Journal of Engineering and Applied Sciences 14 (15): 5183-5189, 2019
ISSN: 1816-949X

Extraction of Essential Genes Based on Network Attributes

[1]Mohammad-Ashraf Ottom, [2]Khalid M.O. Nahar and [3]Izzat Alsmadi
[1]Department of Computer Information System, Yarmouk Univeristy, Shafiq-Irshidat St., Irbid, Jordan
ottom.ma@yu.edu.jo, +962799442111
[2]Department of Computer Science, Yarmouk Univeristy, Shafiq-Irshidat St., Irbid, Jordan
khalids@yu.edu.jo, +962772747892
[3]Department of Computing and Cyber Security, University of Texas A&M, San Antonio, TX, USA
ialsmadi@tamusa.edu, +12089726299

Abstract: Protein and DNA feature's extraction represents an interesting research subject for a wide range of relevant applications. In this study, we studied different methods of grouping a large number of genes based on relations with other genes. We used different network metrics such as centrality degree and betweenness to find essential genes. We proposed and developed an algorithm to extract the total and weighted strengths in associating gene's relations with each other. The results showed that such group related metrics can be used to effectively extract knowledge about genes and their associations with other genes as well as with diseases.

Key words: Genes, genes groups, maximum cliques, genecards, group centrality metrics, network metrics

INTRODUCTION

Networks have been studied in many domains such as social networks, computer networks and biological networks. The biological networks gained a significant attention lately due to the recent research emphasis on human body and related health issues. The primary types of networks that attracted attention in biological networks are metabolic networks, genetic regulatory networks and protein-protein interaction networks. A Protein-Protein Interaction network (PPI) is a graph that represents the interaction between all protein-protein interactions. The vertices in the graph represent proteins and two vertices are connected by an undirected edge when there is a connection or interaction between the proteins (vertices). Wang *et al*. (2016) study shows a sample graph that represents PPI in Fig. 1. In biological networks, four centrality variables are usually studied to characterize essential genes based on their relations with other genes.

Degree centrality or connectedness: Which defined as the number of interactions of a gene. This includes both degree out from the gene to other genes (Out degree) or degree in from other genes to the subject gene (In degree). Nieminen introduced a general measure of degree centrality by counting the number of adjacencies for a certain node, p_k:

$$C_D(p_k) = \sum_{i=0}^{n} a(p_i, p_k) \quad (1)$$

where, a (p_i, p_k) if and only if p_i and p_k are connected by a line and 0 when there is no direction between them. C_D (p_k) is large when p_k has many adjacencies and C_D (p_k) when p_k is not connected to any node in the network.

Betweenness (or shortest path) centrality: Defined as the ratio of the number of the shortest paths that pass through a gene to the number of the shortest path between any pair of genes in the genes network. While genes in the top degree centrality are considered as "Hubs", genes in the top betweenness metric are considered as bottlenecks. The betweenness for p_k node in a network is determined by computing the shortest path between each pair of nodes (p_s, p_t) in the network, then summing the fractions for all nodes pairs (p_s, p_t) that pass through the nodes as shown in the following Eq. 2:

$$C_B(p_k) = \sum_{p_s \neq p_{t \neq} p_k}^{n} \frac{\sigma_{(p_s, p_t)}(p_k)}{\sigma_{(p_s, p_t)}} \quad (2)$$

Closeness centrality: It is an indicator of how "Close" a gene is connected to all other genes of the genes network. It represents the mean number of links connecting a gene to all other genes in the genes network in other words,

Corresponding Author: Mohammad-Ashraf Ottom, Department of Computer Information System, Yarmouk Univeristy, Shafiq-Irshidat St., Irbid, Jordan

