Time Series Gene Expression Data Clustering and Pattern Extraction in *Arabidopsis thaliana* Phosphatase-encoding Genes

Pooya Sobhe Bidari, *Student Member, IEEE,* Roozbeh Manshaei, *Student Member, IEEE,*
Tahmineh Lohrasebi, Amir Feizi, Mohammad Ali Malboobi, Javad Alirezaie, *Senior Member, IEEE*

***Abstract*— Clustering of genes using their expression data has been a major topic in recent years. A large amount of gene expression data even in time series are obtained by microarray technology. Finding gene clusters with similar functions and interconnecting genes by networks has an important role in mining biological gene functional analysis. In this paper, Two Phase Functional Clustering has been presented as a new approach in gene clustering. The proposed approach is based on finding functional patterns of time series gene expression data by Fuzzy C-Means (FCM) and K-means methods. The gene function similarities over a number of experimental conditions are extracted using Pearson correlation between expression patterns of genes. This leads to visualize genes interconnections.**

I. Introduction

In recent years, gene expression data is widely used to discover genes behavior under different conditions. Many experiments are performed to increase the available time series gene expression data sets. By using computational methods, these data sets can be employed for extracting useful information such as gene expression patterns and interactions among genes.

Clustering, classification and function prediction are data mining techniques that are vastly applied in gene expression data analysis. Clustering algorithms used for gene expression data analysis include K-means clustering [10], hierarchical clustering [11], [12], and clique clustering [13]. The literatures concerning the clustering of gene expression data are mainly separated into two groups: identification of genes with similar expression patterns and comparing expression profiles of samples [1].

Methods developed for the first purpose provide some clusters including genes with similar functions over a specific experimental condition. The underlying assumption in clustering gene expression data is that co-expression indicates co-regulation [2]. By using time series gene expression data sets, more thorough clustering analyses can be obtained. In the second type of clustering, samples with more similarity are grouped together. Here, the features are the genes and the samples are the experiments. An example of this type is diagnosing AML/ALL leukemia cancer from a gene expression profile of a patient [1] in which gene expression values over some experiments are considered as features and the patient as the sample.

In this paper, a new approach is suggested for clustering of genes under several experimental conditions. Here, the principle goal of clustering is to find genes with similar behavior in more than one condition which will lead to groups of genes probably with similar genetic functions in the cell. The novel system is designed for clustering and cluster analysis of expression data of 54 genes in *Arabidopsis thaliana* which are expected to be involved in Phosphates control process of the plant in the root and the shoot [7]. The data set used for the clustering includes 7 time points of the expression values of the genes in 10 experimental conditions: Pi-/Pi+, heat, osmotic, salt, drought, genotoxic, oxidative, UV-B, wounding and cold stresses [17], [18]. In an overall view, the algorithm works in two phases:

In the first phase, the genes are clustered to 7, 8 or 9 groups in any experimental conditions by K-means and FCM algorithms. Then the groups are described by some gene expression patterns which define the behavior of genes in each cluster. In addition to finding the patterns of the clusters, the time series expressions of genes in each cluster are compared together by calculating the Pearson correlation. The correlations indicate similarity of genes from a functional aspect. In each cluster, the genes with functional similarity of more than a specific threshold are extracted; these genes are more probably participating in a specific genetic function through the related experiment condition.

After finding patterns for all of the genes, the second phase is performing a new clustering in which the features are the patterns extracted from the last phase and the samples are the genes. Finding the clusters in this phase is done by using Pearson correlation, so any cluster indicates the genes with similar functions through most of applied experimental conditions.

It should be considered that the time series gene expression data which is used in this paper, has few number of time points (less than 8); in addition, most clustering algorithms are unable to distinguish between real and random patterns for short data sets [3]. For example

Manuscript received July 5, 2008. This work was supported by the K.N.Toosi University of Technology.

