Original research **Journal of Medical Physiology. 2019; 4: e8.**

Gene network analysis using measurement indices in microarray data

Atefeh Talebi[1], Seyyed Mohammad Tabatabaei[2], Nasibeh Khayer[3], Mazaher Azizpour[4], Abolfazl Akbari[1], Atieh Khaleghi[5], Hamid Alavi Majd[6*]

1. Colorectal Research Center, Iran University of Medical Sciences, Tehran, Iran.
2. Imam Reza Hospital Clinical Research Unit, Mashhad University of Medical Sciences, Mashhad, Iran.
3. Skull Base Research Center, Hazrat Rasoul Hospital, the Five Senses Institute, Iran University of Medical Sciences, Tehran, Iran.
4. Department of Orthopedic Surgery, Aalborg University Hospital, Aalborg, Denmark.
5. Master Student of Statistics and Machine Learning, Liköping University, Liköping, Sweden.
6. Department of Biostatistics, School of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Received: October 2019; Accepted: November 2019

Abstract:

Background: A great number of data mining methods have been widely made such as gene regulatory networks and gene set analyses to connect genes that reveal similar expression patterns. These methods generally fail to unveil gene-gene interactions in the same cluster. The aim of this study is to use several nonparametric correlation coefficient methods to transform the linear rank statistics into distance metrics on a Saccharomyces cerevisiae data set.

Methods: These nonparametric correlation coefficients, Kendall's tau index and Gini rank correlation, were compared with common Pearson correlation method. The reliability and advantages of our proposed is satisfied using genetic website, http://www.yeast genome .org/. To address the interactions and characterize the gene–gene biological processes explicitly, the gene relationships are shown as a Pajek graph topology.

Result: The results of biological interactions and characteristics demonstrated that the proposed nonparametric correlation coefficient methods have a strong capability to identify interaction genes. Moreover, suggested techniques could accurately detect the main genes and functional interactions in comparison to generally used Pearson correlation coefficient.

Conclusion: The two non-linear correlation coefficient techniques are proposed to measure the gene interactions more precisely.

Keywords: Gene-Gene Interaction; Gini Index; Kendall's tau; Microarray Data

Cite this article as: Talebi A, Tabatabaei SM, Khayer N, Azizpour M, Akbari A, Khaleghi A, Alavi Majd H. Gene network analysis using measurement indices in microarray data. J Med Physiol. 2019; 4: e8.

1. Introduction

In recent years, high-throughput screening techniques has been analyzed the vast amount of microarray data (1). A large number of practical approaches have been done to assess the relationship between genes (2). The cluster and gene set analyses are the most frequent method in such data. A clustering algorithm group the data into classes or clusters with the same characteristics. The clustering algorithms identify groups of objects, or clusters that are more similar to each other than to other clusters (3). Also, probabilistic metric, known as hierarchical, provided computational analysis on microarray data (4). Although, clustering methods cannot identify analyze high level functions or molecular networks, the clustering methods do not take into account the relationships between genes within each cluster and those across different clusters (5-6). Thus, there is a great requirement for new algorithms for gene interactions and pathway-based analyses of gene expression data.

***Corresponding author:** Hamid Alavi majd, Department of Biostatistics, School of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran; Tel/Fax: +982122707347; E-mail: alavimajd@gmail.com

A genome-wide in vivo screen for protein-protein interactions in the Saccharomyces cerevisiae data set has been applied (7-10). Moreover, some non-linear methods have already been made, including support vector machines (SVM) (11), kernel correlation model heuristics (KCH) (11, 12), and non-linear kernel correlation coefficient (KCC) (13). A new multivariate dependence coefficient has been introduced that measures all types of dependence between random vectors in arbitrary dimensions (14). The advantages of these methods are their nonlinear and nonmonotone dependence. Reshef et al. (2011) presented a measure of dependence, which is the maximal information coefficient (MIC) (15). Some other techniques have been made to find gene interaction networks of gene expression data; such as Boolean networks (16), Bayesian networks (17) and graphical Gaussian model (6).

