An integrated Bayesian network framework for reconstructing representative genetic regulatory networks.

Philhyoun Lee, Doheon Lee*, Kwanghyung Lee

Department of BioSystems, KAIST, Daejeon, Korea

*To whom correspondence should be addressed. E-mail: dhlee@kaist.ac.kr

Abstract

In this paper, we propose the integrated Bayesian network framework to reconstruct genetic regulatory networks from genome expression data. The proposed model overcomes the dimensionality problem of multivariate analysis by building coherent sub-networks from confined gene clusters and combining these networks via intermediary points. Gene Shaving algorithm is used to cluster genes that share a common function or co-regulation. Retrieved clusters incorporate prior biological knowledge such as Gene Ontology, pathway, and protein protein interaction information for extracting other related genes. With these extended gene list, system builds genetic sub-networks using Bayesian network with MDL score and Sparse Candidate algorithm. Identifying functional modules of genes is done by not only microarray data itself but also well-proved biological knowledge. This integrated approach can improve the reliability of a network in that false relations due to the lack of data can be reduced. Another advantage is the decreased computational complexity by constrained gene sets. To evaluate the proposed system, *S. Cerevisiae* cell cycle data [1] is applied. The result analysis presents new hypotheses about novel genetic interactions as well as typical relationships known by previous researches [2].

Introduction

Cells have tremendous diversity in its shape and function but share exactly same genetic blueprint. To understand these distinct cellular activities of an identity, we need to understand the mechanisms of protein synthesis regulation. Formally, these relationships are represented as *a genetic regulatory network*, the set of mutually activating and repressing genes and gene products and their interactions [3].

Currently, microarray data is being widely used for reconstructing a genetic regulatory network [4,5,6,7]. Inferring relationships from these transcript levels of thousands of genes confronts several challenges. First, not all of the expression levels of genes can be measured. Intercellular components like hormone, and regulation part by post-translation or small molecules are missed, too. Most of all, measured data is very noisy and the number of experiment sample is far from sufficient for drawing a network structure reliably. A variety of techniques have been applied to overcome these obstacles. These includes discrete models, such as a Boolean network [8, 9] and continuous models based on differential equation, such as an continuous recurrent neural network [10] or power-law formalisms [11]. Probabilistic graphical models like a Bayesian network, also known as a causal network, have been used, too [2, 12, 13, 14].

A Bayesian network [15] is a directed and acyclic graph that encodes a joint probability distribution based on the properties of conditional independence between variables. It can describe complex stochastic processes, thus appropriate for learning from noisy observations. Also Bayesian

This work was supported by the Korean Systems Biology Research Grant (M10309020000-03B5002-00000) from the Ministry of Science and Technology.

network is particularly useful for dealing structures composed of locally interacting components like biological processes. Previous research done by Friedman et al. [2] tried to overcome the dimensionality problem by seeking the set of plausible networks and characterizing features that are common to most of these networks rather than picking up a single model. Here, we present the model uses this partial model of Bayesian network as a framework and expands

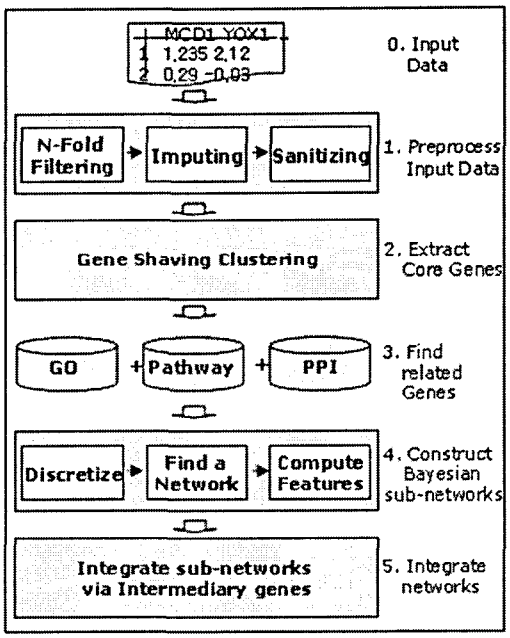

Fig. 1. System architecture of the integrated Bayesian network framework

it by incorporating clustering methods and well-known biological knowledge, together (**Fig. 1**).