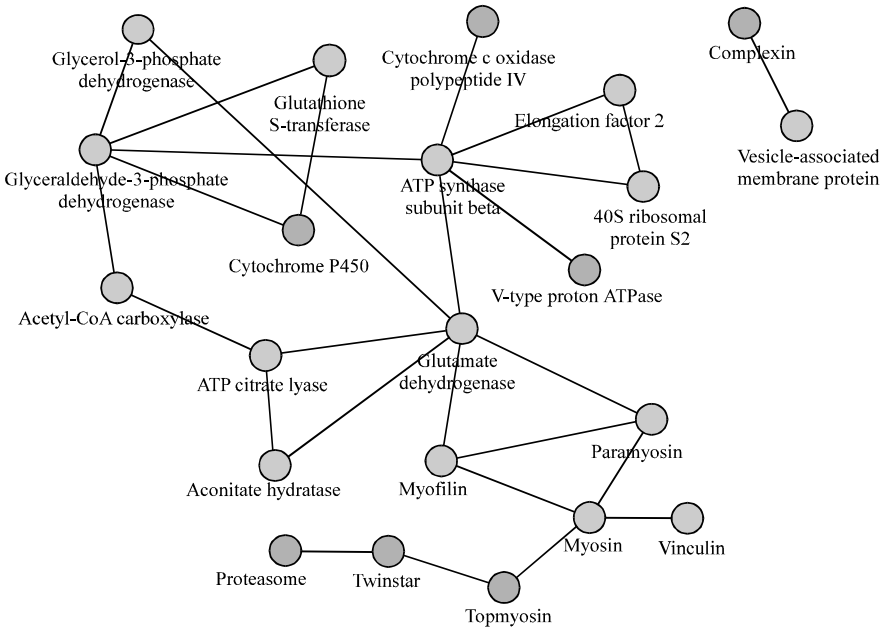

Fig. 1: Sample Protein-Protein Interaction network (PPI)

closeness is the average length of the shortest path between one node and all other nodes in the network as shown in Eq. 3:

$$C(p_k) = \frac{1}{\sum_{p_y} d(p_y, p_k)} \quad (3)$$

where, d (p_y, p_k) is the distance between nodes p_y and p_k.

Eigenvector centrality: This is a measure of the influence of a gene in a genes network. This metric is similar to Google PageRank and differentiates connecting to low scoring or high scoring nodes or genes. The main idea of eigenvector centrality is to show that important node in the network is connected to other significant adjacent nodes in the network. Let, G (E, V) a network of genes where V are nodes of genes and E are the edges between genes. Let, A be the adjacency matrix for this network then the centrality of a gene is proportional to the sum of the centralities of genes which it is connected. The λ is the largest eigenvalue of A and n is the number of genes:

$$A\lambda = \lambda x, \quad \lambda x_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1,\ldots,n$$

Many reasons stand behind studying such connectivity variables. Moreover, many reasons made researchers interested to find out such details about the different genes. Here are some examples clarify what we have declared. Genes with high centrality, betweenness or closeness evolve slowly while evolve fast, if those variables are low (Ragan, 2009).

In regular networks, protein bottlenecks have a higher tendency to be genes which makes betweenness good predictor to choose a potential target (Yu *et al*., 2007). In gene networks, largest eigenvalues usually are preferred since it grants a better accuracy. This is because that the eigenvector is a factor of the matrix (being that symmetric) and the eigenvalues measures the precision that which can reproduce that matrix (Grando, 2015).

In the information flow graph, nodes with low closeness scores, tends to have shorter distance from others. This is mean that sooner it will receive information assuming information flow from all nodes in equal probability and along the shortest path. This indicates that low closeness scores are well-positioned to receive early and novel information (Borgatti, 2005).

Literature review: Several studies investigated genes network to discover the relationship between genes and human diseases. A recent study by Liu and Pan (2016) examined the use of genes network to predict cancer associated genes in human using genes network centrality. They used centrality metrics to forecast the most cancer-causing genes. The study revealed 6 new genes that are possible cancer associated genes which may not discovered yet by medical researchers as a cancer associated genes and showed the potential and capability of using centrality metrics to discover

candidate human diseases-associated genes. The study relied on four centrality metrics, degree, betweenness, closeness and PageRank. Each metric produced scored genes and the researchers decided to select the highest 47 genes for every individual metric (188 genes). Among the 188 genes, there were 89 common gens where 58 of them were cancer associated genes, 4 genes were not translated into a protein and 27 genes were inferred genes. Medical databases and researches confirmed that 21 genes of inferred genes were cancer associated genes and 6 genes were likely to be cancer associated genes.

Another study (Ozgur *et al*., 2008) proposed a new approach to discover diseases-associated genes. They evaluated the approach on prostate cancer and found the ability of centrality metrics to discover unknown candidate's genes associated with prostate cancer as well as other already known genes associated with the disease. In this approach, initial known disease-genes collected and then disease-gene interaction network constructed by mining the interactions between the initial known genes and their neighbors based on the biomedical disease literature using Support Vector Machine (SVM). On the other hand, they used centrality metrics to score genes in the network based on their relevance to the disease. The main assumption in the approach was to consider genes at the central of network as more likely to be associated with the disease. The result of testing the approach on prostate cancer showed that 95% of the top 20 genes are related to the disease by using degree and eigenvector centrality metrics while closeness and betweenness centrality metrics presented genes that are currently unknown to be related to the disease.

