Journal of Theoretical and Applied Information Technology
10th November 2014. Vol. 69 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRIORITIZATION OF CENTRALITY MEASURES IN PROTEIN-PROTEIN INTERACTON NETWORK FOR DISEASE GENE IDENTIFICATION

APICHAT SURATANEE
Department of Mathematics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
E-mail: apichatsu@kmutnb.ac.th

ABSTRACT

For several years, many studies attempt to discover biological processes of disease mechanisms. Nevertheless, they are still far from completeness of understanding. This problem is caused by the complexity of complex diseases. To solve this problem, many computational methods have been developed to predict uncovered disease genes. A lot of genetic information from protein interaction network, gene expression, and genetic sequences has been integrated. With these approaches, a large number of candidate genes are produced increasingly. Therefore, a technique that can select only relevant genes is needed. Ranking techniques have been developed to prioritize the candidate genes. Still, the results are inconsistent among different methods. These incompatibilities might be caused from different types of features. In this study, we performed a prioritization analysis for investigating network topology features for predicting disease-related genes. Four standard network topological features were calculated on a protein-protein interaction network and examined with 46 groups of diseases. The features were ranked independently according to their values for a disease. Then, the performance of disease gene classification with each feature was calculated. The results showed high classification performance in three diseases with different network features. The closeness centrality showed a superior ability to classify disease genes in overall disease groups. Selecting relevant features can greatly improve the performance in disease gene classification.

Keywords: *Feature Prioritization, Disease Gene Identification, Protein-Protein Interaction Network, Network Topology Features*

1. INTRODUCTION

To improve disease treatment, understanding of disease mechanisms need to be fulfilled. However, it is difficult because of the complexity of complex disease. A way to get insight into the mechanisms is to find key genes that are important to the diseases. Disease gene identification has been widely studied to better understanding biological mechanisms for diseases and to improve medical treatments [1-5]. Many computational methods have been developed to identify disease-related genes and integrated several types of biological data, e.g., protein-protein interaction, gene expression data, functional annotation, sequence analysis features, text mining [6-9]. In addition, these approaches were based on concept of gene similarity. Genes, that show similar phenotypes, or are direct partners, are often shared similar functions [10, 11]. With this concept, novel disease genes were predicted by finding the similarity between a candidate gene and a known disease gene. The similarity can be computed from several types of biological data.

The standard information used to compute the similarity is the topological features from biological network. Several network centrality measures, such as degree, hub, betweenness, closeness, and eigenvector centrality metrics, were calculated in the biological network. Candidate genes that show similar values in term of network centrality as the known disease genes might be considered as promising targets for the disease. However, a large number of candidate genes have been predicted. To find only relevant genes, prioritization methods were developed to rank these candidate genes and select only high ranking genes to be promising targets. Tranchevent et al. [12] and Aerts et al. [13] developed a gene prioritization method based on the similarity concept using more than 10 genomic data source. Candidate genes were ranked independently according to the separate data source and finally these rankings were combined into a final ranking. Some method used gene information and functional relationship between genes to reconstruct a functional network to rank positional candidate genes [8].

Instead of using simple nearest neighbors, Kohler et al. [14] used global network distance measure, random walk analysis to define similarities in protein-protein interaction network to prioritize candidate disease genes. Shi et al. [15] used random walk analysis to find functional association between microRNA targets and disease genes in protein-protein interaction network and constructed bipartite miRNA-disease network to analyze the properties of miRNA regulation of the disease gene. Li et al. [16] developed a framework to construct disease-specific drug-protein interactions by integrating gene/proteins and drug connectivity from protein interaction network and text mining technique. The initial seed proteins were improved by expanding using ranked protein interaction data in the online predicted human interaction database and a developed nearest-neighbor protein interaction expansion method. The expanded proteins were ranked based on a scoring model. Genome-wide association study (GWAS) discovers correlation between genetic variants and diseases or traits [17, 18]. Several studies prioritized candidate genes generated from GWAS data [5, 19-21]. Ballouz et al. [5] used domain-based sequence homology analysis to infer function of genes and used candidate gene from GWAS. These studies used standard and developed features from several genomic data source to rank candidate disease genes obtained from GWAS or experimental results. However, the features might be irrelevant to the specific disease identification.