The definition of measure pair-wise dependence between genes is so significant in gene interaction networks; moreover, similarity in constructing the gene networks (18, 19). Generally, the linear similarity indices, such as the Pearson correlation coefficient and the Euclidean distance, are applied in an ad hoc procedure (20). However, when multiple complex gene interactions are present, the relationship may be nonlinear, and therefore a linear metric and normal distribution, such as the Pearson correlation coefficient, may not be accurate enough in describing the gene interaction relationships. As nonparametric correlation coefficients do not require specific distribution and these methods are robust to outliers, unlike Pearson's correlation coefficient which is sensitive to outliers, we used nonparametric correlation coefficients. Moreover, due to the large number of microarray data, linear methods are not able to identify the relationship between genes, so the mutual information criterion is known as a nonlinear and useful criterion for these data. In this present study, the Gini index and the Kendall rank correlation, known as nonparametric linear correlation coefficient methods, are first performed to find strength and direction of association that exists between two variables; next the matrices of nonparametric correlations across the microarray data are calculated to display the pair-wise interaction of genes. Then, a topological graph is made to capture the gene interactions using Pajek software. Furthermore, the reliability of their relationships is assessed by comparing with the Pearson correlation coefficient and also Saccharomyces Genome Database website related with gene networks (http://www.yeastgenome.org/).

2. Method

2.1. Microarray data set

The download of Saccharomyces cerevisiae data was performed using the NCBI database that considered as chips, probes, and offspring samples [GEO: GDS1115] (13, 21, 22). This data set comprised 113 chip expression values and 6229 genes; moreover, missing values of the data set included 27633 (3.93%) that K-nearest neighbors algorithm (k=10) was applied to impute the missing data (23). Our goal was to follow a more accurate inference extraction, corresponding to the underlying principle of rank relationships, as opposed to a common linear measure for the Saccharomyces cerevisiae data. Level of significance for statistical analysis was 0.05. The R version 3.1.2 software was done to statistical methods, and the topological graph network was drawn using Pajek version 4.

2.2. Conception of network

The interaction of a network was defined as a set of vertices, nodes, together with a set of edges, links, that connect various pairs of vertices. In this case, nodes reveal genes or proteins, and edges show interactions (1). The degree of a node refers to the number of edges (links) to which it is connected. A clique is a set of three or more vertices in which each vertex is directly connected to all other vertices. The size of a clique is the number of its vertices, and it is the strictest structural form of a cohesive subgroup (24-25). It can be inferred that stronger links between vertices demonstrate greater dominance where there is biological significance (13). The function of the node degree follows a descending trend or a power-law function; that is, there are a small number of high degree nodes (hubs), although a lot of nodes have only a few relationships (26). A general property of large networks is that their vertex relationships have a scale-free power-law distribution. This characteristic was found to be a result of two generic approaches: (i) networks become continuously larger by the addition of new vertices, and (ii) new vertices fix preferentially to places that are already well connected (27- 29). The power-law function is the probability P(x) of a gene interacting with x other genes, which roughly decrease with increasing x, according to the equation: $P(x) = \frac{k}{x^{-\gamma}}$

where, in a biological network, $2<\gamma<3$, k is a constant.

The following techniques were applied as nonparametric methods to make the gene networks.

2.3. Correlation coefficient methods

Pearson correlation coefficient, which assumes a linear relationship between two random variables, is applied to measure the correlation between continues variables. The Pearson correlation coefficient is calculated from the samples X= (X1,…, Xn) and Y= (Y1,…,Yn) of two variables (e.g. genes in genomics):

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (1)$$

To test the null-hypotheses H0: $\rho=0$ versus H1: $\rho\neq0$, one would then reject H0, when r is far from zero, through the quantity T which is defined as:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \qquad (2)$$

As the Pearson correlation coefficient is a linear estimator, it is not applied to estimate a nonlinear relationship. Instead, we focused on measures of dependence based on ranks. Several nonlinear correlation coefficients are based on ranks. Ranks have a number of desirable properties: they are invariant under monotonic transformations of the individual variables; moreover, robust towards outliers and unexpected observations. They are able to discover not only the linear relationships, but also any kind of monotone relation, without making any extra assumptions on the distributions of variables (30).

The Kendall's tau, τ, between two gene vectors X= (X1,…, Xn) and Y= (Y1,…,Yn) for any pair of observations and are provided to be concordant if Xi< Xj and Yi<Yj or if Xi> Xj and Yi>Yj. They are called discordant, if Xi> Xj and Yi<Yj or if Xi< Xj and Yi>Yj. The Kendall's tau is calculated as:

$$t = \frac{c-d}{c+d} \qquad (3)$$

where c shows the number of concordant pairs and d the number of discordant pairs.