Another application of centrality metrics is to identify the genes associated with osteoporotic fracture healing. The study Gao *et al*. (2017) obtained the data from the gene expression omnibus database accession number GSE51686. Hub (centrally) genes refer to the relatively the most important genes in the network. In this study, centrally genes were identified using three centrality metrics, degree, betweenness and subgraph centrality methods. The ranks gained from the degree, betweenness and subgraph methods were calculated using the CytoNCA plug-in (Version 2.1.6) in cytoscape (a software for complex molecular network analysis and visualization). Higher ranks for the centrality metrics demonstrate that the genes were more significant in the network. The study claimed as the first study to determine that Sdc2, Fkbp10, Oasl2, Ifit1 and Ifit2 may be associated with osteoporotic fracture healing which open the door for medical researcher to investigate the issue further in the lab.

Ozgur *et al*. (2008) developed literature mining strategy called Centrality and Ontology-based Network Discovery using Literature data (CONDL). The strategy used centralities metrics to rank genes in the literature-mined gene networks and to obtain a score of importance of genes. For instance, a gene is considered important when occurs on many shortest path between other genes (betweenness centrality) and the gene could be marked with high level of importance when connected to many other genes (degree centrality). Based on CONDL (Ozgur *et al*., 2008) established a web-based literature mining database system called Ignet to keep track of the interaction between genes, based on medical publications abstracts on PubMed. Ignet can be used to extract gene interactions and to generate new hypothesis about genes and related diseases.

A study by Sun and Zhongming examined network characteristics of the proteins encoded by cancer in human PPI. They initiated that network structure of cancer proteins was much unlike from proteins encoded by essential proteins or control proteins and cancer proteins were characterized with higher degree, higher betweenness, shorter shortest-path distance and weaker clustering coefficient in the human PPI. They also studied two categorizes of cancer proteins, recessive and dominant cancer proteins PPI. They found that recessive cancer proteins had higher betweenness than dominant cancer proteins while their degree distribution and characteristic shortest path distance were also significantly different. They concluded that cancer proteins in human PPI are highly correlated.

MATERIALS AND METHODS

STRING-Search Tool for the Retrieval of Interacting Genes/proteins: STRING (Search Tool for the Retrieval of Interacting Genes/proteins) is genes repository and web system for displaying the known relationships between genes and to predict the interactions between genes. STRING is free genes resource where the datasets updated regularly. The interactions in STRING is derived and imported by many sources such as genes literature experiments, interactions from primary/curated databases, text-mining the published articles and interactions that are observed in one organism are systematically transferred to other organisms. STRING integrates and ranks these associations to produce rank evidence and confidence.

RESULTS AND DISCUSSION

The study of genes as network models: Network graphs are collection of nodes or vertices with inter-connecting edges or arcs. In our models genes represent the individual nodes. As our evaluated model is to consider relations between genes, different attributes from (GenesLikeMe) website are collected to represent the weights between the different genes. Edges are directed attributes that can vary from one side to the other.

Our first dataset includes 238 genes randomly selected. For each gene, we extracted the top 100 similar genes along with similarity metrics from GeneCards website. Our second dataset includes top hubs or genes that have highest degree values (Chen *et al*., 2013).

Degree of centrality: The degree of centrality represents the number of links of a node (i.e., gene) with other nodes (neighbors) in the network. Nodes with very high degree are called "Hubs", since, they are connected to many neighbors. The removal of "Hubs" nodes has a great impact on the network topology and may negatively affect its flow. This is the first category of "Essential genes". We started by calculating the total weighted degree centrality for each gene. Results will be shown based on two categories.

The top hubs they are the genes with the degree of centrality for those genes that were selected from the dataset. Table 1, shows those genes hubs from our first randomly collected 238 genes. As we have selected top 100 peer genes all selected genes will show (The out degree, number of connections out from a Gene G) to be as 100. The attribute (The in degree, number of connections coming into Gene G) indicates other calls or edges from other nodes in the dataset that are listing the subject gene as a top peer.