Several methods have been developed to find a set of relevant genes for a specific disease. However, some methods were complicated and difficult to implement. In addition, the ranked results from several methods seem to be redundant. In this study, we developed a simple analysis using a standard network topology features to analyze the features and diseases. Our prioritization analysis based on the protein-protein interaction network and GWAS data to investigate the relationship of the isolated standard centrality features and disease genes. The analysis was performed with each disease separately. This analysis is useful to find out relevant features for a specific disease. It could be a good source of guidance that could further improve classification performance to obtain accurate disease gene.

2. MATERIAL AND METHOD

To investigate the relationship of centralities and disease genes, the sets of proteins related to a disease were prepared and the protein-protein interaction network was reconstructed. Then the centrality measures were computed in the network. In this section, we described the data source, centrality measures, ranking processes, and performance measurement.

2.1 Data source and gold standard

Disease genes were obtained from GWAS catalog (http://www.genome.gov/ gwastudies/) [22] and used as gold standard. Only diseases containing more than 50 related genes were investigated in our study. In this study, protein-protein interaction network was reconstructed from STRING database version 9.05 [23]. Highly scored interactions (greater than 900) from the database were selected to obtain only reliable interactions in our analysis. The network consists of nodes representing proteins and edges representing interactions between two proteins.

2.2 Centrality measures

Network topology properties were calculated in this reconstructed protein-protein interaction network. The network topology described the relationship of a protein and other proteins in the network. We investigated four standard topology features consisting of degree, closeness, betweenness, and Kleinberg’s centrality. Note that in this study we named Kleinberg’s centrality as hub. Several researches have demonstrated that these centralities are related to the essentiality of protein in networks [24-26]. The definitions of these centralities are as follows:

An undirected graph $G = (V, E)$ consists of a set of nodes V and a set of edges E. Each node i, $j \in V$ represents a protein, and each edge $e(i, j) \in E$ represents an interaction between two proteins i and j. Let $\mathbf{A}$ be the adjacency matrix of the network. Thus $A(i, j) = 1$ when there is a connecting edge between node i and node j, and $A(i, j) = 0$ otherwise. The degree centrality $C_D(i)$ of node i is the number of its incident edges and is given by

$$C_D(i) = \sum_j A(i, j) \qquad (1).$$

The closeness centrality $C_C(i)$ for a node i is given by

$$C_C(i) = \frac{N-1}{\sum_j d(i,j)} \qquad (2),$$

in which $d(i,j)$ is the shortest distance from node i to node j, and N is the number of nodes in the network.

The betweenness centrality (C_B) measures the frequency that a node is in the shortest part of all of the pairs of nodes. The C_B is given by

$$C_B(k) = \sum_i \sum_j \frac{\delta(i,k,j)}{\delta(i,j)},\ i \neq j \neq k \qquad (3),$$

in which $\delta(i,j)$ denotes the total number of the shortest paths between i and j, and $\delta(i,k,j)$ denotes the total number of the shortest paths between i and j that pass through k. The sum is composed of all of the pairs (i, j) of nodes in the network.

The Kleinberg's centrality, or hub, computes the principal eigenvector of **A*t(A)**, where **t(A)** denotes the transposition of matrix **A**. The i^{th} component of the principal eigenvector is defined as Kleinberg's centrality $C_K(i)$ of node i. The eigenvector centrality is based on the assumption that an important node is usually connected to important neighbors. Therefore, each node's centrality is determined by the centrality values of the neighboring nodes. This value is higher if the node is connected to high-scoring nodes.

2.3 Ranking of centrality score

Each protein positioned in the protein-protein interaction network was calculated the centrality values according to the definition described in previous section. Therefore, a list of each centrality measure and corresponding proteins were constructed. These values were ranked in descending order. It means that the higher value of centrality measure, the more important is the protein in protein-protein interaction network. The centrality values of the four centrality measures were ranked independently. These centrality values were used as features of each protein. For a disease, a protein was labeled as 1 if it is in a protein set of the considered disease defined by GWAS, while another protein that is not a member in any disease gene sets was labeled as -1.