The Gini index, g, between two gene vectors (X1,…, Xn) and Y= (Y1,…,Yn) with the respective ranks (R1,…, Rn) and (S1,…,Sn) is calculated as:

$$g = \frac{1}{\left[\frac{n^2}{2}\right]} \sum_{i=1}^{n} \{|R_i + S_i - n - 1| - |R_i - S_i|\} \qquad (4)$$

2.4. Statistical methods

The pair-wise correlation coefficients ((6229¦2)=19397106 pair-wise) were calculated by nonparametric rank correlation coefficients [equations (1,3,4)] in yeast genes. Moreover, these methods were compared with website, http://www.yeastgenome.org. The pair-wise interaction gene networks for each method are represented graphically using the Pajek software. Also, R software was applied to statistical analysis. The nonparametric rank correlation coefficients and the Pearson correlation coefficient graphs were plotted based on a threshold of 0.8 corresponding to a p-value < 0.05.

3. Result

The pair-wise gene interaction network is equivalent to the threshold 0.8. Figure 1 represents the gene networks produced using the Kendall's tau rank correlation, Gini index, and the Pearson correlation coefficient. As shown in Figure 1, the topologies of the gene networks are the same, with all of them having three cliques. According to Figure 1A indicates that the Kendall's tau rank correlation consists of 1066 genes in an interaction network and 6899 gene pairs. The Kendall's tau interaction gene network is composed of 54 negative correlations and 7057 positive correlations. The network average degree is 12.6, and the node maximum degree is 91. Figure 1B demonstrates that the Gini index consists of 1054 genes in an interaction network and 5985 gene pairs. The Gini index interaction gene network is composed of 81 negative correlations and 5701 positive correlations. The network average degree is 11.4, and the node maximum degree is 66. Figure 1C, the Pearson correlation coefficient consists of 1136 genes in an interaction network and 8672 gene pairs. The Pearson interaction gene network is composed of 94 negative correlations and 8578 positive correlations. The network average degree is 15.27, and the node maximum degree is 116.

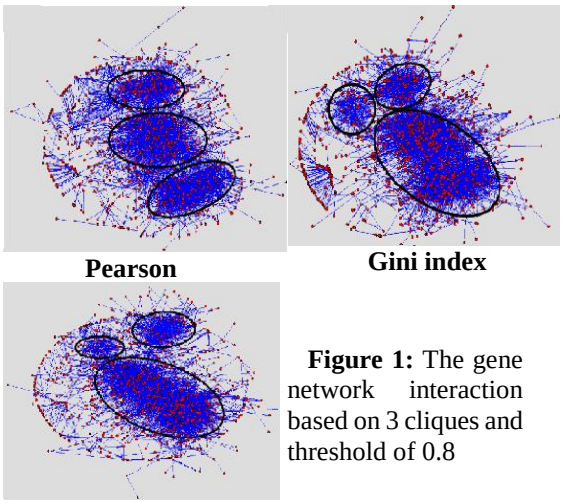

Figure 1: The gene network interaction based on 3 cliques and threshold of 0.8

As can be seen, for nonparametric correlation coefficients, the network average degrees and node maximum degrees are less than those for the Pearson correlation coefficient, which indicates a high interrelationship among the genes. Also, all Figures show that the correlation coefficients consist of both positive and negative values. The number of negative correlations in the Pearson correlation coefficient is greater than in the nonparametric methods.

The network topological characteristics are given in Table 1, where these characteristics are explained in detail.

Figure 2 illustrates the scatter plot of the most important gene pairs, where the Pearson correlation coefficient is smaller than the nonparametric correlations. As shown in Figure 2, there are noisy experimental data, and points of special significance, that greatly influence the p-value of the Pearson correlation coefficient (this topic will be examined in future work). On the other hand, nonparametric correlations are able to reveal the linear relationships quite well. For example, in Figure 2, YDR060W and YKL172W, the value of the Pearson correlation coefficient is only 0.57, whereas the other correlations are greater than 0.81.

