Inference of Gene Regulatory Networks via Boolean Networks Using Regression Coefficients

Ha Seong Kim[1] Ho Sik Choi[2] Jae K. Lee[3] Taesung Park[2]

[1] *Interdisciplinary Program in Bioinformatics , Seoul National University, Seoul, Korea*

[2] *Department of Statistics, Seoul National University, Seoul, Korea*

[3] *Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, USA*

Email : khs@biostats.snu.ac.kr, pinebird@biostats.snu.ac.kr, jaeklee@virginia.edu, tspark@stats.snu.ac.kr

ABSTRACT: Boolean networks(BN) construction is one of the commonly used methods for building gene networks from time series microarray data. However, BN has two major drawbacks. First, it requires heavy computing times. Second, the binary transformation of the microarray data may cause a loss of information. This paper propose two methods using liner regression to construct gene regulatory networks. The first proposed method uses regression based BN variable selection method, which reduces the computing time significantly in the BN construction. The second method is the regression based network method that can flexibly incorporate the interaction of the genes using continuous gene expression data. We construct the network structure from the simulated data to compare the computing times between Boolean networks and the proposed method. The regression based network method is evaluated using a microarray data of cell cycle in *Caulobacter crescentus*.

1 INTRODUCION

Gene regulatory network plays a key role to describe biological phenomena. Recently, a variety of high-throughput experimental techniques have been developed with the ability to observe the experssion of many genes simultaneously. For example, some recent microarray studies have been performed for investigating multiple time-point pathways, including yeast sporulation [1], yeast cell cycle [2], E. coli heat shock [3], Caulobacter crescentus cell cycle [4]. Using these high-throughput time series microarray data, a number of different approaches to gene regulatory network modeling have been introduced, including linear models [5], Boolean networks [6-8], Bayesian networks [9,10] and neural networks [11]. In this paper, a new variable selection method is proposed to dramatically reduce the computing times in Boolean network construction. In addition, a new regression based network method is proposed to build gene regulatory networks.

1.1 Boolean networks

Boolean networks as models of gene regulatory networks were first introduced by Kauffman in 1969 [6]. In this model, gene expression is quantized to only two levels: ON and OFF. A Boolean network $G(V,F)$ is defined by a set of nodes $V=\{x_1,\ldots,x_n\}$ and a list of Boolean functions $F=\{f_1,\ldots f_n\}$. A Boolean function $f_i(x_1, \ldots, x_k)$ with k specified input nodes(indegree) is assigned to node x_i [7, 8]. Regulation of nodes is defined by the set F of Boolean functions. In detail, given the value of the nodes V at time t, the Boolean functions are used to update the value of the nodes at time t+1. We use the Consistency Problem [7] and Best-Fit Extension Problem [8,12] method to find Boolean functions.

1.2 Advantages of Boolean networks

- Boolean network model can be used to explain the dynamic behavior of living systems. Simplistic Boolean formalism can represent the realistic complex biological phenomena by reducing the noise level in biological systems [13].
- Boolean algebra is a prosperous science, providing a vary rich set of algorithms already available for supervised learning in the binary domain, such as logical analysis of data [14,15].

1.3 Drawbacks of Boolean networks

- To construct accurate networks, Boolean networks require a large number experimental data sets that have different initial status.
- Boolean network construction requires heavier computing time as more genes are included.
- Binarizing gene expression data results in significant loss of information from the observed expression levels are concerned.

1.4 Variable selection in Boolean networks using linear regression

In Boolean networks, the Boolean function, f_i , consists of the k variables that affect the ith gene expression level. The ith gene and k genes are highly correlated in Boolean function f_i. We can select the significant genes for ith gene using significant test with simple regression method. Figure 1 (a) is a simple network and (b) is a wiring diagram which gives an explicit way of implementing the update procedure. This network is represented by a Boolean functions (1), f_1 and f_2.

$$\begin{aligned} X_{1,t+1} &= f_1(X_2, X_3) = X_{2,t} \wedge X_{3,t} \\ X_{2,t+1} &= f_2(X_3) = X_{3,t} \end{aligned} \qquad (1)$$