Table 1: List of 46 groups of diseases which have disease genes members more than 50 proteins.

Number	Disease name
1	Systemic lupus erythematosus
2	Multiple sclerosis
3	Cholesterol
4	Red blood cell traits
5	Vitiligo
6	Type 1 diabetes
7	Attention deficit hyperactivity disorder
8	Prostate cancer (gene x gene interaction)
9	Breast size
10	Obesity-related traits
11	Inflammatory bowel disease
12	Type 2 diabetes
13	Blood pressure
14	Dental caries
15	Celiac disease
16	Height
17	Asthma
18	Metabolite levels
19	Ulcerative colitis
20	Visceral adipose tissue/subcutaneous adipose tissue ratio
21	Bipolar disorder
22	Visceral adipose tissue adjusted for BMI
23	Bone mineral density
24	Body mass index
25	Major depressive disorder
26	HDL cholesterol
27	Prostate cancer
28	Triglycerides
29	LDL cholesterol
30	Schizophrenia
31	Chronic kidney disease
32	Protein quantitative trait loci
33	Coronary heart disease
34	Rheumatoid arthritis
35	Pulmonary function (interaction)
36	Cognitive performance
37	Breast cancer
38	Menarche (age at onset)
39	Platelet counts
40	Parkinson's disease
41	Response to statin therapy
42	Crohn's disease
43	Urate levels
44	Pulmonary function
45	Crohn's disease, Ulcerative colitis and IBD
46	Combined all disease genes

2.4 Performance measurement

With the four centrality measures and label sets of all proteins, classification of disease genes could be performed. Performance of the classification could be estimated by receiver operating characteristic (ROC) curve. The curve was depicted by the true positive rate (TPR) and false positive rate (FPR). The true positive rate is defined as follows:

$$TPR = \frac{TP}{TP + FN} \qquad (4),$$

where TP is the number of true positive and FN is the number of false positive. The false positive rate is defined as follows:

$$FPR = \frac{FP}{FP + TN} \qquad (5),$$

where FP is the number of false positive and TN is the number of true negative. At a cut-off of centrality values, TPR and FPR were calculated.

By varying the cut-offs through the centrality values, the ROC curve was generated. The area under the curve (AUC) was calculated and estimated as the performance of the disease gene classification. For a disease, the AUC of a centrality measure was calculated. The same scheme was applied to all centrality measures.

3. RESULT AND DISCUSSION

In this section, the results of our analysis are described. Table 1 shows the list of 46 groups of diseases that have disease gene in the group greater than 50 genes.

3.1 Three diseases obtaining high performance

Performance of disease gene classification for each disease was measured by the AUC. This calculation was done for each feature separately. Therefore, for a disease classification, we can compare these four standard centrality measures and observe which centrality measure is suitable for the specific disease. The same scheme was performed for all diseases with all different centralities. Selecting only high performance results with the AUC value greater than 0.7, we yielded 5 entries consisting of 3 different diseases, Celiac disease, Rheumatoid arthritis, and Inflammatory bowel disease. The results show in Table 2. Interestingly, 3 out of 5 centralities found in this selection was the closeness centrality and the other centralities were the hub and degree centrality. Celiac disease yielded good classification results with closeness, hub, and degree centrality. Network topologies might be suitable to indicate important proteins for Celiac disease. Figure 1, 2, and 3 illustrate the ROC curve of the Celiac disease, rheumatoid arthritis, and inflammatory bowel disease, respectively.

Table 2: The diseases and centrality measures showing high values of AUC (more than 0.7).

