A New Method to Build Gene Regulation Network Based on Fuzzy Hierarchical Clustering Methods

Maghsoodi, Masoume

Islamic Azad University, Iran

1 June 2016

International Academic Journal of Science and Engineering
Vol. 3, No. 6, 2016, pp. 169-176.

A New Method to Build Gene Regulation Network Based on Fuzzy Hierarchical Clustering Methods

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

Abstract
The construction of genetic regulatory networks is understanding the relationship among genes or circuits which regulate the conditions of cells in response to internal or external stimuli. In fact, the objective is to understand the network of relationship among genes which determine which genes are responsible for activating other genes. The understanding of relationships may help to identify the genes which are involved in a disease and design the drugs. The most important limitations in gene regulatory network inference are low number of samples, noise penetration possibility, and large number of genes. There are different models to build gene regulatory network. This study used fuzzy hierarchical clustering method to infer gene regulatory network. Using clustering, the similar genes will be in a cluster. Many edges therefore will be removed. The final assessments showed that the genes clustering increased the efficiency of gene regulation network inference methods.

Keywords: Principal Component Analysis, Head Cluster, Clustering, Gene Regulation Network.

Introduction:
Our body is made up of hundreds different cells. All of them are created by proliferation of a basic cell. Therefore, their constituent elements are same. But, only some genes are expressed in each cell type; that's why the cells shape and task are different. Derisi et al. [1] studied the body metabolism and genetic control of gene expression in genome-scale. 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 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. Skelet and Brazma [2] studied the common approaches of gene regulatory network modeling of BMC Bioinformatics. In recent years, there are provided many methods to estimate the gene regulation networks by gene expression profiles including logical networks, graphical Gaussian model (GGM), differential equation models, and Bayesian networks (BN). Each of these methods has its own problems. The composed graphs in logical network are zero and one and there is no weighting graph. The Gaussian models and Bayesian networks need prior knowledge to build gene regulation network. If the prior knowledge on building gene regulation networks will be less, the experts in this field will be

needed. In differential networks, the number of considered edges is too high and makes low the detection rate in edges creation. The proposed method requires prior knowledge. The data clustering and obtaining the representatives of each cluster lead to obtaining optimal edges. The common methods of gene regulation network building include BMALR, GINIE3, PCA-CMI, and CLR. In BMALR method, the Bayesian network and linear regression are widely used in reconstruction of regulation networks. Huang and Zhik [3] studied the mean cell regulatory network inference of Bayesian model for BMALR linear regression. The [4] used BMALR method to infer the biological systems interaction. The MRNET method uses mutual information of gene expression profile and a method of feature selection (reduce redundancy – enhanced relationship of MRMR) to derive the interaction between genes [5]. The GENIE method is similar to MRNET method in that a gene plays the target role and the remaining genes play the role of regulators and in using feature selection method. Unlike MRNET method, however, it uses random forests and additional trees for regression and feature selection. Patrick and colleagues [5] studied the inference from gene networks using backward elimination. In CLR method, the MI values of RN method were developed with regard to background distribution. In this method, the possible interactions are those which have more deviation of background distribution. A maximum z-score is calculated for each gene (i) as follows. Yazdanpanah and colleagues [6], [7] studied the tuning of PID gains in Fuzzy PID controllers. In their methods, the controllers' gains can be optimized to reduce the time and cost of path planning for aerial and surface vessels. Huneh [8] studied the regulatory network inference from data expression using a tree-based method.

$$Z_i = \max_j \left(0, \frac{I(X_i, X_j) - \mu_i}{\sigma_i}\right) \quad (1)$$

The identification of gene relationships is applicable not only to understand the disease, but also to provide appropriate treatment. The clustering techniques are helpful in this regard. In the proposed method, first the genes are divided into a number of clusters. The hierarchical classification method is used to build clusters. In the next step, the gene regulation network is built on each cluster individually based on existing methods such as correlation. To connect different clusters and the networks of each cluster, an appropriate agent should be selected for each cluster and the agents should be connected to each other. The principal component analysis is used to calculate the head clusters. Sinai e al [10] used fuzzy clustering law for gene expression data. Shelens [11] studied the teaching of main component analysis. After calculating head clusters based on calculating the correlation between head clusters, the head clusters with correlations above 0.7 were bound to each other.