We can see that looking at the "weight" factor that can define the "Strength" of the relation in one aspect or attribute can give us more details or the nature of the relation. For example, while the first top 2 hubs in our dataset have the same total degree of centrality (i.e., 201) their weighted degree of centrality is very different. We have calculated the weighted degree of centrality by dividing the total degree from all 100 nodes by 100. For the gene, ADD3-AS1 in particular, the weighted degree is high as 9 genes have a total similarity score with ADD3-AS1 of more than 1 Table 2.

The top genes with degree of centrality for genes that were not selected in the dataset Table 3. This list is also from our first 238 gene's dataset. Table 3 shows that while those genes were not selected in our dataset (i.e., Degree out = zero), yet they showed up in the dataset because of their high (Degree In Values). This means that in general for all genes those genes will most likely be ranked as top hubs or essential genes.

Top hubs dataset, taken from (Chen *et al*., 2013) Table 4. Those genes are known as "Top hubs". That's why their total degrees are higher than our randomly selected genes. On the other hand, when we calculate the weighted degree (which can give us more details on the strength of such top hubs) not only some of the top peers score lower than many others but they also score lower than our randomly selected genes.

Table 1: Top genes/hubs based on degree centrality

Genes	Degree	Degree (in)	Degree (out)	Weighted degree
ADD3-AS1	201	101	100	2.04
AGBL1-AS1	201	101	100	1.21
A1BG-AS1	200	100	100	1.00
A2M-AS1	200	100	100	1.00
AIRN	200	100	100	1.03

Table 2: ADD3-AS1 gene (top 10 gene peers)

Weights

Ranks	Gene symbol	Total
1	ATXN80S	1.61
2	LINC00328	1.60
3	PWAR5	1.58
4	DDRI-ASI	1.58
5	ANP32A-1T1	1.57
6	ERC2-1T1	1.56
7	CPSI-ITI	1.51
8	MCM3AP-AS1	1.51
9	CEBPA-ASI	1.51
10	ENSG00000267942	1.00

Table 3: Top hubs based gene in-links only

Genes	Degree	Degree (in)	Degree(out)
PROC	60	60	0
CBS	52	52	0
MBL2	50	50	0
F10	48	48	0
AR	47	47	0
ABCG5	43	43	0
DNMT3B	43	43	0
CYP1A2	43	43	0
CEBPA	41	41	0
NR1I2	41	41	0
HNF1A	40	40	0

Table 4: Top 15 gene hubs

Genes	Degree	Degree (in)	Degree (out)	Weighted degree
SPSB1	1278	1217	61	1.83
TRIM21	994	926	68	1.93
TROVE2	951	902	49	2.33
YWHAB	856	778	78	1.10
GRB2	822	751	71	2.52
GSK3B	756	682	74	2.08
CDK2	727	664	63	2.24
CDK1	707	652	55	2.48
ESR1	676	577	99	1.69
SLC2A4	676	630	46	1.59
CSNK2A1	633	559	74	2.26
UBC	626	538	88	1.65
MAPK14	609	543	66	2.19
PRKCA	608	536	72	1.93
SRC	583	497	86	2.15

SPSB1 is the gene with the highest degree of centrality. It means that it is listed with many other genes as similar to them. However, if we look at the weighted degree score that, we took from the top 100 peer genes, the value is 1.83 (Table 4). On the other hand, there are many of the top 15 hubs wither a higher weighted degree than SPSB1 (e.g., SRC, MAPK14, CSNK2A1, CDK1, CDK2, GSK3B, GRB2, GRB2 and TROVE2) (Table 4). All those genes scored a weighted score of more than 2.

GRB2 has the highest weighted score of 2.52. From Table 4, top 5 genes with highest weighted degree of centrality are: GRB2, CDK1, TROVE2, CSNK2A1 and CDK2, respectively.

Table 5 shows why GRB2 scored higher in weighted degree of centrality than SPSB1, although, SPSB1 has a larger number of peer genes. The total score that indicates the strength of connectivity between the gene and its peers is much higher in most of the top 20 peer genes.

Betweenness centrality: The betweenness centrality is the measure of how a node is in the center of the network.