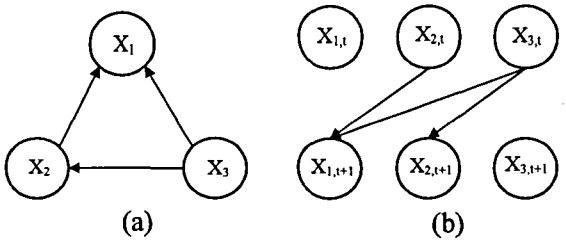


Figure 1: An example of simple network

In a regression analysis, fit a simple regression model with $X_{1,t+1}$ as a dependent variable and $X_{1,t}$, $X_{2,t}$, $X_{3,t}$ as independent variables. Then, all variables except $X_{1,t}$ have high significant p-values. In the same way, fit a simple regression model with $X_{2,t+1}$ as a dependent variable and $X_{1,t}$, $X_{2,t}$, $X_{3,t}$ as independent variables . Then we can obtain only one significant variable, $X_{3,t}$. Based on this idea, we introduce a variable selection method to reduce computing time in Boolean networks. Our proposed method will be detailed in the section 2.1.

1.5 Representation of gene regulatory networks using regression method

The idea of constructing the network diagram via regression model is similar to that of path analysis. Path analysis is a method to analysis the relationship between variables in a path diagram. In particular, each coefficient of pathway connections can be inferred by using regression coefficients *[16]*. Path analysis was developed as a method of decomposing correlations into different pieces for interpretation of effects from a interaction diagram. However, our network algorithm is proposed to construct diagram from the selected regression model. For example, consider the following regression equations (2).

$$\begin{aligned} X_{1,t+1} &= b_{1,2}X_{2,t} + b_{1,3}X_{3,t} + e_1 \\ X_{2,t+1} &= b_{2,3}X_{3,t} + e_2 \\ X_{3,t+1} &= e_3 \end{aligned} \qquad (2)$$

Equations (2) can be represented as network diagram in Figure 2. Proposed regression based network method select regression models (2) and represent network using selected models, as shown in Figure 2.

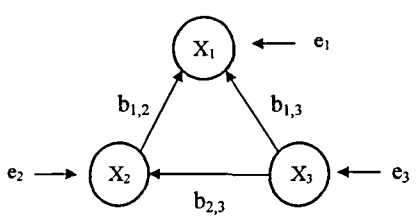


Figure 2: Network diagram for equation (2)

In regression model (2), e1, e2, e3 are disturbances, or, residual terms. These means stray causes, or causes outside the model. e_i does not stand for measurement error, which is assumed to be zero in path analysis. In this paper, we utilize the regression coefficient to represent interactions between genes. Its method will be described in next section.

2 METHOD

Notation

Let n be the total number of genes, i=1,2,...,n
k is number of indegree, l=1,2,...,k
m is total number of time point, t=1,2,...,m
$X_{i,t}$ is ith node at time t

2.1 Regression based variable selection method in Boolean networks

In Boolean function, $X_i = f_i(X_{i1},...,X_{ik})$, essential variables $(X_{i1},...,X_{ik})$ have significant coefficient when fit simple regression model which X_i as response variable and X_{il} as predict variable. If the Boolean network algorithm use the selected variables to find Boolean function, the computing time of the algorithm will be reduced significantly.

1. For the ith gene, fit simple regression model (3) for n pair of (i, j), j=1,2,...,n
$$X_{i,t+1} = \beta_0 + \beta_1 X_{j,t} \qquad (3)$$
2. Select all genes having significant coefficient with ith gene

In our Boolean network program, best-fit extension problem method is used to find Boolean function [8,12]. Instead of searching for the connections among all (n) genes, we only use selected variables to find the Boolean function for ith gene. This method effectively eliminates the less informative genes which do not affect the ith gene.

2.2 Regression based network method

Regression-based network construction is a network building method using regression model which have the largest adjusted R-square value(s).

1. Set the maximum indegree k
2. For the ith gene, construct all possible ${}_nC_l$ (l=1,..,k) models, where the model has following form (4) : with the ith gene as response variable and k genes as predictor variables
$$X_{i,t+1} = \beta_0 + \sum_{l=1}^{k} \beta_{il} X_{l,t} \qquad (4)$$
3. Select a model having the largest adjusted R-square value.
4. Determine the directions and positive, negative effects between the variables

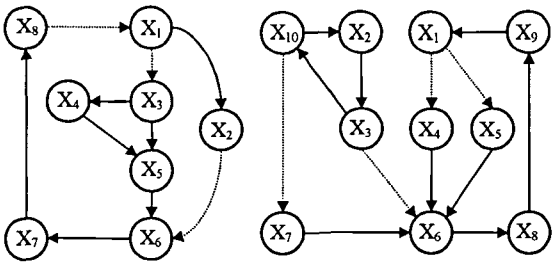
$\beta_{il} > 0$ means positive effect $(X_{i,t+1} \longleftarrow X_{l,t})$

