International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 1, No. 1, 2014, pp. 104-111.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

A new method to determine cluster heads based on principal component analysis in gene regulation process

Masoume Maghsoodi
Graduate student, Computer, Software, Islamic Azad University, Zanjan Branch, Department of Engineering and Technology, Zanjan, Iran

Abstract

There is a vital chemical in the cells of every living being which is called nucleic acid; it makes up the genetic material of cells. The genetic material is a factor which transfers the characteristics of a species from one generation to another. The nucleic acids in cells have two types: RNA and DNA. The genetic mutations or environmental toxins cause the cells react differently; the reaction is mostly performed in undesirable way. Therefore, the identification of gene regulation process is applicable not only to understand the disease, but also to provide appropriate treatment. However, this study aimed to increase the efficiency of gene regulation network using new methods of determining cluster heads based on principal component analysis in gene regulation process. After creating the network of gene regulation, if we connect different clusters and connect the networks of each cluster, an appropriate agent should be selected for each cluster and the agents should be connected to each other. Then, the network is made among these agents. The results of tests showed that the proposed method improved detection rate.

Keywords: Principal Component Analysis, Cluster Heads, Gene Regulation Network Inference.

Introduction:

Given the importance of gene regulation network construction and inference and its effective applications in drug design, follow-up of a disease treatment, and etc., the objective is to provide a method to build a gene regulation network which will have acceptable performance. The process in which DNA codes are transcript to RNA codes is called gene expression. The basic process of gene function should be learned to understand how genes behave under different internal and external conditions. The gene regulation is the process of determining the function of genes. In response to stimuli, a large number of molecules which are very different in terms of performance communicate selectively with each other over time and space. However, many issues including the timing and details of relationship among genes, RNA, and proteins are still unknown. In many cases, the malfunction of components which are in the way of one of the network paths causes illness. The genetic mutations or environmental toxins cause the cells react differently; the reaction is mostly performed in undesirable way. Therefore, the identification of gene regulation process is applicable not only to understand the disease, but also to provide appropriate treatment. Derisi et al. [1] studied the body metabolism and genetic control of gene expression in

genome-scale. The vital protein structures in living organisms can be things like muscle, hemoglobin, enzymes, antibodies, transcription factors, and so on. The type of these proteins and their mixtures determine the structure of organisms and their adaptation with different environmental conditions [2]. The gene expression is a time-dependent activity. The cells need different proteins to perform different tasks at different times and conditions. At exposure to new situations such as lack of food, stress, infection, heat shock, cold shock, and etc., or to new state such as cell cycle, the cells react by changing transcription program and production of needed proteins [3]. The differentiation of different body tissues cells and cell shape changes also occur due to regulation of gene expression by factors in response to symptoms and stimulants. In regard to importance of transcription factors, we may refer to their role in regulation of cell growth cycle, prevention of excessive proliferation of aberrant cells, and prevention of tumors and cancer formation. Krayjun and colleagues [4] studied a repository for microarray data which were generated by trans-NASC services. Developments of computational modeling can reveal the properties and behaviour of these biomolecules similar to the work done on blood components like (HDL) and living cell receptors as integrin. [5,6,7].

Challenges and Problems:

1. The gene expression data contain a large number of genes and a small number of sample which is equivalent to tight structure of data in a space with very high dimensions. This is called high dimension problem.
2. The uncertainty in gene expression data is another problem. The dynamicity of biological process is the most important reason of this uncertainty. Another reason for this uncertainty is that laboratory processes are inherently changeable. This means that the genes expression levels are changed by many factors such as temperature, pressure, and light.
3. The data are often impregnated with noise. The noise in data makes difficult the processing of data.
4. The proposed methods for gene regulation networks have high time complexities.

The following solutions are proposed to achieve the stated objectives and solve the above problems:

1. Fuzzy hierarchical clustering to build a network of gene regulation
2. Presenting a new method to determine cluster heads based on principal component analysis.

