3.2 CONSTRUCTION OF PPI NETWORK AND SUBNETWORK FROM ALL SHORTEST PATH

Construction of PPI network from 2,053 PPI data generated one large and five small networks from 5 candidates protein.Since the small networks did not connect directly to the large one, they were not included in the process of subnetwork construction from PPI network.The component of large network had 426 nodes (58 nodes were labeled candidates protein) with 1,857 interactions between them (Figure 1).

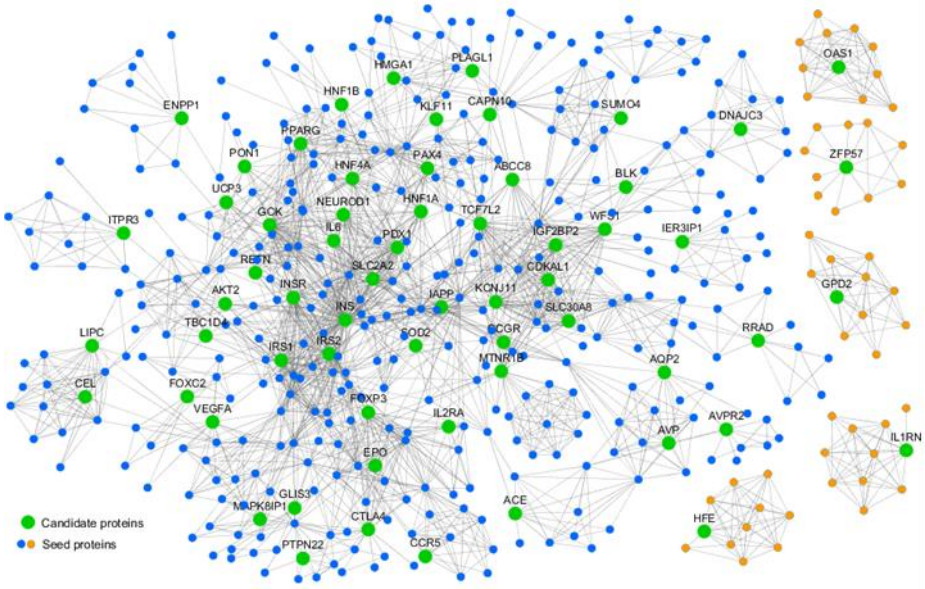

FIGURE 1. Protein Protein Interaction Network

Subnetwork was constructed to obtain simpler PPI network than the large network shown in Figure 1. Subnetwork construction was started by tracing all shortest paths from all binary pairs between candidates protein in the large networks, then chosing the shortest paths that ran through the nodes with the largest degree value[10] for each pairs of candidates protein.

The choice of the shortest paths through the nodes with the largest degree value aimed to avoid the false nodes of the false interactions[22]. The trace result selected 1,596 out of 7,169 shortest paths formed. Then, the subnetwork was constructed

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

from all chosen shortest paths (Figure 2). Subnetwork formed consisted of 91 nodes with 259 interactions between them.

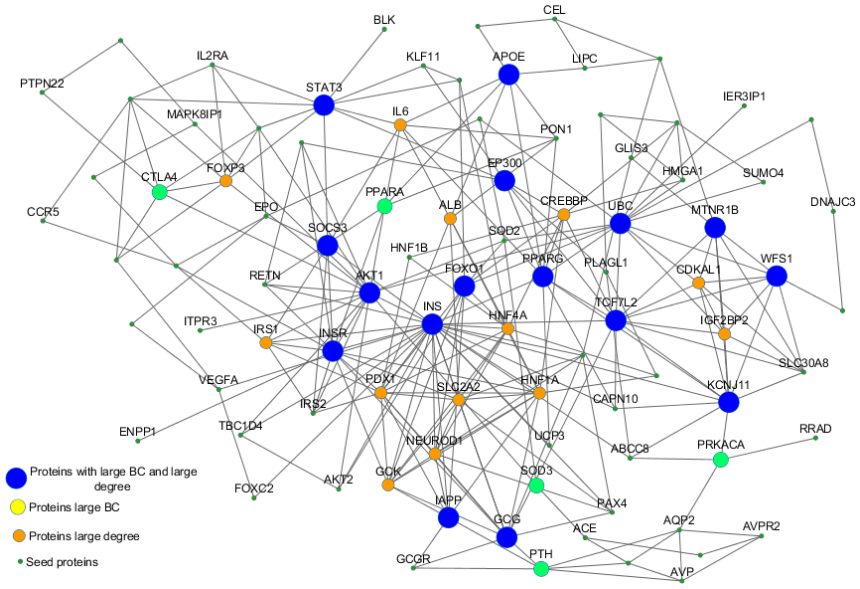

FIGURE 2. Subnetwork formed of all the shortest paths of binary pairs between candidates protein.