$\beta_{il} < 0$ means negative effect $(X_{i,t+1} \longleftarrow\!\cdots\!\cdots X_{l,t})$

3 RESULT

3.1 Simulation data

We simulate two set of artificially generated network with binary data (Figure 3). Each of these data sets contains eight genes with ten time points and it has four different initial experiments without noise. The only difference of network 1 and network 2 is maximum indegree. Network 1 has two maximum indegree and network 2 has four maximum indegree level.

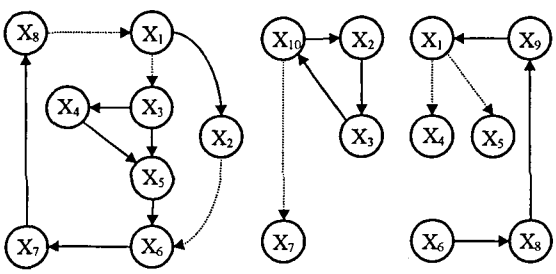


(a) Network1 (b) Network2

Figure 3 : Generated simulation network

3.1.1 Regression based variable selection method in Boolean networks

We set the maximum indegree, k=3, in Boolean network program. Figure 4 is the result of network examples using regression based variable selection method in Boolean network. Network 1 shows exactly same structure with generated network (Figure 3 (a)). In network 2, however, the edges having more than maximum indegree, k=3, were not found.

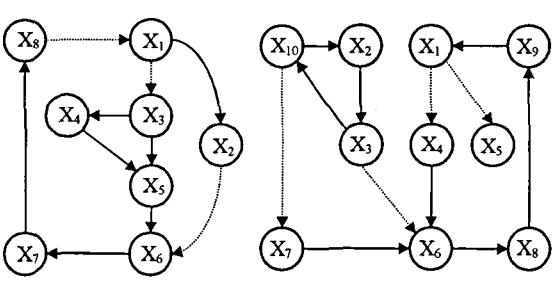


(a) Network 1 (b) Network 2

Figure 4 : Result of networks using Boolean network with regression based variable selection method.

3.1.2 Regression based network

Figure 5 shows the result of network structure using regression based network method. Network 1 (Figure 4 (a)) has the same topology with the original network structure. In network 2 (Figure 4 (b)), the edge between X_5 and X_6 was not found because the maximum indegree k=3, but it provided more accurate result than that of Boolean network.

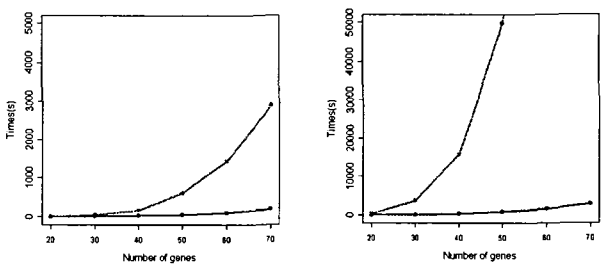


(a) Network1 (b) Network2

Figure 5 : Result of networks using regression based network method

3.2 Compariosn of computing times

To compare the computing times using proposed variable selection method, we simulate the binary data with several genes. The simulation was performed when the maximum indegree k=3 and 4. Figure 6 shows the result of time consumption using the original Boolean networks (b) and Boolean networks with proposed variable selection method (a). The result of original Boolean network algorithm shows 50000 second when it contains total 50 genes with maximum indegree k=4. But the Boolean network algorithm with regression based variable selection method has about 600 second. Although the proposed method still depends on the relationship among individual genes, the method could reduce the computing time in Boolean network construction significantly.



(a) (b)

Figure 6 : Computing times using Boolean network with regression based variable selection method (a) and original Boolean networks (b). Red line and blue line represent maximum indegree k=4 and 3, respectively

3.3 *Caulobacter crescentus* example

Our regression-based network method is applied to using *Caulobacter crescentus* gene expression data. *Caulobacter crescentus*, an innocuous, single celled organism that lives in water, was been sequenced in 2001. The microbe has one circular chromosome containing some 3,700 genes. Dispite this simplicity, a single *Caulobacter crescentus* cell divides into two cells that differ in structure and function – it is an ideal model system for the mechanisms of asymmetric cell division and has been studied thoroughly. Like stem cells, the stalked cell continually gives rise to a new swarmer cell at each cell division. The stalk is a thin cylinder growing out of one pole of the cell. The motile swarmer cell has a flagellum. Caulobacter is an good model system to study the mechanism of an asymmetric cell division [17,18]. Also, a strong correlation between cell-cycle dependent transcription and protein synthesis was reported [19]. We thus experiment with our regression-based network method for use this time series microarray gene expression data set which containing 11 time point slides on 1444 ORF probes [4].

