GMR
Genetics and Molecular Research
ISSN 16765680

Constructing gene network based on biclusters of expression data

F. Liu[1], L. Yang[1], Z.Z. Tian[1], P. Wu[2] and S.L. Sun[1]

[1]International Software School, Wuhan University, Wuhan, Hubei, China
[2]Department of Oncology, Xiangyang Central Hospital, Hubei, China

Corresponding authors: F. Liu / S.L. Sun
E-mail: wolflf@126.com / sunsl@whu.edu.cn

Genet. Mol. Res. 15 (2): gmr.15028784
Received May 16, 2016
Accepted June 3, 2016
Published July 15, 2016
DOI http://dx.doi.org/10.4238/gmr.15028784

ABSTRACT. Two genes can be co-regulated and possibly have the similar function if they are similarly expressed, which provides a theoretical basis for construction of gene regulatory networks using gene expression data. Herein, a new method of gene regulatory network was constructed based on biclusters in this paper. Given a bicluster, this paper analyzes the correlation between genes in the clusters and then constructs the gene regulatory network by selecting genes with a correlation coefficient.

Key words: Biclustering of gene expression; Gene regulatory network; Gene expression data; Gene analysis methods

INTRODUCTION

With the completion of genome sequencing, microarray technology development, and the emergence of grid computing, the large-scale study of gene expression regulation using computational methods has exploded, with many researchers attempting to draw the regulatory network that controls whole organism gene expression (Friedman, 2004; Hecker et al., 2009; Petricka and Benfey, 2011; Godsey, 2013). The expression regulatory network is a process of how a group of regulatory factors modulates gene expression (Wyrick and Young, 2002). The major elements involved in gene expression regulatory networks include cDNA, mRNA, proteins, and small molecules. The regulatory network can be represented by a directed graphical structure as shown in Figure 1, in which the nodes represent the control element and edges represent the regulatory role. Lee et al. (2002) concluded the six types of the gene regulatory networks shown in Figure 2.

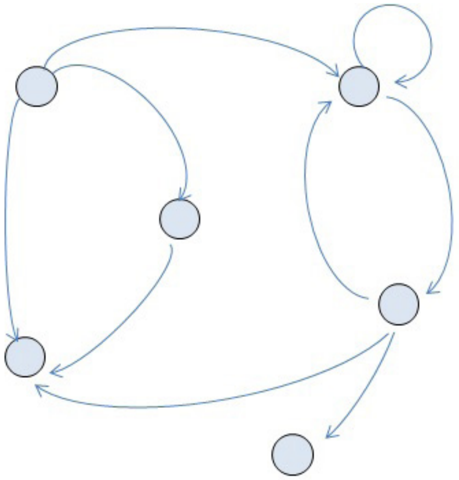

Figure 1. Simple Gene Network Illustration Diagram. Each circle in the figure represents a node, which is the element regulating network.

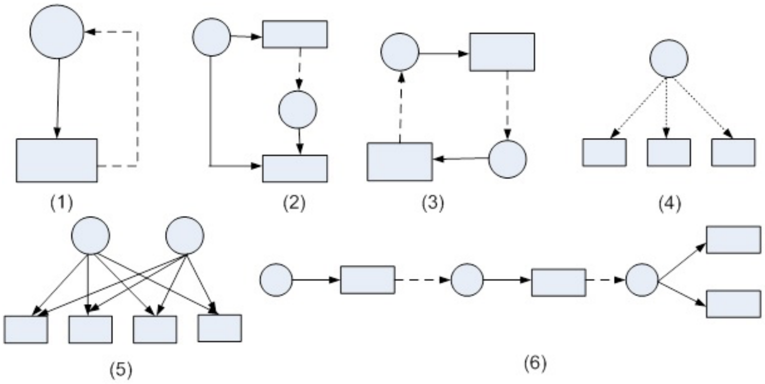

Figure 2. Gene Regulatory Network Category.