Table 5: Top 20 peer genes for GRB2 and SPSB1

	GRB2		SPSB1	
Ranks	Gene symbol	Total score	Gene symbol	Total score
1	SRC	4.38	SPSB2	3.84
2	CSK	4.3	SPSB4	3.84
3	VAV1	4.2	CTSB	2.38
4	LYN	4.15	TRIM21	2.27
5	PTPN11	4.11	WSB1	2.18
6	ABL1	3.8	ASB13	2.18
7	HCK	3.66	ASB9	2.18
8	FYN	3.63	ASB16	2.18
9	PIK3R1	3.53	FBXO45	2.11
10	FGR	3.47	SOCS1	2.11
11	SHC1	3.34	MID1	2.11
12	RAF1	3.29	TRIM11	2.04
13	LCK	3.28	ITGB5	2.02
14	STAT1	3.24	AKT1	2.00
15	JAK2	3.16	ASB8	1.96
16	MAP2K1	3.1	ASB10	1.96
17	AKT1	3.08	ASB6	1.96
18	SYK	3.04	SOCS3	1.96
19	PLCG1	3.01	ASB2	1.95
20	MAPK1	3.00	CDC20	1.93

It represents the fraction of shortest paths inside the network which utilize a Gene G. It shows important nodes that lie on a high proportion of paths between other nodes in the network. Proteins with high betweenness centrality have been termed "Bottlenecks" for their role as key connectors of proteins with essential functional and and dynamic properties. Table 6 shows top 10 genes in terms of betweenness in our dataset. Some of those showed as top betweenness in other research publications, e.g. (Wang *et al*., 2016).

Eigenvector centrality: The eigenvector centrality (importance) measure used to evaluate and compare the importance of nodes between different conditions (Davidson and Oshlack, 2018). From Table 7, we can see that genes such as, JUN, TNFAIP3, NFKBIA and IL6 continuously appear in the top of the different metrics. Those are clearly essential genes to investigate for any association with diseases, anomaly behaviors, etc. (Table 5 and 6).

Figure 2 shows the interaction genes network between genes JUN, TNFAIP3, NFKBIA and IL6,

Table 6: Top 10 genes in betweenness

Genes	Degree	Betweenness
JUN	110	1928
NFKBIA	107	1600
IL6	106	974
NFKBIE	101	694
TNFAIP3	102	371
REL	104	236
ATF3	100	0
CH25H	100	0
CREB5	100	0
DUSP8	100	0

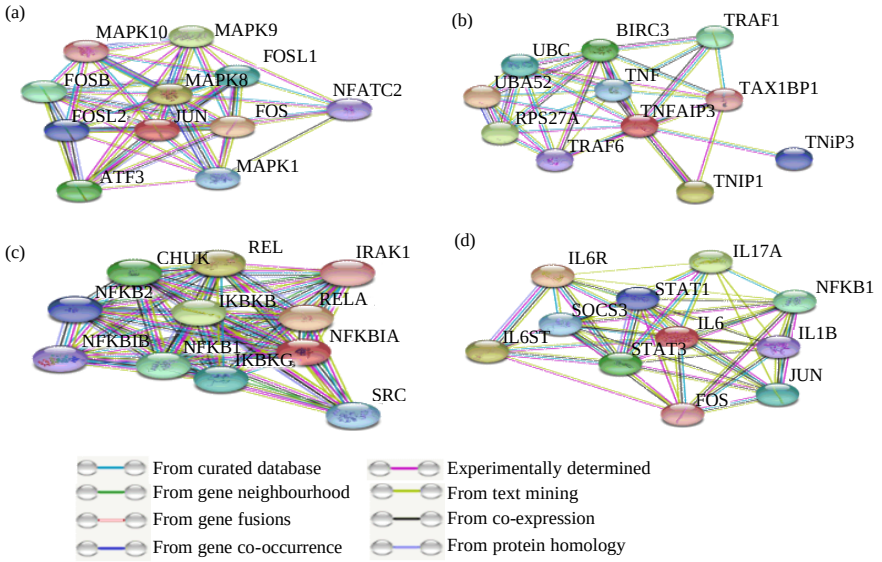

Fig. 2: Evidence based gene-gene interaction network for genes (using STRING): a) JUN; b) TNFAIP3, c) NFKBIA and d) IL6