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 [12].

This study aims to create a hierarchy of gene regulation networks. First using clustering, the similar genes place in a cluster; this will increase the accuracy rate. In the next step, the gene regulation network is built on each cluster individually based on existing methods such as correlation. Based on principal component analysis, the agents are built in each cluster. The network is finally made among these agents.

Fuzzy clusters:

The complexity of biological networks and large numbers of genes in microarray data sets have caused many challenges in analysis of gene expression data. The clustering technique is one of the methods to overcome these challenges; the similar genes are grouped in a cluster to analyze the function of genes. Because of the overlap among biological groups and existence of noise data in microarray data sets, the fuzzy clustering is the most suitable method to analyze gene expression data.

The fuzzy clustering method divides the data into several clusters. Any data in data set belongs to all or some of clusters with a degree of membership. The data that is close to center of a cluster has higher membership degree than the data which are away from center. In fact, the data which are far from center have smaller membership degree.

The clustering aims to place the data set X = {x1, x2, ... , XN} in clusters with cluster center as V = {vi, 1≤ i ≤C}. Assume that Uij refers to data degree of membership of xi in cluster vi. Therefore, the matrix U with C components on given data set is defined as following.

$$U = \{u_{ij} \mid 1 \le i \le C, 1 \le j \le N\} \qquad (2)$$

Also, the wik weight is attributed for every gene for more accurate clustering procedure. This weight shows the importance of K-th gene in i-th cluster. These weights are determined by weight matrix W = {wik | 1 ≤ i ≤ C, 1 ≤ k ≤ D}. In this matrix, D is the dimension of data and C is the number of classes.

According to above definitions, the function in fuzzy clustering should be below minimum.

$$\text{J=} \; J = \sum_{j=1}^{N} \sum_{i=1}^{C} u_{ij}^{m} \sum_{K=1}^{D} w_{ik}^{\tau} (x_{jk} - v_{ik})^2 \; (3)$$

$$0 \le W_{ij} \le 1, \sum_{i=1}^{D} W_{ij} = 1 \quad \text{و} \quad 0 \le U_{ij} \le 1, \sum_{i=1}^{C} U_{ij} = 1 \qquad (4)$$

So, the following equation is used for fuzzy weighting and membership degree.

$$d_{ij} = \sum_{k=1}^{D} w_{ik}^{\tau} (x_{jk} - v_{ik})^2 \qquad (5)$$

$$v_{ij} = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_{jk}}{\sum_{j=1}^{N} u_{ij}^{m}} \qquad (6)$$

$$w_{ik} = \frac{(q_{ik})^{\frac{-1}{r-1}}}{\sum_{s=1}^{D}(q_{ik})^{\frac{-1}{r-1}}} \quad (7)$$

$$q_{ik} = \sum_{j=1}^{N} u_{ij}^{m}(x_{jk} - v_{ik})^2 \quad (8)$$

$$U_{ij} = \frac{(d_{ij})^{\frac{-1}{m-1}}}{\sum_{S=1}^{C}(d_{Sj})^{\frac{-1}{m-1}}} \quad (9)$$

In fact, the mean of cluster centers are taken to calculate the membership of other data in adjacent clusters. By calculating the distance of each data to centers of each cluster, the cluster fuzzy membership function is obtained. In fact, this represents the degree of membership of data in cluster.

Data set:

The DREAM3 challenge dataset was used in this study to infer silico regulation 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.

Evaluation and results:

The DREAM3 dataset was used to assess the proposed method. 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.

Table 1: Evaluation of common gene regulation network inference methods without clustering

Method	NET1		NET2		NET3		NET4		NET5	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
BMALR	0.155	0.745	0.166	0.737	0.231	0.792	0.234	.0808	0.214	0.778
GINIE3	0.154	0.745	0.155	0.733	0.231	0.775	0.208	0.791	0.197	0.798
PCA-CMI	0.124	0.712	0.103	0.661	0.196	0.708	0.163	0.702	0.184	0.722
CLR	0.152	0.732	0.137	0.691	0.2	0.744	0.186	0.738	0.189	0.726

Importance of clustering:

The common gene regulation network construction methods including BMALR, GINIE3, PCA-CMI, and CLR are applied on entire gene expression matrix to show the importance of clustering. Table 1 and Figure 2 show the results of gene regulation network construction using these methods. Table 2 and Figure 3 show the results of gene regulation network construction using clustering and BMALR, GINIE3, PCA-CMI, and CLR methods.