The calculation of gene regulation networks topology map is one of the problems of biological systems. The experimental techniques such as protein binding assays may determine the genes regulation, but they require more time and antibodies for each transcription factor. Pierre and colleagues [8] studied the principles and strategies to develop network models in cancer. There are many algorithmic methods for gene regulation networks. The [9] compared the accuracy of prediction in 8 unsupervised methods and one supervised method in 38 datasets. The findings showed that although the unsupervised methods parameters were selected as optimal, the prediction accuracy of supervised methods was better. The findings showed different data types including knockout, knockdown, and multifactorial. The [10] reported the strengths and weaknesses of gene network inference methods. The findings of DREAM showed that the network inference is a challenge. In 2010, a method was provided to solve the challenges of DREAM3 which includes identification of cascade signaling, predicting signal response, prediction of gene expression, and challenge of DREAM3 in silico network [11]. Therefore, it is necessary to introduce a good method to infer gene regulation networks and optimize them. Some common methods of gene regulation network construction include Correlation, BMALR, GINIE3, and CLR. However, the main objective of this paper is to optimize these methods with clustering technique and selecting good representative of each cluster based on principal component analysis [12]. The gene expression is the proposed method input of matrix. Based on Multifactorial data, this matrix is a 100 in 100 matrix which includes the expression of 100 genes in 100 samples. The objective is to build a network which includes 100 nodes. Each node represents a gene. The edges among nodes represent the regulatory relationships

among genes. The built network will be compared with standard existing network to determine the effectiveness of proposed method.

In summary, the proposed method involves the following steps: clustering of data and genes in different clusters, gene regulation network construction in each cluster based on existing procedures, getting the cluster heads in each cluster based on independent components analysis, building networks among cluster heads, and obtaining final network based on built networks in previous steps. The principal component analysis is used to calculate the cluster heads. The principal component analysis (PCA) is a data analysis tool which may be used for reducing data dimensions, data compression by data loss, feature extraction, and data representation [13]. There are two common but different definitions for PCA which lead to same algorithm. In the first definition, the PCA is an orthogonal data image in a linear space with less dimensions; the variance of imaged data is maximum. Also, the PCA may be defined as a linear image which minimizes the average image cost. The average image cost is the average distance between data and image. Suppose we have N samples with D dimensions. The objective is that with data in a space with dimension M <D, the M will be mapped in a way that maximizes the variance of new data. Let's suppose that we want to map the data in one-dimensional space (M = 1). This space may be represented using a D dimensional vector. We assume that the size of vectors are equal, without loss of generality.

This means that u = 1 QUOTE u. To obtain data image (x_n), it is necessary to obtain its internal multiplication with vector u_1 which assumed here (M = 1). The imaged data average is obtained by u_1x; where QUOTE x is the basic data average which is obtained by equation (1):

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad (1)$$

To calculate the variance:

$$\frac{1}{N}\sum_{n=1}^{N}\{u_1 x_n - u_1\bar{x}\}^2 = u_1{}^T S u_1 \qquad (2)$$

In above equation, S is the basic data variance and is obtained as follows:

$$S = \frac{1}{N}\sum_{n=1}^{N}(x-\bar{x})(x-\bar{x})^T \qquad (3)$$

The PCA aims to maximize the amount of variance in equation (2-3) by indicating u = 1 u QUOTE u^T:

$$u_1{}^T S u_1 + \lambda_1(1-u_1{}^T u_1) \qquad (4)$$

Derived from above equation, we have:

$$Su_1 = \lambda_1 u_1 \qquad (5)$$

The above equation indicates that the u1 is one of the special vectors of matrix S and QUOTE ë λ_1 is the value corresponding to this vector. If we multiply the two parts of above equation in QUOTE U $u^T{}_1$ and given the u = 1 u QUOTE u^T, we have:

$$u_1{}^T S u_1 = \lambda_1 \qquad (6)$$

The above equation indicates that the new data variance (which is created by mapping on vector u1) is equal to λ_1 **.** **T**he special vector with largest eigenvalue should be used to maximize this variance. However, it is desirable that the M dimensional space will be used instead of one-dimensional space in mapping of data. The variance in this M dimensional space should be highest. To do this, the vectors with highest eigenvalues should be selected. In summary, the mean and variance of samples should be obtained for mapping data in new space with maximum variance. This objective will be achieved by obtaining special vectors of covariance matrix and selecting M vectors with largest eigenvalues [14]. The data transmission to other space allows identification of best agents. The cluster heads are in fact the

representatives of classes and should be most like to cluster data. The mean is used to calculate the most similarity. The genes which are closer to cluster center or mean will be selected as representatives. Figure 1 shows the distribution of each cluster data based on principal component of PCA. The cluster heads are calculated based on this distribution. The data which are close to center of cluster are considered as cluster heads.