The ctrA gene regulate many genes in several functional categories and provide the clusterd gene set [4]. Figure 7 is the ctrA regulatory network governing Caulobacter cell cycle progression. We construct the gene regulatory networks using regression based network method for three different set of genes: flagella biogenesis, DNA methylation and cell division. Before build the network structure (Figure 7), we include the ctrA gene in each group

to compare the published information. Figure 8 is the result of constructed network for three groups. The edges in Figure 8 show the ctrA gene regulate the fla, ccrM, ftsZ genes as a transcription facter [17,18].

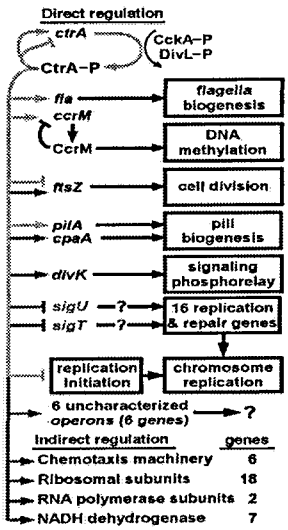
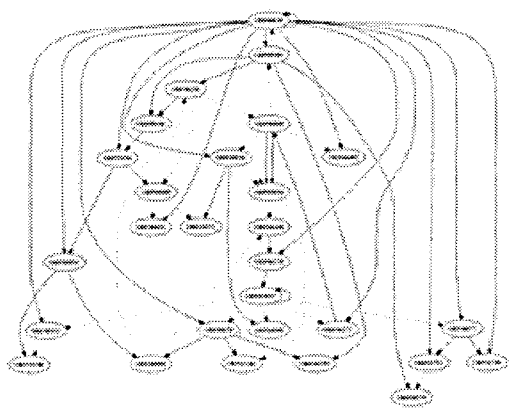
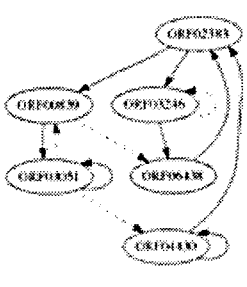
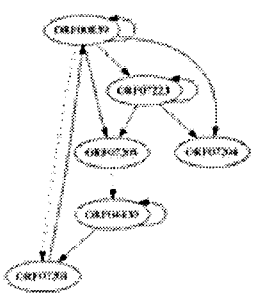


Figure 7 : The CtrA regulatory network governing Caulobacter cell cycle progression. Phosphorylated CtrA autoregulates its own transcription and activates or represses the transcription of multiple sets of genes. CtrA bound to sites in the origin of replication inhibits replication initiation [4].



(a) Fragella biogenesis





(b)DNA methylation (c) Cell division

Figure 8 : Result of networks using regression based network method in Caulobacter cell-cycle progression. (a) Interaction of genes in fragella biogenesis. ctrA:ORF00839, fla:ORF02752. (b) Interaction of genes in DNA methylation ctrA:ORF00839, ccrM:ORF03051 (c) Interaction of genes in Cell division. ctrA:ORF00839, ftsZ:ORF07201

4 DISCUSSION

Boolean network construction is useful to build a gene regulatory network because if gene expression data contain a considerable amount of noise, the binary transformation of the expression data can significantly reduce the error (with proper normalization methods) [14]. However, in Boolean networks, the higher indegree value is, the heavier is its computing time (exponentially). Therefore, our proposed variable selection method is effective in studying the large scale gene regulatory network analysis due to its computational advantage. But Boolean network construction still has a drawback that may cause a loss of information for the data binarization.

Regression based network method is simple and efficient in the sense that this method directly utilizes the continuous gene expression ratio data both to improve the network estimation accuracy without loss of information and to reduce computing time using statistical approach. It could be useful to compare the dynamics between experiments because this approach does not need several experiments with different initial condition to fit a single network model.