Table 2. Evaluation of common gene regulation network inference methods with genes clustering

Method	NET1		NET2		NET3		NET4		NET5	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
BMALR	0.159	0.764	0.171	0.757	0.243	0.802	0.252	.0817	0.232	0.782
GINIE3	0.156	0.755	0.161	0.748	0.241	0.789	0.221	0.818	0.211	0.811
PCA-CMI	0.131	0.721	0.114	0.683	0.203	0.725	0.179	0.722	0.191	0.742
CLR	0.155	0.738	0.147	0.712	0.212	0.762	0.195	0.753	0.192	0.741

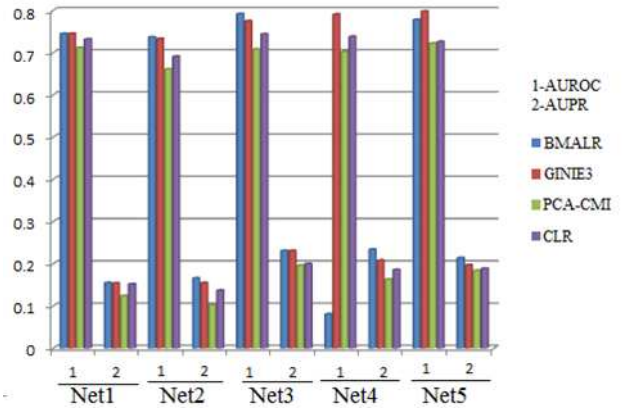

Figure 2. Results of common gene regulation network inference methods without clustering

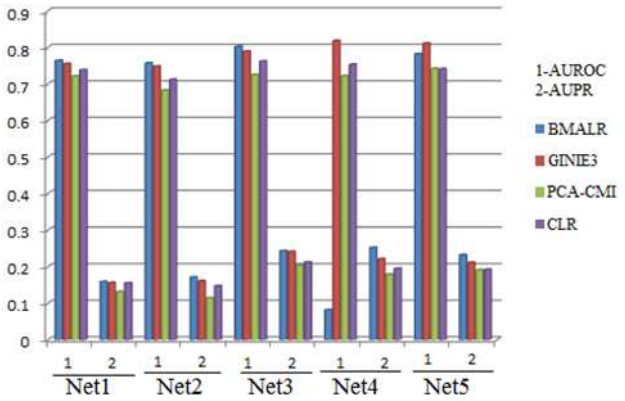

Figure 3. Results of common gene regulation network inference methods with clustering

As it can be seen, the AUPR and AUROC criteria increased in all methods. The reason for this increase is that using data clustering, many edges are deleted. Given that gene regulation network is in dispersed space in terms of number of edges, the method worked well. It should be noted that the number of cluster was considered to be 10 clusters; this number is obtained based on multiple tests. When the number of

Table 3: Assessment of common gene regulation network inference methods with hierarchical clustering

Method	NET1		NET2		NET3		NET4		NET5	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
BMALR	0.159	0.764	0.171	0.757	0.243	0.802	0.252	.0817	0.232	0.782
GINIE3	0.156	0.755	0.161	0.748	0.241	0.789	0.221	0.818	0.211	0.811

clusters is 10, the evaluation criteria are better. The clustering method is one of the important parameters in proposed method. The K-means and c-means were clustered to identify the best method of clustering gene expression data using hierarchical clustering methods. The results of clustering methods with gene regulation networks construction methods are displayed in table below.