Disease	Centrality Measure	AUC
Celiac disease	Closeness	0.7386
Rheumatoid arthritis	Closeness	0.7248
Celiac disease	Hub	0.7152
Inflammatory bowel disease	Closeness	0.7111
Celiac disease	Degree	0.7015

3.2 Celiac disease and network topologies

We further investigated the relationship between all four network topologies and Celiac disease genes. As shown in Table 2, Celiac disease was presented with three centralities. Another centrality that missed in Table 2 for the disease is the betweenness centrality. Similarly, in Table 3, the classification result of the betweenness centrality showed a bit inferior AUC than 0.7 (the AUC of 0.69). To verify the result, we compared with a random technique. The random labels with the same number of positive and negative sets were assigned and the scheme for measuring performances was performed for all centrality measures. The results showed that, in random case, the AUCs were approximately close to 0.5. The complete results are shown in Table 3.

Table 3: AUC of Celiac disease with different centrality measures comparing with randomness.

Centrality Measures	Celiac disease	Randomness
Degree	0.7015	0.4951
Closeness	0.7386	0.5051
Betweenness	0.6906	0.5006
Hub	0.7152	0.4759

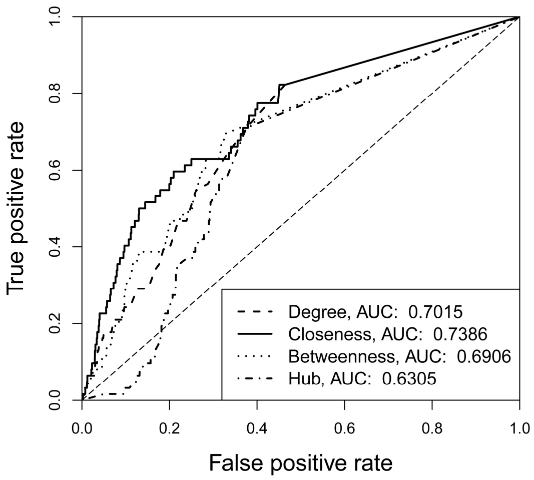

Figure 1: ROC of the Celiac disease with different centrality measures

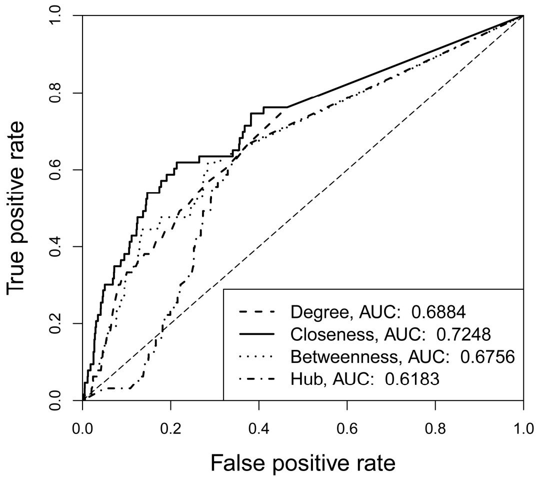

Figure 2: ROC of the Rheumatoid arthritis with different centrality measures

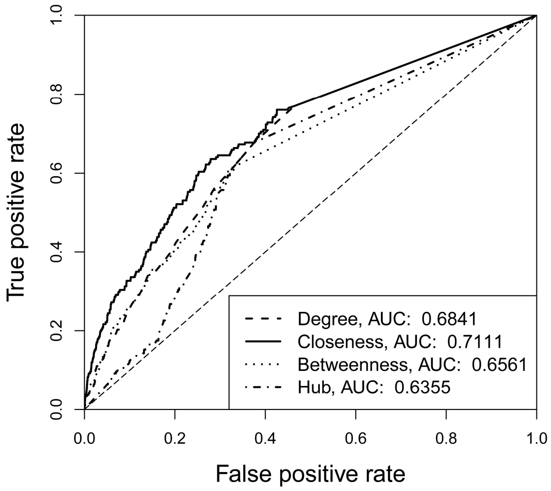

Figure 3: ROC of the inflammatory bowel disease with different centrality measures