Fig. 3 displays the power-law function and regression line of the Pearson and the other nonparametric correlation coefficients for a threshold equal to 0.8. The value (γ) shows that the gene interaction networks of nonparametric

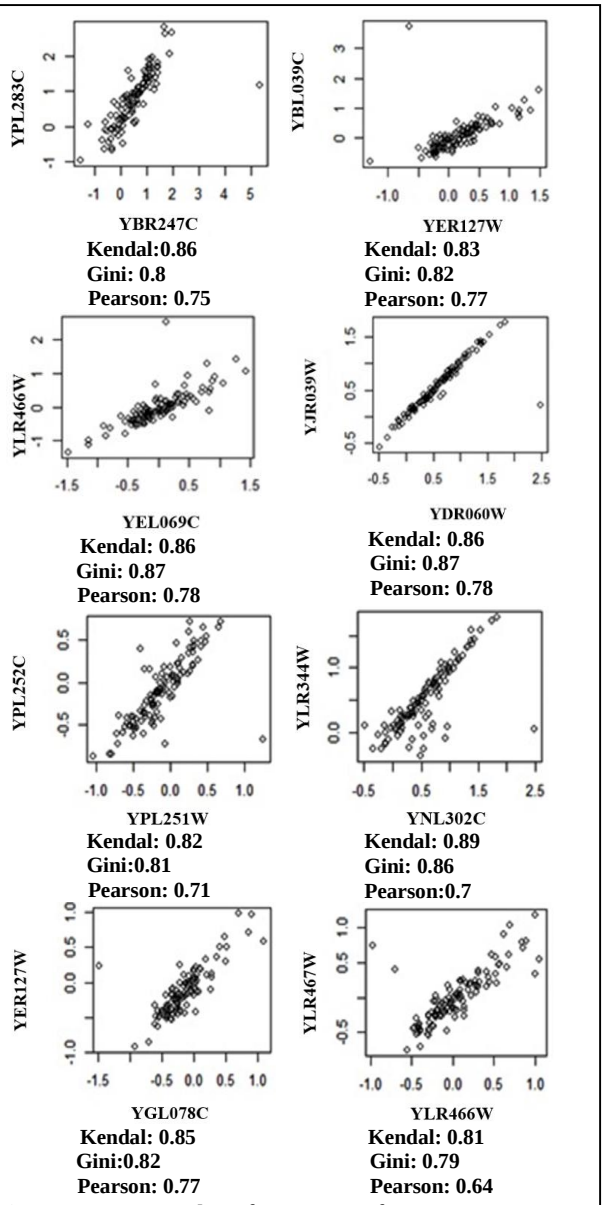

Figure 2: Scatter plot of some significant pair genes.

correlations are similar in degree of distribution to the Pearson correlation coefficient, and that they are scale-free. Table 2 shows the nine-degree genes' biological interactions and their characteristics, taken from the http://www.yeastgenome.org/ website. It shows that they are related to a certain protein component of the ribosomal subunit (large or small). They charge the core transcriptions along with the other genes, and they also have large interactions with other genes. Other approaches are introduced on the website, based on the relationships between genes, such as examining the chromosome sequence and the protein products.
Table 3 shows the maximum nine-degree genes.

4. Discussion

A great number of gene products are recognized to treat in a highly modified method (20). Several non-linear

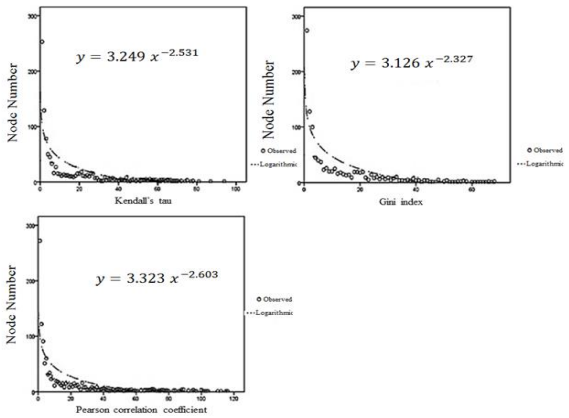

Figure 3: The gene network degree distributions.

correlation approaches are applied to identify the gene expression characteristics (20). In this study, various nonparametric methods have been performed to construct the gene networks in Saccharomyces cerevisiae data set. To address the interactions and characterize the gene–gene biological processes explicitly, the gene relationships are displayed as a graph topology. Using a yeast gene relevance experiment we have compared the Pearson correlation with the nonparametric correlation coefficients and have verified experimentally that the nonparametric methods behave more precisely. Additionally, these methods were compared with genetic website. The results show that the nonparametric methods can promote and enhance the gene interaction prediction accuracy quite significantly. The aim of the study is to identify more precise inference methods, equivalent to the underlying fundamentals of rank measures, which could be compared with other linear relationships in the gene expression data set. A higher correlation coefficients were observed in nonparametric methods than linear approach. These results propose that the nonparametric methods have strong capability in identifying interaction genes, and also that the proposed methods can discover accurately the key genes and functional interactions, cliques, as compared to the frequently used Pearson correlation.