P. Sobhe Bidari, R. Manshaei, J. Alirezaie are with the Department of Biomedical Engineering at K.N. Toosi University of Technology, Tehran, Iran (phone: (+98)9121725910; e-mail: {pooya.sobhebidari, r.manshaei}@ee.kntu.ac.ir, alireza@eetd.kntu.ac.ir). J. Alirezaie is also with the Department of Electrical and Computer Engineering at Ryerson University, Toronto, ON, Canada (e-mail: javad@ee.ryerson.ca) and with the Department of Systems Design Engineering at University of Waterloo, ON, Canada (e-mail: javad@rousseau.uwaterloo.ca).

T. Lohrasebi, A. Feizi, M. A. Malboobi are with National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran (email: {narcis,malboobi}@nigeb.ac.ir, afeizi@gmail.com).

clustering algorithm based on the dynamics of the expression patterns [4], clustering using the continuous representation of the profile [5] and clustering using a Hidden Markov Model (HMM) [6] are not appropriate for short time series data set because they over-fit the data when the number of time points is small. K-means and FCM are introduced as appropriate methods for clustering this kind of data [3]. Therefore, in this paper, K-means and FCM algorithms are employed to cluster the data set in the first phase and Pearson correlation is used in the second phase.

II. Gene Expression Data

A. Locus Numbers

The Locus number of a gene is its fixed position on the chromosome, which describes its location on the chromosome. In this paper, the mathematical algorithms are applied to the time series gene expression data of 54 genes which are expected to be stimulated in phosphatase activity of *Arabidopsis thaliana* plant which has approximately 28000 genes [16]. 59 genes out of 28000 are introduced as controllers of phosphatase condition in the plant in [7]. In this reference, the genes are studied from the Sequence Alignment aspect. In our paper, the time series expression data of the genes in the root and shoot of plant are studied to find out their behavior on the specified conditions. In reliable data sets [17], [18], the time series data of only 54 genes out of 59 could be found; therefore, in this paper only these 54 genes are studied; their locus numbers are depicted in Table 1.

B. Data Format

The function of each gene in the root and shoot of *Arabidopsis thaliana* has been studied in 10 experimental conditions. In each condition, gene expression value is calculated for 7 time points. The expression values applied to the clustering algorithms are the ratio of values measured from each experiment in each time point to the value measured from the relevant control experiment.

The experimental conditions are Pi-/Pi+, heat, osmotic, salt, drought, genotoxic, oxidative, UV-B, wounding and cold stresses, and the time points for all of conditions except Pi-/Pi+ are 0, 0.5, 1, 3, 6, 12, and 24 hours, passed from the beginning of the experiments. The time points for Pi-/Pi+ condition are 0, 3, 7 and 14 days. In order to have the same number of time points in all conditions, before performing proposed method other points which are 5, 9 and 11 days are estimated by cubic interpolation.

The data set for Pi-/Pi+ condition used in this paper is produced by National Institute of Genetic Engineering and Biotechnology (NIGEB) which is under publication [17], and the data set for other 9 experimental conditions on *Arabidopsis thaliana* used in this paper are downloaded from the database of the University of Toronto [18]. Times series gene expression data sets are studied separately in the root and shoot of *Arabidopsis thaliana*.

TABLE 1
Locus Number Of 54 Genes Used In This Paper

	Locus Number		Locus Number
1	At1g04040	28	At3g10150
2	At1g09870	29	At3g15820
3	At1g13750	30	At3g15830
4	At1g13900	31	At3g17790
5	At1g14290	32	At3g20500
6	At1g14700	33	At3g46120
7	At1g15080	34	At3g50920
8	At1g17710	35	At3g52780
9	At1g25230	36	At3g52810
10	At1g52940	37	At3g52820
11	At1g56360	38	At4g13700
12	At1g69640	39	At4g14930
13	At1g73010	40	At4g24890
14	At2g01180	41	At4g25150
15	At2g01880	42	At4g29260
16	At2g01890	43	At4g29270
17	At2g03450	44	At4g36350
18	At2g16430	45	At5g03080
19	At2g18130	46	At5g15070
20	At2g27190	47	At5g24770
21	At2g32770	48	At5g24780
22	At2g38600	49	At5g34850
23	At2g39920	50	At5g44020
24	At2g46880	51	At5g50400
25	At3g01310	52	At5g51260
26	At3g02600	53	At5g57140
27	At3g07130	54	At5g63140

Features / Samples	F1	F2	…	F7
The i^{th} Gene (i = 1,…,54)	Time 1 (0 hour)	Time 2 (0.5 hour)	…	Time 7 (24 hour)

Fig. 1. Structure of Samples and Features in Phase 1. Features related to each experiment are the gene expression values over times: 0, 0.5, 1, 3, 6, 12 and 24 hours after applying 9 conditions or 0, 3, 5, 7, 9, 11, 14 days after applying Pi-/Pi+ condition to *Arabidopsis thaliana*.