3.3 DETERMINING KEY PROTEIN SUBNETWORK

The BC value for every node in the network was used as the base in determining the key node (protein).The key node was selected from high value of BC with the critical point of the number of the node was set in 5% of the total node in the network[20][21]. The topology analysis in the subnetwork of the PPI network resulted 21 proteins as nodewith BC and 9 proteinsamong them were candidates protein (Table 1).

TABLE 1.
Proteins with high value of BC in the subnetwork

Protein Symbols	BC Value	CC Value	Protein Symbols	BC Value	CC Value
INS*	0,255551	0,486486	PRKACA	0,04657	0,309278
AKT1	0,234001	0,452261	FOXO1	0,043799	0,382979
UBC	0,141515	0,401786	EP300	0,043525	0,378151
TCF7L2*	0,091839	0,430622	SOD3	0,042913	0,333333
INSR*	0,082576	0,428571	WFS1*	0,033569	0,351563
KCNJ11*	0,072136	0,403587	MTNR1B*	0,032198	0,327273
GCG	0,069506	0,405405	CTLA4*	0,031935	0,334572
PPARG*	0,06617	0,410959	PPARA	0,031583	0,362903
STAT3	0,065138	0,348837	SOCS3	0,030828	0,378151
PTH	0,05519	0,3125	IAPP*	0,030167	0,375
APOE	0,054121	0,342205	* : CandidatesProtein		

All proteins with high BC value were the key nodes and those 21 nodes were constructed to be the network buffer[10].The interesting thing is that the Table 2 shows there are only 9 candidates protein included in 63 candidates protein from the total of 21 key proteins with the high value of BC and the remaining 12 proteins are the seed proteins obtained from PPI.

Whereas the nodes with large degree value have 29 nodes dan 19 nodes among them are Candidates Protein (Table 2).

TABLE 2.
Proteins with large degree value in the subnetwork

Protein Symbols	Degree Value	CC Value	Protein Symbols	Degree Value	CC Value
INS*	26	0,486486	GCK*	10	0,378151
AKT1	17	0,452261	CREBBP	9	0,368852
INSR*	15	0,428571	IAPP*	8	0,375
PDX1*	14	0,394737	WFS1*	8	0,351563
TCF7L2*	13	0,430622	MTNR1B*	8	0,327273
HNF1A*	12	0,38961	CDKAL1*	7	0,348837
UBC	12	0,401786	ALB	7	0,328467
HNF4A*	12	0,387931	IGF2BP2*	7	0,348837
GCG	12	0,405405	FOXP3*	7	0,346154
SLC2A2*	11	0,381356	IL6*	7	0,357143
PPARG*	11	0,410959	APOE	7	0,342205
KCNJ11*	11	0,403587	IRS1*	7	0,387931
NEUROD1*	11	0,375	SOCS3	7	0,378151
STAT3	10	0,348837	EP300	7	0,378151
FOXO1	10	0,382979	* : CandidatesProtein		

Result on the Table 1 and 2 show that INS (insulin) is a node which is in the network center because INS has the lasgest CC value[19] and the highest BC value as well. Thus, INS is the most important nodefrom 21 keys protein with high BC value. Both tables also show 16 proteins as nodes with high BC value and large degree (Table 3).