Main idea is to reduce the number of target genes for learning one Bayesian network. For this, the system organizes genes into coherent sub-network modules. Other members of these clustered functional sub-sets are retrieved using diverse biological resources such as Gene Ontology [16], pathway [17], and protein protein interaction [18] information. With these reduced number of genes in each set, the proposed system can construct the genetic sub-network more reliably. Bayesian networks are learned by MDL score [19] and Sparse Candidate algorithm [20] for search efficiency.

We use Gene Shaving algorithm [21,22] to identify subsets of genes that are co-regulated or have similar function. Gene Shaving identifies genes with coherent expression patterns and large variation across samples. This method differs from other clustering algorithms in that genes are permitted to belong to more than one cluster. Our system utilizes these overlapped genes as intermediary points to combine coherent sub-networks as a whole at the last step. Also the number of genes in each result cluster is relatively small, approximately several to dozens. So even after expanding genes from biological resources, the number of genes in one set to be reconstructed together is maintained below two hundreds.

Methods (Materials and Methods/ Systems and Methods etc)

Data Preprocessing

If the number of genes in microarray expression data is N and the number of samples is p, data can be considered as an N x p matrix, $X=x_{ij}$ (i=1,..,N, j=1,..,p) The proposed system selects genes that show certain amount of expression change (here 4-fold). For imputing missing values, simple row average estimation is used [23]. The genes having relatively large consecutive omissions (in the experiments, we set 3) are sanitized.

Clustering using Gene Shaving algorithm

Step 1 : The system seeks nested clusters S_k of size k having the highest variance of the signed column mean. Since an eigen gene denotes the normalized linear combination of genes with largest variance across the samples, the largest principal component of the genes is computed. To get the nested clusters from size k=N to k=1, α fraction of the genes having lowest correlation with the leading principal component are shaved off from the current gene set. This whole process is iterated until only one gene remains.

Step 2 : The optimal cluster size k is estimated in the direction of high coherence between members of the cluster. The between and within variances of each cluster are calculated as [21,22]:

$$V_W = \frac{1}{p}\sum_{j=1}^{p}\left[\frac{1}{k}\sum_{i=1}^{k}(x_{ij}-\bar{x}_j)^2\right]$$

$$V_B = \frac{1}{p}\sum_{j=1}^{p}(\bar{x}_j-\bar{\bar{x}})^2$$

$$R_2 = 100\,\frac{V_B/V_W}{1+V_B/V_W}$$

A large value of variance ratio R_2 implies that a cluster is composed of tightly consistent genes. For differentiating real patterns from

spurious ones, the cluster that shows the largest gap with the randomly permuted matrix is selected as the optimum.

$$\text{argmax}_k \left(\text{Gap}(\text{org}_k(R_2) - \text{mean}(\text{permuted}_k(R_2)) \right)$$

Step 3 : To remove the effect of genes in the previously chosen clusters, the original data matrix is orthogonalized and step 1 and step 2 are repeated until the desired number of clusters, C, is found

Extending gene lists by Biological Knowledge

Using well-proven biological knowledge, the proposed system looks for genes that may have functional relationship with each clustered genes. Gene Ontology from SGD [16], pathway information from KEGG (Kyoto Encyclopedia of Genes and Genomes) [17], and protein protein interaction information from DIP (Database of Interacting Proteins) [18] are used. Each cluster includes found genes as a member.

Reconstruct the genetic regulation network by Bayesian network

Step 1 : In this stage, the proposed system draws a Bayesian network for each cluster. Let's define cluster j as a finite set $U_j=\{X_1, .., X_n\}$ of discrete random variables where a variable X_i denotes the expression level of a gene in the cluster. First, the system discretizes gene expression measurements into 3 categories, under-expressed, constant, and over-expressed (-0.5 and +0.5 is used as a threshold). Computed values are stored as an N_j x p_j data matrix, D_j. Then, the system search for an equivalence class of networks [24] that consist of random variables in U_j and best matches D_j.