Features / Samples	F1	F2	…	F10
The i^{th} Gene (i = 1,…,54)	Pattern in Condition1	Pattern in Condition2	…	Pattern in Condition10

Fig. 2. Structure of Samples and Features in Phase 2. For each gene, features 1, 2, …, 10 are the pattern labels that the times series behavior of the gene follow under experimental conditions 1, 2, …, 10 which is provided in phase 1.

III. Proposed Method

A. Two Phase Functional Clustering

The new method proposed in this paper for functional clustering of genes includes two phases. In phase 1, each experimental condition is studied separately; so that, clustering of genes are performed according to their dynamic behavior over time. In this phase, as shown in Fig. 1, genes are considered as samples and their expression values over the mentioned times are considered as features.

As it will be discussed in the next section, the output of phase 1 is new features for each gene which describe the pattern of gene behavior over the experimental conditions.

In phase 2, genes with new features are studied, and the similarities of their overall functions are extracted by calculating Pearson correlation between all pairs of genes. As shown in Fig. 2, genes and the patterns extracted over each condition are considered as samples and features, respectively.

B. *Details of Phase 1*

The steps of phase 1 are as follows:

1) Loading data: The time series data is loaded into a 3 dimensional matrix named *Data(g,p,c)* in which rows, columns and depths indicate genes, time points and experimental conditions, respectively.
2) Preprocessing: *NewData(g,p,c)* is calculated from *Data(g,p,c)* by (1):

$$NewData\ (g, p, c) = \frac{Data\ (g, p, c) - Data\ (g, p-1, c)}{t(p) - t(p-1)} \quad (1)$$

$$t(p) = 0, 0.5, 1, 3, 6, 12, 24 \quad , p = 1, ..., 7$$

$$or\ \ t(p) = 0, 3, 5, 7, 9, 11, 14 \quad for\ Pi - / Pi +$$

where *g = 1, …, 54, c = 1, …, 10, t(p)* is the time passed from the beginning of experiment related to column *p*.

3) Estimating the best numbers of clusters for all conditions in the way that the followings terms be confirmed:
   - Having no cluster with 1 gene,
   - Having no cluster with many genes.

   For reaching this target, the number of clusters has been changed from 3 to 10 and after applying clustering algorithms, the best number for each experimental condition has been chosen.
4) Selecting the first experimental condition and putting the related data in a matrix named *C-Data*.
5) Applying clustering methods: In this step, *C-Data* is clustered by K-means and FCM algorithms to the best number of clusters obtained in step 3.

   K-means is an unsupervised learning algorithm for solving clustering problems to find certain number of clusters fixed a priori. The main idea is to define *k* centroids, one for each cluster. Different location of centroids causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early iteration is done. At this point, *k* new centroids are re-calculated as the mean of the clusters resulting from the previous step and the members of *k* clusters are also re-calculated. The iterations are repeated until the time that the centroids are fixed and an objective function is minimized. The objective function *J* used in this paper is the sum of distance between data points of each cluster and their centroid, as shown in (2):