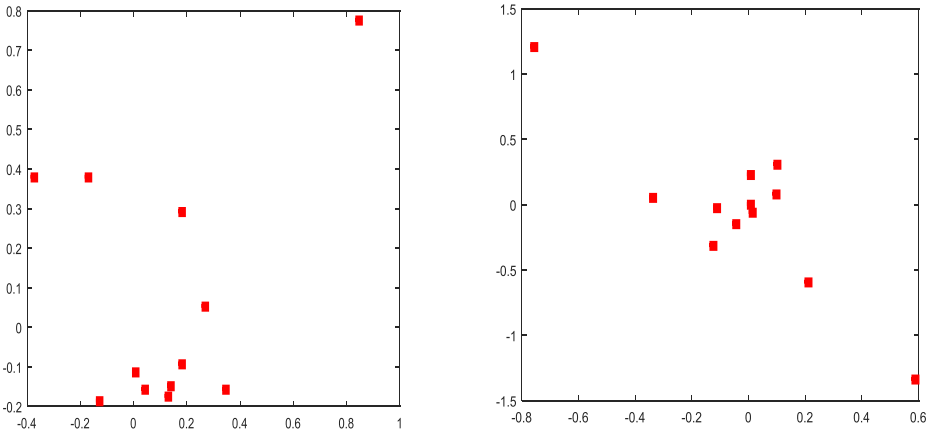


Figure 1. The distribution of each cluster data based on principal components of PCA method

Using the genetic regulatory networks relationships, the genes involved in a disease may be identified. In deriving these networks, the different parameters such as clustering type, number of clusters, evaluation criteria, clusters representatives, and clusters centers impact on validity of proposed method. The appropriate algorithms may be used for optimal regulation of these parameters.

Binding of cluster heads:

After building gene regulation network in each cluster, we have a set of separate networks. For example, if we have ten clusters, a gene regulation network is built in each cluster and we have ten separate gene regulation networks. Therefore, the cluster head must be identified to bind these networks. In above, we explained the method of calculating head cluster. After identifying the cluster heads (genes), the mutual information will be calculated between pairs of cluster heads.

The mutual information (MI) between two cluster heads is calculated as follows:

$$(\curlyvee)\quad I(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) \log\left(\frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}\right)$$

If the information will be more than a specified limit that we've considered to be 0.5, the biding between cluster heads will be established. Otherwise, the binding will not be established between two pairs of cluster heads. Figure 2 shows the binding of cluster heads.

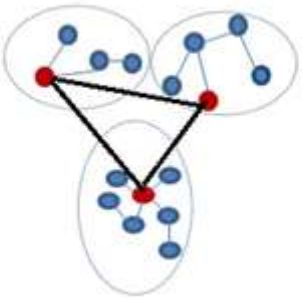


Figure 2: Binding of cluster heads to each other

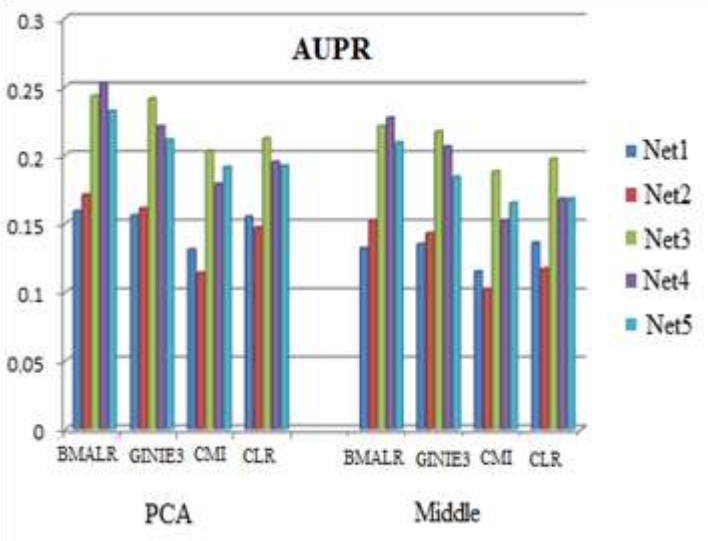
Therefore, the clustering is first performed on genes; the related genes will be in same clusters. By doing this, a large number of edges between clusters may be removed. Then, the principal component analysis is used to determine the cluster heads to bind the networks of each cluster. After determining the cluster heads, they will bind together based on mutual information and the final network will be achieved.