We uses MDL score as an evaluation function for the posterior probability of a network [25]. Since an optimal one balances the complexity of the network with the degree of fitness, score can be calculated as [19]:

$$DL(U, D, G) = DL_{net}(U, G) + DL_{data}(D, G)$$

$$DL_{net}(U, G) = \sum_i (\log \|X_i\| + (1+|Pa(X_i)|)\log n) + \frac{\log N}{2} \sum_i \| Pa(X_i) \| (\|X_i\|-1)$$

$$DL_{data}(D, G) = N \sum H(X_i) - N \sum I(X_i; Pa(X_i))$$

$\|X\|$ = the cardinality of values each gene
$H(X|Y)$ = the conditional entropy of X given Y
$I(X;Y)$ = the mutual information between variables X and Y

A greedy hill climbing with random restart algorithm is used as a heuristic search strategy. For efficient learning, we restrict our search to the small number of candidate parents based on correlation between two variables. The limitation caused from pair-wise selection is compromised by measuring the discrepancy between estimated $P_B(X_i, X_j)$ and empirical estimate $P(X_i, X_j)$ [20].

$$M_{Disc}(Xi, Xj \mid B) = D_{KL}(P(Xi, Xj) \| P_B(Xi, Xj))$$

Step 2 : With the network built at step 1, 100-fold Bootstrap method [26] is used for finding confident Markov relations. The system samples p_j column vectors from D_j with replacement and makes m new N_jxp_j data matrix, D_{Bj} (B=1,..m, m=100). Learning procedure is applied for each new data set and reliability is estimated as:

$$\text{conf}(f) = \frac{1}{m} \sum_{i=1}^{m} f(G_i)$$

f(G) is 1 if f is a feature in G, and 0 otherwise. System draws the network with features that have higher confidence than 0.8 and connects the new component to sub-graphs with dotted line if edges between them have confidence above 0.6. For each cluster, network reconstruction is done via step1 and step2.

Step 3 : Networks are combined together if they share genes as components. For example, networks from cluster 1 and cluster 3 share gene CLN2 and MCD1. Two networks can be linked via a intermediary.

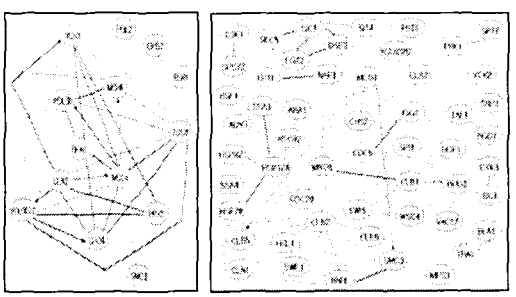

Fig 2. networks from cluster 1 and cluster 3

Results

We used an *S.cerevisiae* gene expression data set consisting of 79 microarray samples that measured under different cell cycle synchronization methods [1]. The data contains mRNA levels of 6177 ORFs. From this set, we reconstructed genetic regulatory networks and analyzed the results. Supplementary data are available at WWW site:

http://bioif.kaist.ac.kr/~phlee

Experimental Result

The system computed 30 clusters based on Gene Shaving algorithm. 10% of genes were shaved each step and permutation no was 5 (**Fig. 3**). For determining the optimal size for each cluster, Gap value was calculated. Cluster 2 is illustrated as an example (**Fig. 4**). For each cluster, prior biological knowledge was incorporated. Some clusters have distinguishable characters based on GO. Cluster 2 consists of mainly "cell wall organization and biogenesis" genes. Cluster 4 has only genes related with "galactose metabolism". Cluster 8 includes genes functioning "response to copper ion". And cluster 11 has "chromatin assembly" genes and so on. Each learned sub-network is a partial directed acyclic graph. As an example, sub-networks from cluster 2 are shown (**Fig. 5**)