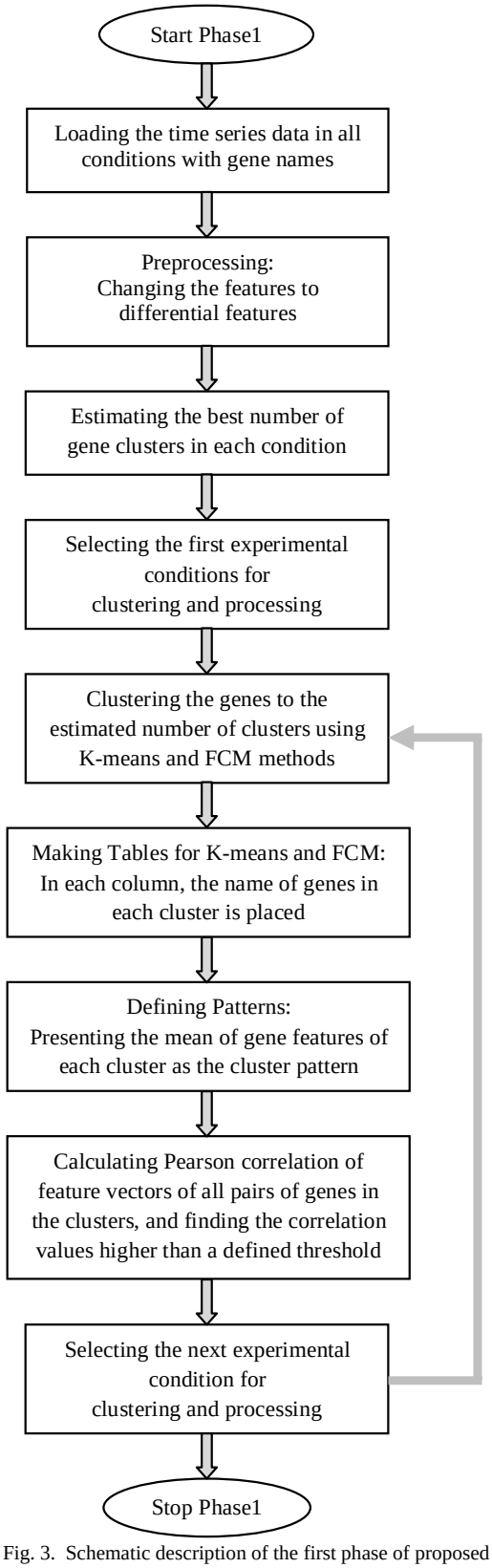

Fig. 3. Schematic description of the first phase of proposed algorithm.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (2)$$

where k is the number of clusters, n is the number of members in cluster j, $x_i^{(j)}$ is an data point and c_j is the cluster centroids [8].

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is based on minimization of an objective function which is expressing the similarity between any measured data and the center shown in (3):

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 \quad ,1 \le m < \infty \quad (3)$$

where N is the number of all data members, C is the number of clusters, m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} sample and c_j is the center of the cluster.

FCM is carried out through an iterative optimization of the objective function, with the update of membership u_{ij} by (4) and the cluster centers c_j by (5):

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \quad (4)$$

$$u_{ij} = \frac{\sum_{i=1}^{N} u_{ij}^m . x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (5)$$

This iteration will stop when (6) is satisfied:

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon \quad (6)$$

where ε is a termination criterion between 0 and 1, and k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m [9], [14].

6) Making Tables: Names of genes in each cluster which are obtained in step 5 are put in 2 tables for K-means and FCM to be used in phase 2.
7) Defining patterns: The pattern of each cluster is the time series values which are the means of expression values of the genes in that cluster. Patterns as the symbol of clusters are used as the new features of the genes in phase 2.
8) Calculating Pearson correlation: The Pearson correlation is commonly used as the similarity measure between genes [15]. In this paper, the correlation is applied to all pairs of genes in each cluster which leads to find the genes with higher functional similarity. If the correlation between the time series feature vector of two genes is higher than 0.9, they are considered as two genes with similar function in the related experimental condition. Pearson correlation is shown in (7):