Our proposed regression based network construction selects the regression model with the largest adjusted R-square. It, however, does not use other models that may have similar high adjusted R-square values. In the future studies, we plan to extend the probabilistic approach that can consider such multiple competing regression models [20]. The use of probabilistic framework [20], promoter sequence analysis [21] and time-lagged clustering [22] may also improve the regression based network model to capture the complex dynamics in biological systems.

REFERENCES

[1] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, The transcription program of sporulation in budding yeast, *Science*, 282, 699, 1999

[2] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein & B. Futcher, Comparehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9, 3273-3297, 1998

[3] C. S. Richmod, J. D. Glasner, R. Mau, H. jin and F. R. Blattner, Genomi-wide expression profiling in Escherichia coli K-12, *Nucleic Acids Research*, 27 3821, 1999

[4] M. T. Laub, H. H. McAdams, T.Feldblyum, C.M. Fraser, L. Shapiro, Global Analysis of the Genetic Network Controlling a Bacterial Cell Cycle, *Science*, Vol 290, 2000

[5] P. D'Haeseleer, X. Wen, S. Fuhrman, & R. Somogyi, Linear modeling of mRNA expression levels during CNS development and injury, *Pacific Symposium on Biocomputing*, 4, 41-52, 1999

[6] S. A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets, *J. Theor. Biol.*, 22, 437-467, 1969

[7] T. Akutsu, S. Kuhara, O. Maruyama, & S. Miyano,

Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions, *the 9th Annual ACM-SIAM symposium on Discrete Algorithms (SODA'98)*, pp. 695-702, 1998

[8] I. Shmulevich, O. Yli-Harja, J. Astola, Inference of Genetic Regulatory Networks Under the Best-fit Extension Paradigm, Proc. *Workshop on Nonlinear Signal and Image Processing*, 2001

[9] K. Murphy, & S. Main, Modelling gene expression data using dynamic Bayesian networks, Technical Report, University of California, Berkeley, 1999

[10] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, Using bayesian networks to analyze expression data, *J. Computational Biology*, 7(3), 601-620, 2000

[11] D. C. Weaver, C. T. Workman, & G. D. Stormo, Modeling regulatory networks with weight matrices. *Pacific Symposium on Biocomputing*, 4, 112-123, 1999

[12] E. Boros, T. Ibaraki, & K. Makino, Error-Free and Best-Fit Extensions of partially defined Boolean functions, *Information and Computation*, 140, 254-283, 1998

[13] S. Huang, Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery, *J. Mol. Med.*, 77 469-489, 1999

[14] I. Shmulevich, W. Zang, Binary analysis and optimization-based normalization of gene expression data, *Bioinformatics*, Vol. 18 no. 4 pages 555-565, 2002

[15] E. Boros, P. L. Hammer, J. N. Hooker, Boolean Regression, *GSIA Working Papers*, 1991

[16] Otis Dudley, Introduction to structural equation models, *New York: Academic Press.* ISBN: 0-1222-41509, 1975

[17] C. Jacobs, Regulatory proteins with a sense of direction: cell cycle signalling network in Caulobacter, *Molecular Microbiology* , 51(1) 7-13, 2004

[18] U. Jenal, C. Stephens, The Caulobacter cell cycle: timing, spatial organization and checkpoints, *Curr Opin Microbiol,* 5: 558-563, 2002

[19] B. Grunenfelder, G. Rummel, J. Vohradsky, D. Roder, H. Langen, U. Jenal, Proteomic analysis of the bacterial cell cycle. *Proc Natl Acad Sci USA*, 98:4681-4686, 2001

[20] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zang, Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, vol 8, no. 2, pp. 261-274, 2002

[21] P. M. Haverty, U. Hansen, and Z. Weng, Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification, *Nucleic Acids Res*, 32(1): 179 - 188, 2004

[22] L. Ji and K.-L. Tan, Identifying time-lagged gene clusters using gene expression data, *Bioinformatics*, 21(4): 509 – 516, 2005

[23] T. Akutsu, S. Miyano, Identification of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model, Proc. *Pacific Symposium on Biocomputing'99* (PSB'99), 17-28, 1999

[24] K. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak & J.J. Tyson, Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle, *Molecular Biology of the Cell*, 11, 369-391, 2000

[25] R. Somogyi, & C. Sniegoski, Modeling the complexity of gene networks: Understanding multigenic and pleiotropic regulation, *Complexity, I*, 45-63, 1996