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

TABLE 3.
Proteins with High BC Value and Large Degree

Protein Symbols	Descriptions
INS	Insulin; Insulin decreases blood glucose concentration.
AKT1	v-Akt murine thymoma viral oncogene homolog 1; Regulate many processes including metabolism, proliferation, cell survival, growth and angiogenesis.
UBC	Ubiquitin C
TCF7L2	Transcription factor 7-like 2 (T-cell specific, HMG-box)
INSR	Insulin receptor; Receptor tyrosine kinase which mediates the pleiotropic actions of insulin.
KCNJ11	Potassium inwardly-rectifying channel, subfamily J, member 11; This receptor is controlled by G proteins
GCG	Glucagon; Glicentin may modulate gastric acid secretion and the gastro-pyloro-duodenal activity.
PPARG	Peroxisome proliferator-activated receptor gamma
STAT3	Signal transducer and activator of transcription 3; Signal transducer and transcription activator that mediates cellular responses to interleukins, KITLG/SCF and other growth factors.
APOE	Apolipoprotein E; Mediates the binding, internalization, and catabolism of lipoprotein particles.
FOXO1	Forkhead box O1; Transcription factor that is the main target of insulin signaling and regulates metabolic homeostasis in response to oxidative stress.
EP300	E1A binding protein p300; Functions as histone acetyltransferase and regulates transcription via chromatin remodeling.
WFS1	Wolfram syndrome 1 (wolframin); Participates in the regulation of cellular Ca(2+) homeostasis, at least partly, by modulating the filling state of the endoplasmic reticulum Ca(2+) store
MTNR1B	Melatonin receptor 1B; High affinity receptor for melatonin. The activity of this receptor is mediated by pertussis toxin sensitive G proteins that inhibit adenylate cyclase activity.
SOCS3	Suppressor of cytokine signaling 3; SOCS3 is involved in negative regulation of cytokines that signal through the JAK/STAT pathway

TABLE 4.
Proteins with High BC Value and Large Degree

Protein Symbols	Descriptions
IAPP	Islet amyloid polypeptide; Selectively inhibits insulin-stimulated glucose utilization and glycogen deposition in muscle, while not affecting adipocyte glucose metabolism

And there are 5 proteinswhich only have high BC (Table 4). To distinguish the roles of the nodeson the table above, they are described with different colours and measurements (Figure 2).

TABLE 5.
Proteins with High BC Value without Large Degree Value

Protein Symbols	Descriptions
PTH	Parathyroid hormone; PTH elevates calcium level by dissolving the salts in bone and preventing their renal excretion. Stimulates [1-14C]-2- deoxy-D-glucose (2DG) transport and glycogen synthesis in osteoblastic cells.
PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha; Regulates the abundance of compartmentalized pools of its regulatory subunits through phosphorylation of PJA2 which binds and ubiquitinates these subunits, leading to their subsequent proteolysis.
SOD3	Superoxide dismutase 3; Protect the extracellular space from toxic effect of reactive oxygen intermediates by converting superoxide radicals into hydrogen peroxide and oxygen
CTLA4	Cytotoxic T-lymphocyte-associated protein 4.
PPARA	Peroxisome proliferator-activated receptor alpha; Ligand-activated transcription factor.

3.4 NETWORK BUFFER FROM ALL PROTEINS WITH HIGH BC

Buffer network is constructed from all high BC nodes as the results of extraction on the subnetwork (Figure 3). At first, BC value were introduced to measure the centrality of nodes in the network and had function as communication controller between nodes in the network. The buffer network is constructed to know how the interaction of all high BC nodes in the network. The result of this buffer network construction has 21 nodes with 38 interactions among them.

TABLE 6.
Proteins with High BC Value in the Buffer Network

Protein Symbols	BC Value	CC Value	Protein Symbols	BC Value	CC Value
INS	0,321111	0,625	PPARA	0,03114	0,408163
AKT1	0,24348	0,512821	WFS1	0,018567	0,444444
TCF7L2	0,200292	0,571429	APOE	0,016257	0,384615
KCNJ11	0,134211	0,5	FOXO1	0,009649	0,37037
UBC	0,109678	0,487805	STAT3	0,006579	0,350877
PPARG	0,095205	0,512821	PTH	0,004386	0,350877
GCG	0,077953	0,47619	CTLA4	0	0,344828
INSR	0,077544	0,5	MTNR1B	0	0,392157
IAPP	0,052632	0,434783	PRKACA	0	0,338983
SOCS3	0,051754	0,434783	SOD3	0	0,344828
EP300	0,044298	0,416667			

In this network, INS has the highest BC value and the nodein the network center with the largest CC value (Table 4). There are 8 nodes protein directly adjacent to INS: AKT1, TCF7L2, KCNJ11, PPARG, GCG, INSR, IAPP dan SOCS3 (Figure 3).