PCA-CMI	0.131	0.721	0.114	0.683	0.203	0.725	0.179	0.722	0.191	0.742
CLR	0.155	0.738	0.147	0.712	0.212	0.762	0.195	0.753	0.192	0.741

Table 4: Assessment of common gene regulation network inference methods with k-means clustering

Method	NET1		NET2		NET3		NET4		NET5	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
BMALR	0.156	0.746	0.167	0.739	0.235	0.796	0.241	.0808	0.221	0.779
GINIE3	0.155	0.747	0.157	0.736	0.234	0.781	0.216	0.806	0.206	0.805
PCA-CMI	0.126	0.715	0.11	0.671	0.195	0.711	0.169	0.713	0.189	0.732
CLR	0.151	0.732	0.137	0.694	0.208	0.751	0.189	0.743	0.191	0.734

Table 5: Assessment of common gene regulation network inference methods with c-means clustering

Method	NET1		NET2		NET3		NET4		NET5	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
BMALR	0.158	0.752	0.169	0.743	0.239	0.798	0.247	.0811	0.224	0.779
GINIE3	0.156	0.749	0.158	0.741	0.238	0.784	0.219	0.809	0.208	0.807
PCA-CMI	0.128	0.719	0.112	0.679	0.197	0.717	0.172	0.716	0.189	0.736
CLR	0.154	0.734	0.141	0.705	0.208	0.752	0.19	0.747	0.191	0.736

According to above tables, it is clear that the hierarchical clustering method is better than other clustering methods. According to results shown in table, it is also clear that in all clustering methods, the evaluation criteria improved compared with when the clustering was not performed. These results showed that clustering is effective in gene regulation network construction.

Conclusion:

The understanding of relationships among genes may help to detect the genes which are involved in a disease and design drugs. This study proposed a gene regulation network inference method. In this method, first the genes are divided into different clusters; this makes a lot of edges in network to be removed. According to evaluations, the hierarchical approach is better than other methods. The gene regulation are built in networks based on common procedures. The head clusters should be obtained to converge and connect the achieved networks in each cluster.

The final evaluation showed that the clustering of genes improved the efficiency of gene regulation network inference methods.

Reference:

[1] J. DeRisi 1 and V. Iyer 2 and P. Brown 3 ,'' Exploring the metabolic and genetic control of gene expression on a genomic scale '', Science 1997 ,pp. 680-686.

[2] T. Schlitt 1 and A Brazma 2 ,'' Current approaches to gene regulatory network modelling '', BMC Bioinformatics 2007

[3] X. Huang 1 and Z.Zhike 2 ," Inferring cellular regulatory networks with Bayesian model averaging for linear regression (BMALR)", Mol. BioSyst., 2014, pp. 2023-2030.

[4] X. Huang 1 and Z. Zhike ," Inferring cellular regulatory networks with Bayesian model averaging for linear regression (BMALR)", Mol. BioSyst , 2014.

[5] E. Patrick 1 and et al," Information-Theoretic Inference of Gene Networks Using Backward Elimination", BMC Bioinformatics, 2010,

[6] V. A. HuynhThu 1," Inferring Regulatory Networks from Expression Data Using Tree-Based Methods ", PLoS ONE,2011.

[7] E. Abbasi, M.J. Mahjoob and R. Yazdanpanah. "Controlling of Quadrotor UAV Using a Fuzzy System for Tuning the PID Gains in Hovering Mode," International Workshops in Electrical-Electronics Engineering, ACE-2013, Koc University, 5-7 September 2013.

[8] R. Yazdanpanah, M. J. Mahjoob, and E. Abbasi. "Fuzzy LQR Controller for Heading Control of an Unmanned Surface Vessel." International Conference in Electrical and Electronics Engineering, ACE 2013, Koc University, September 2013.

[9] Faith 1 and B.Hayete 2 and J. Thaden 3 ,'' Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles'', PLoS Biol 2007.

[10] M. Sinaee 1 and G .Eghbal 2 and Mansoori 3 , "Fuzzy Rule Based Clustering for Gene Expression Data",4th International Conference on Intelligent Systems, Modelling and Simulation, IEEE 2013.

[11] J. Shlens 1 ,"A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS", Derivation, Discussion and Singular Value Decomposition. 20…, pp.

[12] The DREAM project.http://wiki.c2b2.columbia.edu/dream /index.php/ The_DREAM_Project. 2008 ,pp. 1–7.