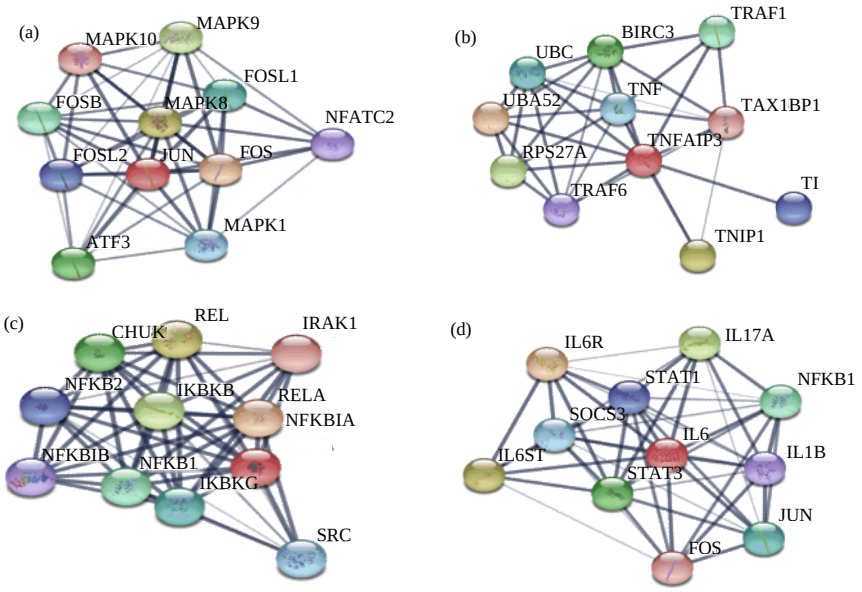

Fig. 3: Confidence based gene-gene interaction network for genes (using STRING): a) JUN; b) TNFAIP3; c) NFKBIA and d) IL6

Table 7: Top 10 genes in eigen centrality

Genes	Degree	Betweenness	Eigen C
NFKBIA	107	1600	1
REL	104	236	0.922239
JUN	110	1928	0.855354
TNFAIP3	102	371	0.736912
ATF3	100	0	0.688358
NFKBIE	101	694	0.5522
CREB5	100	0	0.496886
HES1	100	0	0.426361
IL6	106	974	0.403531
FOS	10	0	0.348625

generated using STRING tool. The interaction between gens depends on many factors such, known interactions from curated databases and known experiments, predicted interaction based on gene neighborhood, gene fusions and gene co-occurrence other sources such as text-mining. However, Fig. 3 expresses the interaction between genes based on the confidence score where line thickness indicates the strength of interaction between genes.

CONCLUSION

In this study we evaluated a large dataset of genes to study inner relations between genes. We focused on studying some centrality metrics, namely degree, betweenness and eigen centrality. Our main goal was to discover essential genes from the evaluated dataset that will show in top of those metrics. We noticed that some genes such as JUN, TNFAIP3, NFKBIA and IL6 continuously appear in the top of all or most centrality metrics. Similar, results were shown in relevant other research publications. Variations can depend on which metric is used to define the similarity criteria between the two genes in any edge-relation association. From the website (GenesLikeMe) we collected several similarity scores while we only showed results of the "Total score" between genes.

ACKNOWLEDGEMENT

We thank faculty of IT&Computer Science at Yarmouk University and the the Department of Computing and Cybersecurity at Texas A&M University for providing support and resources to achieve this research.

REFERENCES

Borgatti, S.P., 2005. Centrality and network flow. Social Netw., 27: 55-71.

Chen, E.Y., C.M. Tan, Y. Kou, Q. Duan and Z. Wang *et al*., 2013. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. BMC. Bioinf., 14: 1-14.

Davidson, N.M. and A. Oshlack, 2018. Necklace: Combining reference and assembled transcriptomes for more comprehensive RNA-Seq analysis. Giga Sci., 7: 1-6.

Gao, F., F. Xu, D. Wu, J. Cheng and P. Xia, 2017. Identification of novel genes associated with fracture healing in osteoporosis induced by Krm2 overexpression or Lrp5 deficiency. Mol. Med. Rep., 15: 3969-3976.

Grando, F., 2015. On the analysis of centrality measures for complex and social networks. Master Thesis, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

Liu, X. and L. Pan, 2016. Predicating candidate cancer-associated genes in the human signaling network using centrality. Curr. Bioinf., 11: 87-92.

Ozgur, A., T. Vu, G. Erkan and D.R. Radev, 2008. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinf., 24: i277-i285.

Ragan, M.A., 2009. Trees and networks before and after Darwin. Biol. Direct, 4: 1-38.

Wang, Y., T. Huang, L. Xie and L. Liu, 2016. Integrative analysis of methylation and transcriptional profiles to predict aging and construct aging specific cross-tissue networks. BMC. Syst. Boil., 10: 413-421.

Yu, H., P.M. Kim, E. Sprecher, V. Trifonov and M. Gerstein, 2007. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLoS Comput. Boil., 3: 0713-0720.