Data set:
The Dream4 data set was used to evaluate the proposed method [14]. This data set includes five networks. In every network, there are 100 genes and the gene expressions have been obtained in 100 different times. According to proposed method, the input is a matrix of 100 by 100. The objective is to deduce five networks with n = 100 genes which are the result of different mutations of genes. There are 100 samples in each network.

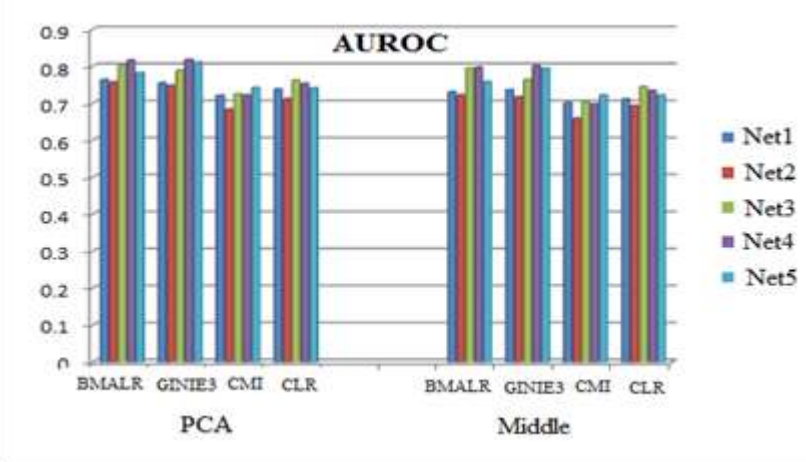
Evaluation criteria:
The proposed method assigns weight to each possible binding between genes. The precision-recall curve (PR) and ROC curve were used to assess these weights.

Impact of cluster heads:
The number of cluster heads and method of obtaining them are important parameters which impact on gene regulation network construction. First, the method of finding cluster heads is investigated. For this purpose, the cluster heads are obtained by cluster center method and principal component analysis. In cluster center method, the genes are arranged in clusters and the median is considered as cluster. In independent components analysis, if the genes in each cluster will be much farther away from mean of cluster, they will be less likely to be selected as cluster heads. The genes which are close to mean of cluster are identified as head cluster.

Graph 1. Evaluation based on AUPR criterion and determining the cluster head based on median

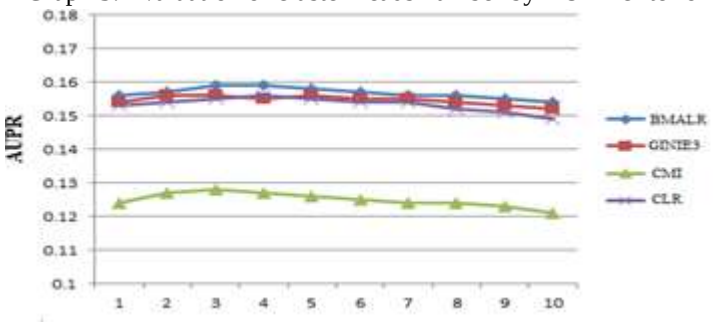


Graph 2. Evaluation based on AUROC criterion and determining the cluster head based on median