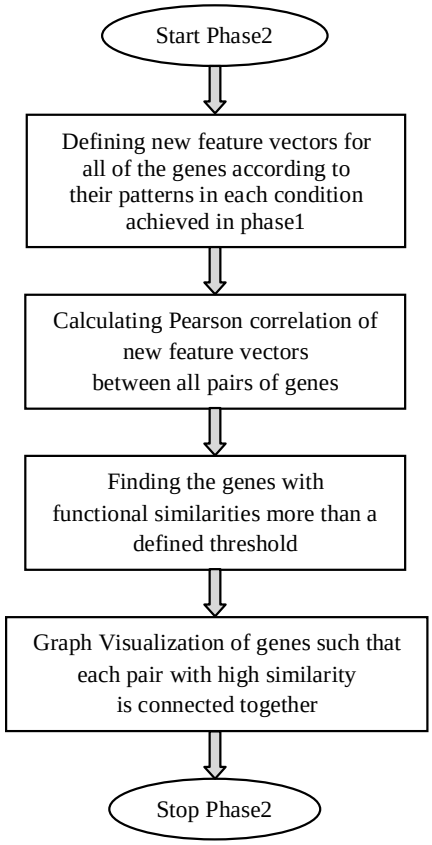

Fig. 4. Schematic description of the second phase of proposed algorithm.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} . \sqrt{\sum (y - \bar{y})^2}} \quad , r \in [-1,1] \quad (7)$$

where r is the Pearson correlation between feature vectors x, y and $\bar{x}$, $\bar{y}$ are the mean values of x, y.

9) Selecting the data of next experimental condition as *C-Data* matrix and returning to step 5.

C. Details of Phase 2

The steps of phase 2 are as follows:

1) New feature vectors: In phase 2, the methods are applied to the new data set obtained from phase 1. As mentioned before, new feature vectors of each gene are the labels of the patterns related to the gene behavior in the experimental conditions.
2) Calculating Pearson Correlation: The concept of calculating Pearson correlation in phase 2 is to find pairs of genes with higher behavior similarity over all conditions.
3) Two genes with correlation coefficient higher than 0.75 are supposed to be functionally similar. This means that they have resemblance behavior at least over 75 percents of the conditions.
4) Graph visualization: All of the genes are visualized as nodes of a graph. Two genes are connected if they are supposed to be similar in step 3. Therefore, by this visualizing, gene interconnection networks can be recognized.

TABLE 2
RESULTS OF APPLYING THE PROPOSED METHOD TO THE DATA SET

Phase	Experimental conditions	Number of Clusters	Number of Achieved Interconnections	
			FCM	K-means
1	Pi-/Pi+	7	266	255
	Heat	7	25	31
	Osmotic	8	42	33
	Salt	8	51	50
	Drought	7	32	32
	Genotoxic	7	22	19
	Oxidative	7	31	35
	UV-B	7	24	29
	Wounding	7	34	37
	Cold	9	58	65
2	All of Conditions	-	90	91

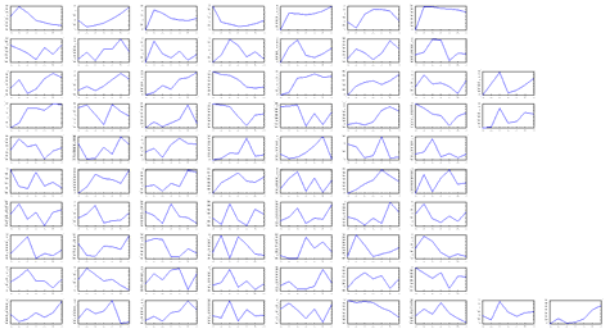

Fig. 5. Time series patterns obtained from clustering of genes over all experimental conditions in the root of plant. Patterns in the same row are obtained from the same experimental condition. Columns indicate the cluster number in the experiments.

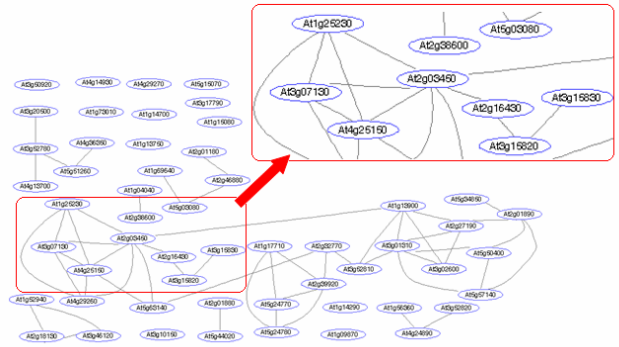

Fig. 6. Gene interconnection network obtained from two phase functional clustering in *Arabidopsis Thaliana* Phosphatase-encoding genes of the root. Two genes are connected if the Pearson correlation of their patterns over all conditions would be higher than 0.75.