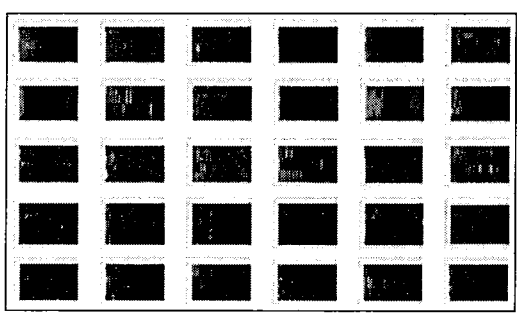

Fig. 3. The result of 30 clusters

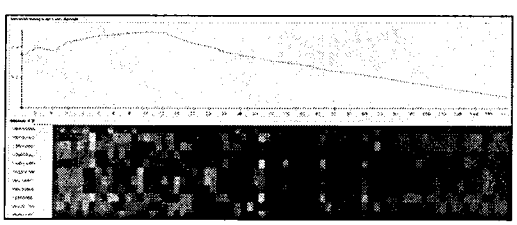

Fig. 4. Gap function value of cluster 2

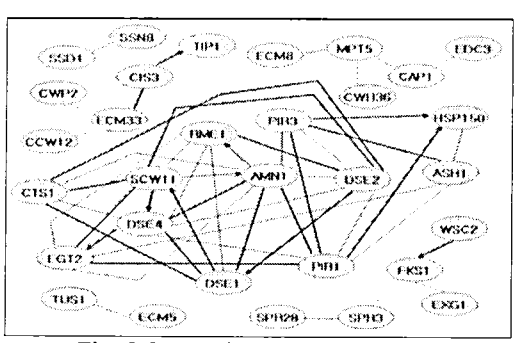

Fig. 5. Learned network of cluster 2

Biological Analysis

With few exceptions (about 20% new relations), each of inferred networks recovered features drawn at the previous research [2]. If mediator gene is missed in the module, indirect dependencies were denoted as a direct link or those relations were lost. This reveals the importance of identifying functional modules.

In the cluster 10, features with confidence 1.0 were found (**Fig. 6**). These relations between four genes (HTA1, HTB1, HHT1, HHF1) were missed at the previous research [2]. They are revealed as members of a 'nucleosome protein complex', which performs a key role for mRNA transcription. This shows that our integrated method can recover exquisite biological relationships which were lost by previous approaches [2].

C	X_i	X_j
1.0	HTA1(YDR225W)	HHF1(YBR009C)
1.0	HTB1(YDR224C)	HHF1(YBR009C)
1.0	HTB1(YDR224C)	HHT1(YBR010W)
1.0	HHT1(YBR010W)	HTA1(YDR225W)
0.99	HTB1(YDR224C)	HTA1(YDR225W)
0.98	HHF1(YBR009C)	HHT1(YBR010W)

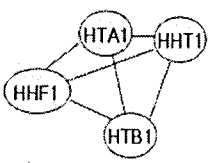

Fig. 6. features with high confidence in cluster 10

Constructed genetic network implies some novel biological hypotheses about genetic interaction. Some of top-ranked dominant genes (MCD1, CLN2, SRO4, YOL007C, YOX1, POL30, MSH6) in the ordering relations [2] were found in a sub-network drawn from cluster 1. Intuitively, these genes are potential inducers of the cell-cycle process according to the characteristic of Gene Shaving clustering [21,22]. TOS4 and PRY2 are members of this sub-network and the only unclassified components currently based on MIPS[27]. We inferred that TOS4 and PRY2 may function as a key regulator like MCD1, RFA1, CLN2, SRO4, YOL007C, YOX1, POL30, and MSH6. From literature [28], we can find that TOS4 is established as a transcription factor that has 230 target genes and also regulated by SBF. PRY2 has strong similarity to PRY1, which is an un-annotated gene itself. But PRY1 was identified in association with YPR086W(SUA7, SOH4), a general RNA polymerase II transcription factor [29]. This strongly indicates that PRY2 can be related with transcription activity. This hypothesis needs to be investigated further computationally and also experimentally.

Complementary evidence of functional annotation is also provided. SCW11 is known to have week similarity to the protein, glucanase but

currently unclassified gene. In the cluster 2, SCW11 is directly linked with DSE2, DSE4, and CTS1. All 3 genes are kinds of Glycosidases (enzymes hydrolysing O- and S-glycosyl compounds) and localize at the cell wall. Strong dependencies among those genes (**Table 1**) drawn from a microarray experiment strongly indicates that SCW11 carries out enzyme functionality as a glycosidase, too.

gene	C	known function
DSE2	0.96	glucan 1,3-beta-glucosidase activity
DSE4	0.93	endo-1,3-beta-glucanase
CTS1	0.95	endochitinase

Table 1. Confidence of Markov relations

Discussion

The proposed system reconstructs networks from microarray gene expression data by combining Gene Shaving clustering [21,22] method and biological knowledge with Bayesian network [15].

The main advantage of this approach is that false positive relations due to the lack of data can be reduced by pre-selecting sets of genes through the microarray data itself as well as the previously well proved biological knowledge, while the trade-off is that sub-network can miss the related gene components. Also, this approach is capable of reducing computational complexity. The result of *S. Cerevisiae* data [1] analysis shows this pros and cons.

The authors are in the initial stage of on-going project. We consider following improvements as further works. I) Identifying functional sub-modules correctly contributes to overall success of process. We are working on improving methods for grouping related genes more reliably. II) Genetic Regulators such as a transcription factor or a signaling molecule need to be stated clearly. And the regulation form like activation or inhibition can be denoted. III) Currently the system discretizes gene expression values into 3 catagories according to the pre-defined threshold. This simple discretization can cause the loss of information in abundance. We plan to apply finder ones [30, 31]. IV) Improved heuristic search can be applied. V) The system currently integrates biological information from KEGG, GO and DIP. In the future, it will cover more biological resources.