Table 1: The network topological characteristics base on the threshold 0.8 ($p<0.05$)

Correlation type	Pearson correlation	Kendal's tau	Gini index
Number of Genes	1136	1066	1054
Pair-wise Genes	8672	6899	5985
Clique Genes	3	3	3
Network average degree	15.27	12.6	11.4
Node maximum degree	116	91	66

Table 2: The characteristics of ten genes based on biological interactions

Gene name	Biological interactions	Biological description
YHL033C	240 total interactions for 189 unique genes characteristics	Ribosomal protein of the large subunit
YER074W	203 total interactions for 192 unique genes characteristics	Ribosomal protein of the large subunit
YDR447C	179 total interactions for 160 unique genes characteristics	Ribosomal protein of the large subunit
YPL079W	113 total interactions for 84 unique genes characteristics	Ribosomal protein of the large subunit
YGL031C	136 total interactions for 124 unique genes characteristics	Ribosomal protein of the large subunit
YNL162W	30 total interactions for 29 unique genes characteristics	Ribosomal protein of the large subunit
YHR141C	64 total interactions for 63 unique genes characteristics	Ribosomal protein of the large subunit
YML024W	83 total interactions for 73 unique genes characteristics	Ribosomal protein of the large subunit
YML100W	55 total interactions for 34 unique genes characteristics	Ribosomal protein of the large subunit

Yu et. al have designed network maps included that several hundred molecular complexes with limited and binary interactions. Their method showed the relationship between YPL252C and YPL251W (20). Tarassovet. al surveyed a new approach that confirmed the relationship between YDR060W and YKL172W was 0.81. They investigated a Vivo map of the yeast protein interactome (20). Collins et. al presented a new metric for protein-protein interactions and find the strong relationship between YDR060W and YKL172W (20). Cheng et. al proposed Kernel correlation coefficient method to find gene-gene interactions in Saccharomyces cerevisiae. They revealed that proposed Kernel correlation coefficient measure has a strong capability to recognize gene interactions. The relationship between YJR039W and YEL069C was 0.81 (20).

Table 3: Maximum nine degree genes based on correlation coefficient methods

Pearson		Kendall		Gini	
YPL142C	116	*YML100W*	70	*YDL063C*	94
YGL031C	114	*YPL142C*	68	*YGL031C*	90
YDL082W	111	*YHL033C*	67	*YPL142C*	81
YDR447C	108	*YLR178C*	66	*YDL082W*	79
YER074W	104	*YER074W*	65	*YNL162W*	78
YML024W	103	*YMR250W*	65	*YDL060W*	77
YOR234C	102	*YML024W*	64	*YHL033C*	75
YHL033C	101	*YDR447C*	63	*YHR170W*	74
YPL081W	100	*YGL031C*	61	*YPL126W*	74

5. Conclusion

It is concluded that based on biological interactions and characteristics, the nonparametric correlation coefficient methods have a strong capability to identify interaction genes than Pearson correlation coefficient.

6. Acknowledgment

We thank the patients who participated in the study.

7. Conflict of interest

No conflict of interest was declared.

8. Funding source

This work was supported by Iran University of Medical Sciences.

9. Author contribution

H.A and, A.T contributed to study design. A.T, and S.T contributed to the selection of patients and data gathering. N.K performed data analysis. A.A write the manuscript. A.T, S.T and NK edit the manuscript.