When two different genes are similar to each other, they will often be co-regulated, possess similar function, and have the same levels of expression (D'haeseleer et al., 2000; Shmulevich et al., 2002). This observation provides a theoretical basis for using gene expression data to construct gene regulatory networks (Lei et al., 2004). Gene expression data obtained from high-density chips and mass spectrometry experiments can be used to study the function of genes, analyze relationships of mutual coordination and constraint between genes, and study gene transcriptional regulatory networks. When a gene is transcribed, a group of transcription factors binds to an initiation site on the gene and regulates the transcription process, and these transcription factors are also the products of other genes. When a set of transcription factors for one gene also bind an initiation region of another gene, the other gene can also be translated. Once enough of the protein product, for example, an enzyme, accumulates, there may be a feedback loop that will close the initiation region of another gene. When a gene that is transcribed and translated forms a functional gene product, it will directly or indirectly affect the expression levels of itself and of other genes. The expression levels of genes continuously change, creating a fluid expression environment for other genes. In summary, the gene expression activities constitute a biological information system of complex and continuously changing gene networks at the molecular level.

Currently, the primary methods for constructing gene regulatory networks can be divided into two categories: the model based on gene sequences and the model based on expression data. Models and algorithms based on gene sequence analysis include methods such as AlignACE (Hughes, 2002). The models based on expression data includes the following (Yi et al., 2003; Lei et al., 2004): weight matrices (Spears, 1996; Weaver et al., 1999; Butte and Kohane, 2000), Boolean algebra models (D'haeseleer et al., 2000; Simon et al., 2001; Shmulevich et al., 2002), Bayesian network models (Xu et al., 2003; Bickel, 2005), correlation coefficient models (McAdams and Arkin, 1997), and linear combination and differential equation models (Butte and Kohane, 2000).

For gene expression matrices of a time series, numerous methods have been presented that utilize biclustering technology, which selects for similar genes under limited conditions (An et al., 2012; Liu, 2013; Das and Borah, 2014). This method can produce biclusters that show similar expression profiles. In this paper, a new method to construct gene regulatory networks that focus on a selected gene in the biclusters is presented.

MATERIAL AND METHODS

Three steps were used to construct gene regulatory networks: 1) consecutive sequence clustering analysis of gene expression data; 2) correlation analysis between genes in a class; and 3) selection of genes with significant correlation coefficients to construct regulatory networks.

The biclusters used in this work have previously been described in literature (Liu F, 2013). In the study, the given gene expression matrix *X* (*G*, *C*), a multiple asynchronous consecutive sequence biclustering class *B* (*I*, *J*) can be found using biclustering analysis of gene expression data.

Next, the correction coefficient method was used to initially filter gene *I* and its expression value into cluster *B* (*I*, *J*). The main idea was to find interrelated pairwise genes based on their expression values such as genes with similar functions or inhibitory functions.

In this study, the Pearson correlation coefficient was used to quantify the global correlation between two gene expression values. Given two genes *X* and *Y*, and the expression values with *n* conditions represented as $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$, respectively, then Pearson correlation coefficients of gene *X* and gene *Y* can be expressed by following equation:

$$r(X, Y) = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n})(\sum Y^2 - \frac{(\sum Y)^2}{n})}} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n})(\sum {y_i}^2 - \frac{(\sum y_i)^2}{n})}} \quad \text{(Equation 1)}$$

Kato et al. (2001) proposed the concept of local correlation of time delay, defined as the time difference that exists in a correlation between genes. Building on the asynchronous consecutive sequence clustering characteristics, this study used the maximum similarity of an asynchronous consecutive sequence to measure the possible existence of local correlations in time delay between two genes. This method is called a local maximum similarity measure.

Assuming that expression values of two genes *X* and *Y* are $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$, the expression correlation of *q* consecutive time points started from t_1 and t_2 was defined as follows:

$$r(X_{t_1}, Y_{t_2}) = \frac{\sum_{0 \le k \le q} x_{(t_1+k)} y_{(t_2+k)} - \frac{\sum_{0 \le k \le q} x_{(t_1+k)} \sum_{0 \le k \le q} y_{(t_2+k)}}{q}}{\sqrt{(\sum_{0 \le k \le q} x^2_{(t_1+k)} - \frac{(\sum_{0 \le k \le q} x_{(t_1+k)})^2}{q})(\sum_{0 \le k \le q} y^2_{(t_2+k)} - \frac{(\sum_{0 \le k \le q} y_{(t_2+k)})^2}{q})}} \quad \text{(Equation 2)}$$

The local maximum similarity of *X* and *Y* can be expressed by the maximum similarity of expression values with *q* consecutive expression events starting at different time points. Thus, local similarity of *X* and *Y* can be defined as follows:

$$r(X, Y) = \left| \max_{0 \le t_1, t_2 \le m} \right| (r(X_{t_1}, Y_{t_2}) \quad \text{(Equation 3)}$$

Here, Xt_1 and Yt_2 represent the expression values of *q* consecutive expression starting from the time points t_1 and t_2 of *X* and *Y*.