Acknowledgements

The authors are grateful to CHUNG Moon Soul Center for BioInformation and BioElectronics for supporting this work. This work was supported by the Korean Systems Biology Research Grant (M10309020000-03B5002-00000) from the Ministry of Science and Technology.

References

[1] Spellman et al. Comprehensive Identification of Cell Cycle-Regulated Genes of the Saccaromyces Cerevisiae. *Molecular Biology of the Cell*, Vol. 9, 3273-3297, December 1998.

[2] Nir Friedman et al. Using Bayesian Networks to Analyze Expression Data, 2000. *Journal of Computational Biology* 7:601-620.

[3] D. Endy and R.Brent. Modelling cellular behaviour. *Nature*, 409 Suppl(6818) : 391-5, 2001.

[4] Hasty, J., McMillen and et al. Computational studies of gene regulatory networks. 2001. *Nature Reviews of Genetics* 2:268-279.

[5] Bower, J. M. et al. *Computational Modeling of Genetic and Biochemical Networks.* 2001. MIT Press, Cambridge, MA.

[6] Isaac S. Hohane et al. *Microarrays for an integrative genomics*, MIT press,Cambridge, 2003.

[7] Smolen, P. et al. Modeling transcriptional control in gene networks: methods, recent results, and future directions. 2000. *B. Math. Biol.*, 62:247-292.

[8] Kauffman, S. A. The large-scale structure and dynamics of gene control circuits: an ensemble approach. 1974. *Journal of Theoretical Biology* 44:167-190.

[9] S.Liang, S.Fuhrman, and R. Somogyi. Reveal a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing.* vol. 3, p18-29, Hawaii, 1998.

[10] Mjolsness, E. Sharp, D.H., and Reinitz, J. A connectionist model of development. 1991. *Journal of Theoretical Biology* 152:429-453.

[11] Hlavacek, W.S., and Savageau, M. S. Completely uncoupled and perfectly coupled gene expression in repressible systems. 1997. *Journal of Molecular Biology* 266:538-558.

[12] Dana peer et al. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, vol.1, no 1, p1-9, 2001.

[13] Yoo et al. Discovery of Causal Relationships in a Gene-Regulation Pathway from a Mixture of Experimental and Observational DNA Microarray. *Data. Pacific Symposium on Biocomputing* **2002**

[14] Eran Segal et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, Vol.34, 2, June 2003.
[15] Jensen, F.V. An introduction to Bayesian Networks, *University College London Press*, 1996.
[16] http://www.genome.ad.jp
[17] http://db.yeastgenome.org
[18] http://dip.doe-mbi.ucla.edu
[19] W. Lam et al. Learning Bayesian Belief networks: An approach based on the MDL principle. *Comp. Int.*, 10:269-293, 1994.
[20] Learning Bayesian Network Structure from Massive Datasets : The "Sparse Candidate" Algorithm by Nir Friedman et al.
[21] Hastie et al. 'Gene Shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, Volumn 1, research0003.1-0003.21.
[22] Hastie et al. Gene Shaving : a new class of clustering methods for expression arrays. Technical report. *Stanford University*.
[23] Olga Troyanskaya et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, Vol. 17, no. 6, p.520-525, 2001.
[24] J.Pearl and T.S. Verma. A theory of inferred causation. Principles of Knowledge Representation and Reasoning: *Proc. Second International Conference(KR '91)*, p 441-452, 1991.
[25] D. Heckerman et al. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197-243, 1995.
[26] B. Efron and R.J.Tibshirani. An Introduction to the Boostrap. *Chapman and Hall*, London, 1993.
[27] http://mips.gsf.de/
[28] Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev* **16**:3017-3033
[29] Ito T et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc *Natl Acad Sci U S A*. 10;98(8):4569-74. Apr 2001.
[30] Seiya Imoto et al. Bayesian Network and Nonparametric Heteroscedastic Regression for Nonlinear Modeling of Genetic Network, *The first IEEE Computer Society's Bioinformatics conference*, USA, 2002
[31] Discretizing Continuous Attributes While Learning Bayesian Networks by Nir Friedman et al.