3.3 Centrality evaluation

To find the best centrality measure, the relationship between these standard centralities and the 46 groups of diseases was observed. The fraction of diseases that have the AUC values of their disease gene classification with a centrality measure greater than a cut-off value was calculated. The percentages are shown in Table 4. We observed five cut-off values consisting of 0.5, 0.55, 0.60, 0.65, and 0.7 for all diseases and centrality measures. In the overall ranges of the AUC cut-off values, the closeness centrality yielded the best percentage comparing with other centrality measures. At an AUC cut-off value of 0.5, all centrality measures showed the same percentages of 95.65%. For a higher AUC cut-off value of 0.55, the closeness centrality showed the best percentage followed by the degree, betweenness, and Kleinberg's centrality, respectively. At the AUC cut-off value of 0.7, there are only 6.52% of diseases with the closeness centrality obtaining the AUC greater than the cut-off value, while the degree and Kleinberg's hub obtained the same percentage of 2.17 and there is no disease obtaining the AUC greater than the cut-off value with the betweenness centrality.

Table 4: The percentage of diseases in overall 46 groups of diseases with centrality measures. Diseases that have area under the curve higher than the cut-off value were counted.

Cut-off	Degree (%)	Closeness (%)	Betweenness (%)	Hub (%)
0.50	95.65	95.65	95.65	95.65
0.55	76.09	78.26	73.91	71.74
0.60	45.65	52.17	45.65	45.65
0.65	17.39	32.61	17.39	15.22
0.70	2.17	6.52	0.00	2.17

4. CONCLUSION AND DISCUSSION

Translating new genomic information to real medical treatment is a challenge task that needs to get insight into disease mechanisms [27]. Nowadays, the opportunities to understand a complete causal pathway of disease mechanisms are limited by the different combinations of multiple variants [4]. To increase such opportunities, several computational methods were developed to predict new disease genes using multiple data source. However, a number of candidate genes were proposed to concern in diseases in several studies. To obtain only promising candidate genes, the prioritization methods were applied. The prioritization can perform well if the scores or features for ranking are reasonable. In this study, we examined four different network topology features consisting of the degree, closeness, betweenness, and Kleinberg's centrality with 46 groups of diseases aggregated from GWAS databases. The centrality measures were calculated in the protein-protein interaction network. Performances of disease gene classification of a disease with different centrality features were measured. Selecting only high performance (AUC greater than 0.7), we yielded 5 entries of 3 diseases and 3 features. Celiac disease with closeness centrality showed the best performance. Investigating overall diseases in our system, we found that the closeness centrality was the best feature for the classification. The results showed that, using a stringent threshold of AUC, the closeness centrality still yielded a superior ability. This result might be a good suggestion to use this feature for disease prediction.

A limitation of this analysis is that only isolated feature was employed for analyzing a specific disease. Even the closeness showed the best result, the other features, e.g., the degree and betweenness centrality also yielded a good classification's performance. In many cases, using only single feature might be not enough to perform the classification. Fusion multiple performance features might be a good option to improve the classification performance that could be developed in further study. In conclusion, this study demonstrated a prioritization analysis to rank features that is valuable to perform to reduce irrelevant features. Therefore, the results from this analysis could be a good suggestion to select features for a disease-gene classification.

ACKNOWLEDGEMENT

This research is supported by the Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Thailand, under grant number: 5742104.

REFERENCES:

[1] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease", *Nature Genetics*, Vol. 33 Suppl, No. 2003, pp. 228-237.

[2] H. G. Brunner and M. A. van Driel, "From syndrome families to functional genomics", *Nature Reviews Genetics*, Vol. 5, No. 7, 2004, pp. 545-551.

[3] X. Xiao, J. L. Min, P. Wang and K. C. Chou, "Predict drug-protein interaction in cellular networking", *Current Topics in Medicinal Chemistry*, Vol. 13, No. 14, 2013, pp. 1707-1712.

[4] A. C. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects", *Human Molecular Genetics*, Vol. 17, No. R2, 2008, pp. R166-173.

[5] S. Ballouz, J. Y. Liu, M. Oti, B. Gaeta, D. Fatkin, M. Bahlo and M. A. Wouters, "Candidate disease gene prediction using Gentrepid: application to a genome-wide association study on coronary artery disease", *Molecular Genetics & Genomic Medicine*, Vol. 2, No. 1, 2013, pp. 44-57.