IV. RESULTS

In this section, the results of applying the proposed algorithms to the data set are presented. In Table 2 the followings are illustrated: the cluster numbers selected for each condition, interconnections achieved separately in all of experimental conditions in phase 1, the number of iterations for clustering, and the number of interconnections obtained after calculating the Pearson correlation for all conditions together in phase 2.

Fig. 5 illustrates 74 time series expression patterns of genes behavior in the root of the plant calculated from the data set in phase1. Patterns in the same row are obtained from the same experimental condition and the columns indicate the cluster number in the experiments.

Gene interconnection networks obtained by comparing the gene functions through all conditions are shown in Fig. 6. As described in the last section, this network is the result of applying Pearson correlation and finding pairs of genes with correlation coefficient more than 0.75.

V. CONCLUSION

In this paper a novel approach for time series gene clustering has been proposed which includes two phases. In this approach, genes are clustered by K-means and FCM methods according to theirs time series expression, then patterns of gene behavior are extracted. Then, new features are defined for the genes and by calculating Pearson correlation between new feature vectors, genes with similar expression behavior are obtained which can lead to find interconnections between genes. This algorithm is applied to *Arabidopsis thaliana* Phosphatase-encoding gene data set of some experimental conditions. At least 75 percent of these genes with similar functions were extracted.

REFERENCES

[1] B. Chandra, S. Shanker, S. Mishra, "A new approach: Interrelated two-way clustering of gene expression data", *ELSEVIER J.*, Statistical Methodology vol. 3, 2006, pp. 93–102.

[2] Y. Yuan, C. T. Li, "Unsupervised Clustering of Gene Expression Time Series with Conditional Random Fields", *Inaugural IEEE-IES*, Digital EcoSystems and Technologies Conference, 2007, pp. 571–576.

[3] J. Ernst, G. J. Nau, Z. Bar-Joseph, "Clustering short time series gene expression data", *Oxfordjournals*, *Bioinformatic*, vol. 21, Suppl. 1, 2005, pp. i159–i168.

[4] M. F. Ramoni, P. Sebastiani, I. S. Kohane, "Cluster analysis of gene expression dynamics", *Proc. Natl. Acad. Sci. USA*, 99, 2002, pp. 9121–9126.

[5] Z. Bar-Joseph, G. Gerber, T. S. Jaakkola, D. K. Gifford, I. Simon, "Continuous representations of time series gene expression data", *J. Comput. Biol.*, 2003, pp. 341–356.

[6] A. Schliep, A. Schonhuth, C. Steinhoff, "Using hidden Markov models to analyze gene expression time course data", *Bioinformatics*, 19, 2003, pp. i264–i272.

[7] A. Feizi, M. A. Malboobi, T. Lohrasebi, S. Kazempour Osaloo, "Highly diverse plant acid phosphatases are the products of both convergent and divergent evolutionary processes", *BMC press*, submitted for publication.

[8] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, *Berkeley*, *University of California Press*, vol. 1, 1967, pp. 281-297.

[9] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics*, vol. 3, 1973, pp. 32-57.

[10] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 1967, pp. 281–282.

[11] S.C. Johnson, "Hierarchical clustering schemes", Psychometrika 2 (1967) 241–254.

[12] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. 95 (1998) 14863–14868.

[13] R. Sharan, R. Shamir, “CLICK: A clustering algorithm with applications to gene expression analysis”, Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, ISMB, 2000, pp. 307–316.
[14] J. C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algoritms”, *Plenum Press, New York*, 1981.
[15] G. Getz, E. Levine, E. Domany, “Coupled two-way clustering analysis of gene microarray data”, *Proc. of the National Academy of Sciences of the United States of America*, vol. 97, 2000, pp. 12079-12084.
[16] H. Lan, R. Carson, N. J. Provart, A. J. Bonner, “Combining classifiers to predict gene function in *Arabidopsis thaliana* using large-scale gene expression measurements”, *BMC Bioinformatics*, 2007, 8-358.
[17] http://www.nrcgeb.ac.ir
[18] http://bbc.botany.utoronto.ca