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

INS nodeis a highly connected nodein the network because INS has the highest BC value and the largest degree value[23]either in the large network, the subnetwork,or the the buffer network. It means that INS nodeis a protein which arranges all interactions traiffic between nodes and binds nodes of other proteins which interact directly with it [23].

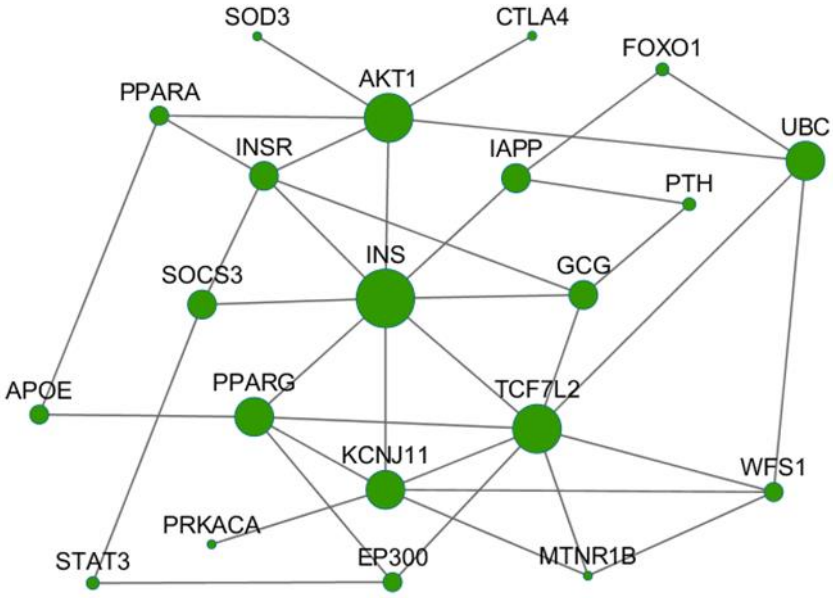

FIGURE 3. The buffer network is constructed from 21 high BC proteins extracted from subnetwork of PPI network, and the size of noderepresents the value of BC

3.5 ROBUSTNESS EVALUATION OF BUFFER NETWORK

From three parameters (BC, CCdan Degree) of network evaluation calculated, there are 3 proteinswith the highest BC value, 2 proteins with the largest CC value and 1 protein with the largest degree value on the test network (Table 6). The frequencies of INS as nodewith the highest BC with 184emergences, nodewith the largest CC with 203 emergences and 235 emergences as nodewith the largest degree dengan Degree show the position of INS as network center, traffic controller of all protein-protein interactionand protein binding other proteins which are directly adjacent with it strongly. They are supported with the accurate large enough value of buffer network, 0.86321839 (Table 7).

TABLE 7.
Frequencies of emergence from the nodeswith the largest/highest BC, CC and Degree value on 235 test network

Numbers of Proteins eliminated	Frequancies of highest BC, CC and Degree value on the Test Network						Accuracy of Buffer Network	Number of Test Network
	BC			CC		Degree		
	INS	AKT1	TCF7L2	INS	TCF7L2	INS		
1	40	16	2	53	5	58	0,926929392	58
2	50	5	2	52	5	57	0,904761905	57
3	25	5	0	28	2	30	0,880952381	30
4	23	2	5	26	4	30	0,849206349	30
5	24	2	4	22	8	30	0,836507937	30
6	22	5	3	22	8	30	0,780952381	30
Total	184	35	16	203	32	235	0,86321839	235

4. CONCLUSION

Identification of significant proteins associated with DM was conducted by extracting node of high BC protein in the PPI network and its number was determined in 5% of the total protein nodes in the network. The buffer network was constructed to find out the interaction between the significant proteins and it can be obtained the protein description which is the most responsible in controlling interaction traffic between proteins and as the network center. From all constructed tests, it can be concluded that INS is the most significant protein compared with 20 significant proteins that exist.

ACKNOWLEDGEMENT

The authors would like to give deep gratitude to Ministry of Education and Culture (DIKTI), Republic of Indonesia for funding this research through the BPP-DN.

REFERENCES

[1] M L Acencio, N Lemke. Towards the prediction of essential genes by integration of information.*BMC Bioinformatics*. 2009; 18: 1–18.

[2] H E & U M Tuomilehto J, Indstrom J, Eriksson J, Valle T. Numb Er 18 Prevention of Type 2 Diabetes Mellitus By Changes in Lifestyle Among Subjects With Impaired Glucose Tolerance.*N. Engl J Med*. 2001; 344(18): 1343–1350.