[6] L. Miozzi, R. M. Piro, F. Rosa, U. Ala, L. Silengo, F. Di Cunto and P. Provero, "Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data", *PLoS One*, Vol. 3, No. 6, 2008, pp. e2439.

[7] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard, "Speeding disease gene discovery by sequence based candidate prioritization", *BMC Bioinformatics*, Vol. 6, No. 2005, pp. 55.

[8] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes", *Am J Hum Genet*, Vol. 78, No. 6, 2006, pp. 1011-1025.

[9] H. Al-Mubaid and R. K. Singh, "A text-mining technique for extracting gene-disease associations from the biomedical literature", *Int J Bioinform Res Appl*, Vol. 6, No. 3, 2010, pp. 270-286.

[10] H. N. Chua, W. K. Sung and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions", *Bioinformatics*, Vol. 22, No. 13, 2006, pp. 1623-1630.

[11] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami and T. Takagi, "Assessment of prediction accuracy of protein function from protein--protein interaction data", *Yeast*, Vol. 18, No. 6, 2001, pp. 523-531.

[12] L. C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts and Y. Moreau, “ENDEAVOUR update: a web resource for gene prioritization in multiple species”, *Nucleic Acids Res*, Vol. 36, No. Web Server issue, 2008, pp. W377-384.

[13] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet and Y. Moreau, “Gene prioritization through genomic data fusion”, *Nat Biotechnol*, Vol. 24, No. 5, 2006, pp. 537-544.

[14] S. Kohler, S. Bauer, D. Horn and P. N. Robinson, “Walking the interactome for prioritization of candidate disease genes”, *Am J Hum Genet*, Vol. 82, No. 4, 2008, pp. 949-958.

[15] H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, Z. Zhao, W. Jiang, Z. Guo and X. Li, “Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes”, *BMC Syst Biol*, Vol. 7, No. 2013, pp. 101.

[16] J. Li, X. Zhu and J. Y. Chen, “Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts”, *PLoS Comput Biol*, Vol. 5, No. 7, 2009, pp. e1000450.

[17] J. N. Hirschhorn and G. Lettre, “Progress in genome-wide association studies of human height”, *Horm Res*, Vol. 71 Suppl 2, No. 2009, pp. 5-13.

[18] G. Lettre and J. D. Rioux, “Autoimmune diseases: insights from genome-wide association studies”, *Human Molecular Genetics*, Vol. 17, No. R2, 2008, pp. R116-121.

[19] M. J. Li, P. C. Sham and J. Wang, “Genetic variant representation, annotation and prioritization in the post-GWAS era”, *Cell Res*, Vol. 22, No. 10, 2012, pp. 1505-1508.

[20] K. Wang, M. Li and H. Hakonarson, “Analysing biological pathways in genome-wide association studies”, *Nat Rev Genet*, Vol. 11, No. 12, 2010, pp. 843-854.

[21] P. Holmans, E. K. Green, J. S. Pahwa, M. A. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan and N. Craddock, “Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder”, *Am J Hum Genet*, Vol. 85, No. 1, 2009, pp. 13-24.

[22] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”, *Proc Natl Acad Sci U S A*, Vol. 106, No. 23, 2009, pp. 9362-9367.

[23] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. von Mering, “The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored”, *Nucleic Acids Res*, Vol. 39, No. Database issue, pp. D561-568.

[24] K. Park and D. Kim, “Localized network centrality and essentiality in the yeast-protein interaction network”, *Proteomics*, Vol. 9, No. 22, 2009, pp. 5143-5154.

[25] G. del Rio, D. Koschutzki and G. Coello, “How to identify essential genes from molecular networks?” *BMC Syst Biol*, Vol. 3, No. 2009, pp. 102.

[26] K. Plaimas, R. Eils and R. Konig, “Identifying essential genes in bacterial metabolic networks with machine learning methods”, *BMC Syst Biol*, Vol. 4, No. 2010, pp. 56.

[27] M. J. Khoury, M. Gwinn, P. W. Yoon, N. Dowling, C. A. Moore and L. Bradley, “The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?” *Genet Med*, Vol. 9, No. 10, 2007, pp. 665-674.