10. References

1. Dehmer M, Emmert-Streib F, Graber A, Salvador A. Applied Statistics for network biology: Wiley Online Library; 2011.
2. Gao W, Wu H, Siddiqui MK, Baig AQ. Study of biological networks using graph theory. Saudi J Biol Sci. 2018;25(6):1212-9.
3. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 2000; 418-29.
4. Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FC, et al. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics. 2007;6(3):439-50.
5. Markowetz F, Troyanskaya OG. Computational identification of cellular networks and pathways. Mol Biosyst. 2007;3(7):478-82.
6. Wu X, Ye Y, Subramanian KR, editors. Interactive analysis of gene interactions using graphical Gaussian model. Proceedings of the 3rd International Conference on Data Mining in Bioinformatics; 2003:

Springer-Verlag. 2003. 63-69.
7. Emmert-Streib F, Glazko G, De Matos Simoes R. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front Genet. 2012;3:8.
8. Santra T. A bayesian framework that integrates heterogeneous data for inferring gene regulatory networks. Front Bioeng Biotechnol. 2014;2:13.
9. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, et al. An in vivo map of the yeast protein interactome. Science. 2008;320(5882):1465-70.
10. Ud-Dean SM, Gunawan R. Ensemble inference and inferability of gene regulatory networks. PLoS One. 2014;9(8):e103812.
11. Chen P. A novel kernel correlation model with the correspondence estimation. J Math Imaging Vis. 2011;39(2):100-20.
12. Land WH, Qiao X, Margolis DE, Ford WS, Paquette CT, Perez-Rogers JF, et al. Kernelized Partial Least Squares for feature reduction and classification of gene microarray data. BMC Syst Biol. 2011;5(3):S13.
13. Cheng L, Khorasani K, Ding Y, Guo X. Gene interaction networks based on kernel correlation metrics. Int J Comput Biol Drug Des. 2013;6(1-2):72-92.
14. Székely GJ, Rizzo ML. Brownian distance covariance. Ann Appl Stat. 2009;3(4):1236-65.
15. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. science. 2011;334(6062):1518-24.
16. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Biocomputing'99: World Scientific; 1999. p. 17-28.
17. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7(3-4):601-20.
18. Khayer N, Mirzaie M, Marashi S-A, Rezaei-Tavirani M, Goshadrou F. Three-way interaction model with switching mechanism as an effective strategy for tracing functionally-related genes. Expert Rev Proteomics. 2019;16(2):161-9.
19. Khayer N, Marashi S-A, Mirzaie M, Goshadrou F. Three-way interaction model to trace the mechanisms involved in Alzheimer's disease transgenic mice. PloS one. 2017;12(9):e0184697.
20. Cho Hc, Hadjiiski L, Sahiner B, Chan HP, Helvie M, Paramagul C, et al. Similarity evaluation in a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images. Med Phys. 2011;38(4):1820-31.
21. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci USA. 2005;102(5):1572-7.
22. Storey J, Tibshirani R. Statistical significance for genomewide studies Proc Natl Acad Sci USA. 2003;100:9440-9445.
23. Nguyen DV, Wang N, Carroll RJ. Evaluation of missing value estimation for microarray data. J Data Sci. 2004;2(4):347-70.
24. De Nooy W, Mrvar A, Batagelj V. Exploratory social network analysis with Pajek: Revised and expanded edition for updated software: Cambridge University Press; 2018.
25. Hudson NJ, Dalrymple BP, Reverter A. Beyond differential expression: the quest for causal mutations and effector molecules. BMC Genomics. 2012;13(1):356.
26. Shipley B. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R: Cambridge University Press; 2016.
27. Romano JP, Lehmann E. Testing statistical hypotheses. Springer New York, NY, USA; 2005.
28. Taylor R. Interpretation of the correlation coefficient: a basic review. J Diagn Med Sonogr. 1990;6(1):35-9.
29. Alavi Majd H, Talebi A, Gilany K, Khayyer N. Nonparametric correlation coefficient methods in gene interactions from microarray data. Int J Analyt, Pharmaceutic Biomed Sci. 2015;4(1):38-46.
30. Alavi Majd H, Talebi A, Gilany K, Tabatabaei SM. Nonparametric correlation coefficient methods constructing three-way gene interactions in microarray data. Int J Analyt, Pharmaceutic Biomed Sci. 2015;4(1): 65-70.
31. Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, Sahalie J, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008;322(5898):104-10.
32. Majd HA, Talebi A, Gilany K, Khayyer N. Two-Way Gene Interaction From Microarray Data Based on Correlation Methods. Iran Red Crescent Med J. 2016;18(6).