[3] H Huang, X Wu, R Pandey, J Li, G Zhao, S Ibrahim, J Y Chen. C2Maps: a

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

network pharmacology database with comprehensive disease-gene-drug connectivity relationships.*BMC Genomics*.2012; 13(6): S17.

[4] J Calles-Escandon,M Cipolla. Diabetes and endothelial dysfunction: a clinical perspective.*Endocr Rev*.2001; 22: 36–52.

[5] J Li, X Zhu, J Y Chen. Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts.*PLoS Comput Biol.*2009; 5(7): e1000450.

[6] S Dow, A Lucau-danila, K Anderson, A P Arkin, A Astromoff, M El Bakkoury, R Bangham, R Benito, S Brachat, B Andre, D F Jaramillo, D E Kelly, S L Kelly, P Ko. Functional profiling of the Saccharomyces cerevisiae genome.*Nature*.2002; 418(6896): 387–91.

[7] L M Cullen, G M Arndt, A T Park, N S Wales. Genome-wide screening for gene function using RNAi in mammalian cells.*Immunol Cell Biol*.2005; 83: 217–223.

[8] 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, H Bussey. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery.*Mol. Microbiol*.2003; 50(1): 167–181.

[9] M Li, H Zhang, J Wang, Y Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol*. 2012; 6(1): 15.

[10] J Ran, H Li, J Fu, L Liu, Y Xing, X Li, H Shen, Y Chen, X Jiang, Y Li, H Li. Construction and analysis of the protein-protein interaction network related to essential hypertension. *BMC Syst. Biol*.2013; 7: 32.

[11] L J Jensen, M Kuhn, M Stark, S Chaffron, C Creevey, J Muller, T Doerks, P Julien, A Roth, M Simonovic, P Bork, C von Mering. STRING 8--a global view on proteins and their functional interactions in 630 organisms.*Nucleic Acids Res*.2009; 37: D412–D416.

[12] H D Herce, W Deng, J Helma, H Leonhardt, M C Cardoso.Interactions in living cells.*Nat Commun.*2013; 4: 1–8.

[13] S Jones, J M Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A.*1996; 93(1): 13–20.

[14] E M Phizicky, S Fields. Protein-protein interactions: Methods for detection and analysis.*Microbiol Rev*.1995; 59(1): 94–123.

[15] Y Xia, H Yu, R Jansen, M Seringhaus, S Baxter, D Greenbaum, H Zhao, M Gerstein. Analyzing Cellular Biochemistry In Terms of Molecular Networks.*Annu Rev Biochem*. 2004; 73: 1051–87.

[16] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, T Ideker. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks.*Cold Spring Harb Lab Pres*. 2003; 13: 2498–2504.

[17] S Hwang, SW Son, S C Kim, Y J Kim, H Jeong, D Lee. A protein interaction network associated with asthma.*J Theor Biol*.2008; 252(4): 722–731.

[18] L C Freeman. Centered Graphs and The Structure of Ego Networks.*Math Soc Sci*.1982; 3: 291–304.

[19] L C Freeman. Centrality in Social Networks Conceptual Clarification.*Elsevier*.1978; 1: 215–239.

[20] J Goñi, F J Esteban, N de Mendizábal, J Sepulcre, S Ardanza-Trevijano, I Agirrezabal, P Villoslada. A computational analysis of protein-protein interaction networks in neurodegenerative diseases.*BMC Syst. Biol*. 2008; 2(52).

[21] M Kim, K Kwon, M L Cyberpsychology. Psychological Characteristics of Internet Dating Service Users : The Effect of Self-Esteem, Involvement, and Sociability on the Use of Internet Dating Services.*Cyberpsychol Behav*.2009; 12(4): 445–449.

[22] G Lima-mendez, J Van Helden. The powerful law of the power law and other myths in network biology.*Mol Biosyst*.2009; 5(12):1482-93.

[23] J D J Han, N Bertin, T Hao, D S Goldberg, G F Berriz, L V Zhang, D Dupuy, A J M Walhout, M E Cusick, F P Roth, M Vida. Evidence for dynamically organized modularity in the yeast protein- protein interaction network. *Nature*. 2004; 430: 88–93.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)