Lastly, genes with significant correlation coefficients were selected to construct regulatory networks.

Some genes were positively correlated while others were negatively correlated, which consequently may have mutual regulatory mechanisms. It was assumed that those genes that did not exhibit a strong correlation have no mutual regulation. Sushmita et al. (2009) proposed a gene selection method based on a given threshold cutoff. Simply stated, given a threshold value ä, if the absolute value of the correlation coefficient between the genes is greater than ä, then a correlation between the genes exists.

Similarly, Equation 3 was applied to construct a correlation matrix A, in which each element a_{ij} of matrix A represented the local maximum similarity. Once the correlation matrix was constructed, k number of largest absolute element values were be selected from matrix A as strong interrelation elements. If the selected element is greater than 0, a positive correlation existed between the two corresponding genes, denoted by 1. If the chosen element was less than 0, a negative correlation between the corresponding two genes was observed, denoted by -1. Otherwise, no relationship between the corresponding two genes was found, denoted by 0. Therefore, this gene interrelationship matrix can be converted into an adjacency matrix C as follows:

$$c_{ij} = \begin{cases} -1 & \text{if the element is selected, then correlation coefficien t is negative} \\ 1 & \text{if the element is selected, then correlation coefficien t is positive} \\ 0 & \text{if the element is not selected} \end{cases} \qquad \text{(Equation 4)}$$

A gene effect network can be constructed using a gene adjacency matrix C. If c_{ij} = 1, then a positive correlation between gene i and gene j exists, and there will be a real edge between node i and node j in the effect network. If c_{ij} = -1, then a negative correlation between gene i and gene j exists, and there will be a virtue edge between node i and node j in the effect network. If c_{ij} = 0, no correlation exists between gene i and gene j, hence there cannot be an edge between node i and node j on the effect network.

RESULTS

Liu (2013) proposed a method to bicluster gene expression data. A gene network was constructed from Liu's biclustering results using the above methods. Figure 3 shows the gene network constructed from the first and 29th output biclusters when q = 20 (Liu, 2013). In the regulatory network, the solid line represents the positive correlation between the genes, indicating that there may exist mutual regulations between these two genes where the expression of one gene would promote the expression of other genes. The dash line represents a negative correlation between the genes, indicating that the expression of a gene can inhibit the expression of other genes.

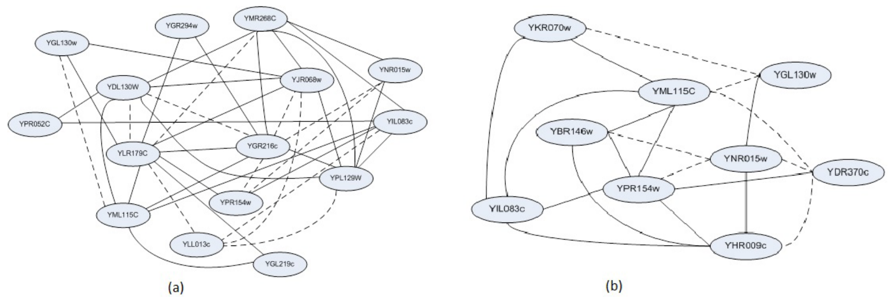

Figure 3. Gene Network constructed by the biclusters.

DISCUSSION

This study briefly described the traditional method of constructing gene regulatory networks, and then preliminarily explored a new approach - the construction of gene regulatory networks based on biclustering results. Based on correlation analysis of biclustering gene expression data, three steps were used to construct a gene regulatory network: 1) Biclustering analysis of gene expression data, 2) correlation analysis between the candidate genes, and 3) construction of gene regulatory networks by correlation analysis. According to the local similarity of gene expression data, this study proposed a new approach for measuring expression similarity relation between genes. Finally, this study used previous biclustering results to construct two gene regulatory network diagrams employing the new method.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation (#61572368).