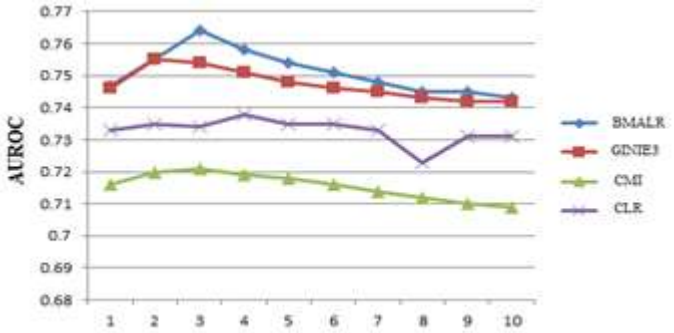


The obtained results and determining cluster heads based on PCA raises both evaluation criteria. The reason is that this method takes data from each cluster into a space in which the data scattering is more. The data which are close together are like being close to mean of class and are considered as head cluster. The number of cluster heads is another parameter affecting the finding of cluster heads. For this purpose, figure 3 and 4 show the value of AUPR and AUROC criteria in first network with different cluster heads.

Graph 3. Evaluation of cluster heads number by AUPR criterion



Graph 4. Evaluation of cluster heads number by AUROC criterion



According to Graph 3 and 4, it is better that the number of cluster heads will be low. This is a logical result; if the number of cluster heads will be more, the creation of edge between clusters will be more possible. Given that the number of edges is low in gene regulation networks, it is better that the number of cluster heads will be low. Based on the tests, the number of cluster heads is 2 to 4. In many methods, the number is considered to be 3.

Conclusion:

The final assessment showed that the cluster heads should be obtained to converge and bind gene regulation networks in each cluster. In this paper, three cluster heads were obtained for each network. These cluster heads were obtained based on principal component analysis. The results of tests showed that the efficiency of this method is better than other methods. After calculating cluster heads based on calculating the correlation between cluster heads, the cluster heads with correlations above 0.7 were bound to each other. Therefore, the selection of cluster heads increases the performance of gene regulation network inference methods.

References:

Craigon , D and James , N and Okyere , J and Higgins , J and Jotham , J and May , S and Nascarrays (2004) ,'' a repository for microarray data generated by NASC's transcriptomics service '', Pp.575–577.

DeRisi , J and Iyer, V and Brown, P (1997) ,'' Exploring the metabolic and genetic control of gene expression on a genomic scale '', pp. 680-686.

Elnitski , L and Jin , V and Farnham , P and et al (2006) ,''Locating mammalian transcription factor binding sites a survey of computational and experimental techniques. Genome Res '', pp 1455–64.

http://rapguidetoevolution.co.uk

Damirchi, B., Saidi, M. S., Rismanian, M., Firoozabadi, B., & Amininasab, M. (2016). ''An alternative mechanism for the formation of high density lipoprotein in peripheral tissue ''. Scientia Iranica. Transaction B, Mechanical Engineering, 23(2), 600.

Damirchi, B., Rouhollahi, A., Sohrabi, S., & Mehr, S. M. N. (2013). ''Modeling and Stability Analysis of Truncated High Density Lipoprotein (HDL) System Using Martini Coarse Grain Technique''. In ASME 2013 International Mechanical Engineering Congress and Exposition (pp. V03AT03A069-V03AT03A069). American Society of Mechanical Engineers.

Shamloo, A., Nikbin, E., Mehboudi, N., & Damirchi, B. (2014). ''Homo-oligomerization of transmembrane α-domain of integrin''. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE (pp. 1162-1165). IEEE.

Jen , C and et al. ,'' The Arabidopsis co-expression tool (ACT)'', pp.336–348.

Jon Shlens ,"A Tutorial on principal cpmponent analysis ".

Madhamshettiwar , P and et al, (2012) ,''Gene regulatory network inference evaluation and application to ovarian cancer allows the prioritization of drug targets'', pp.41.

Peer , D and Hacohen , N (2011) ,'' Principles and strategies for developing network models in cancer '', pp. 864–73.

Stolovitzky, G and Monroe ,D and Califano , A (2007) ,''Dialogue on reverse-engineering assessment and methods the DREAM of high-throughput pathway inference '', pp 1–22.

The DREAM project. http://wiki.c2b2.columbia.edu/dream/index.php/ The_dream Project.

Xun , H and Zhike Zi (2014)," Inferring cellular regulatory networks with Bayesian model averaging for linear regression (BMALR)", pp.