REFERENCES

An J, Liew AW and Nelson CC (2012). Seed-based biclustering of gene expression data. *PLoS One* 7: e42431. http://dx.doi.org/10.1371/journal.pone.0042431

Bickel DR (2005). Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics* 21: 1121-1128. http://dx.doi.org/10.1093/bioinformatics/bti140

Butte AJ and Kohane IS (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Proceedings of Pacific Symposium on Biocomputing (PSB 2000), 5: 415-426.

Das M and Borah B (2014). Biclustering of gene expression data using a two -phase method. *Int. J. Comput. Appl.* 103: 6-10.

D'haeseleer P, Liang S and Somogyi R (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16: 707-726. http://dx.doi.org/10.1093/bioinformatics/16.8.707

Friedman N (2004). Inferring cellular networks using probabilistic graphical models. *Science* 303: 799-805. http://dx.doi.org/10.1126/science.1094068

Godsey B (2013). Improved inference of gene regulatory networks through integrated Bayesian clustering and dynamic modeling of time-course expression data. *PLoS One* 8: e68358. http://dx.doi.org/10.1371/journal.pone.0068358

Hecker M, Lambeck S, Toepfer S, van Someren E, et al. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems* 96: 86-103. http://dx.doi.org/10.1016/j.biosystems.2008.12.004

Hughes JD, Estep PW, Tavazoie S and Church GM (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.* 296: 1205-1214. http://dx.doi.org/10.1006/jmbi.2000.3519

Kato M, Tsunoda T and Takagi T (2001). Lag analysis of genetic networks in the cell cycle of budding yeast. *Genome Informatics* 12: 266-267.

Lei YS, Shi DH and Wang YF (2004). Bioinformatics of gene regulatory networks. *Chinese Journal of Nature* 26: 7-12.

Liu F (2013). Time-lagged co-expression gene analysis based on biclustering technology. *Biotechnol. Biotechnol. Equip.* 27: 4031-4039. http://dx.doi.org/10.5504/BBEQ.2013.0058

McAdams HH and Arkin A (1997). Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* 94: 814-819. http://dx.doi.org/10.1073/pnas.94.3.814

Petricka JJ and Benfey PN (2011). Reconstructing regulatory network transitions. *Trends Cell Biol.* 21: 442-451. http://dx.doi.org/10.1016/j.tcb.2011.05.001

Shmulevich I, Yli-Harja O and Astola J (2002). Inference of genetic regulatory networks under the best fit extension paradigm. In proceedings of the IEEE- EURASP workshop on nonlinear signal and image processing, 3-6.

Simon I, Barnett J, Hannett N, Harbison CT, et al. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* 106: 697-708. http://dx.doi.org/10.1016/S0092-8674(01)00494-9

Spears WM (1996). Simulated annealing for hard satisfiability problems. DIMACS series in discrete mathematics and theoretical computer science 26: 533-558.

Sushmita M, Haider B and Ranajit D (2009). Gene interaction--an evolutionary biclustering approach. *Inf. Fusion* 10: 242-249. http://dx.doi.org/10.1016/j.inffus.2008.11.006

Lee TI, Rinaldi NJ, Robert F, Odom DT, et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298: 799-804. http://dx.doi.org/10.1126/science.1075090

Weaver DC, Workman CT and Stormo GD (1999). Modeling regulatory networks with weight matrices. In: Proceedings of Pacific Symposium on Biocomputing (PSB1999). 4: 112-123.

Wyrick JJ and Young RA (2002). Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.* 12: 130-136. http://dx.doi.org/10.1016/S0959-437X(02)00277-0

Xu XJ, Wang LS and Ding DF (2003). Estimating gene regulatory networks from yeast expression time series. *Acta BioChimica Biophhys. Sin.* 35: 707-716.

Yi D, Yang MS and Li HZ (2003). The construction of gene network by correlation analysis. *Chin. J. Health Statistics* 20: 144-146.