Inference of gene regulatory networks using Bayesian network

Daye Jeong
Department of Computer Science,
Yonsei University
Seoul, Republic of Korea
rabilish@cs.yonsei.ac.kr

Sanghyun Park
Department of Computer Science,
Yonsei University
Seoul, Republic of Korea
sanghyun@cs.yonsei.ac.kr

***Abstract*— It has been attempted to reveal regulatory information from microarray data using Bayesian network [1]. However, due to limitation of microarray, successful result is obtained only under a limited condition. For this reason, Bayesian network from combining microarray with biological knowledge was proposed [2]. In this paper, we proposed Bayesian network learned by genetic algorithm to infer gene regulatory network. We use protein-protein interaction, gene-gene interaction, and protein-DNA interaction data for construction of Bayesian network.**

Keywords—Bayesian network; Genetic algorithm; gene regulatory network; microarray; PPI; GI; Protein-DNA interaction

I. INTRODUCTION

Human body consists of many interacting systems. Each unit including cell, DNA, and chemical compounds interact each other and it makes people alive. Some of the known interactions between them, but many things are still not identified. Human DNA encodes genetic information for humans. They may inhibit, or activate itself, and regulate other genes, RNA, and proteins. This regulation information can be obtained by analyzing the mRNA which is an intermediate material of genes and proteins. mRNA expression value can be obtained through microarray or RNA-seq.

It has been attempted to analyze microarray data using Bayesian network to reveal genetic relationship [1]. However, microarray data is noisy and provide few samples. Application of Bayesian network is successful only under a limited condition. For this reason Bayesian network by combining expression data with biological knowledge was proposed [2].

A Bayesian network is useful for describing processes composed of locally interacting components and provides models of causal influence [1]. In the Bayesian network, the distribution of parent node affects the distribution of child node through conditional probability. Therefore, the distribution of node can be changed according its structure.

In order to learn the structure of Bayesian network, we use genetic algorithm (GA). GA is search algorithm based on genetics and the process of natural selection. Each generation, every individual calculates fitness. Individuals for the next generation are selected using selection algorithm like roulette-wheel algorithm. Each individual in the population evolves to getting a higher fitness value generation by generation. In this paper, we define a fitness function as how the structure of the network reflects microarray data and biological information which includes protein-protein interaction (PPI), gene-gene interaction (GI), and protein-DNA interaction.

In this paper, we have implemented Bayesian network using R that is free software for statistical computation and graphical visualization. In order to generate Bayesian network, we have utilized the graph package [7].

The remainder of this paper is organized as follows. In Section 2, we describe our approach to learn Bayesian network using GA. In section 3, we apply our approach to reconstruct the KEGG pathway. Finally, in section 4, we conclude by summarizing our work and discuss some of our future work.

II. METHOD

In this paper, we proposed methodology to infer gene regulatory network using a Bayesian network from microarray and biological knowledge. Before the network construction, the expression values of microarray are discretized to three expression state: −1(under-expressed), 0(base line), 1(over-expressed) [1]. In order to not overfit the data, we perform leave one out cross validation.

The process of GA begins with an initial population which is randomly generated. Each individual are represented by structure of network. Then, calculate fitness score and select individual using roulette-wheel algorithm. The higher the fitness score is, the more likely an individual will be selected. Selected individuals become next generation through mutation and cross over.

The fitness function of the GA for each individual is defined as follows (1):

$$P(G|D) \propto P(D|G)P(G) \quad (1)$$

Where P(G|D) is posterior probability of graph G given data D, P(D|G) is the likelihood, and P(G) is structure prior probability based on prior knowledge. In this paper, P(D|G) is

 ISSN 1975-4736 ©MITA 2014

calculated by comparing expression state of leaf nodes which is most likely state on local joint probability calculated from training set and test set. We calculate P(G) as average weight of edge from multiple biological knowledge. We use PPI, GI, and protein-DNA interaction. Each weight of knowledge is 0.25, 0.25, and 0.5. Protein-DNA interaction data is directional data unlike other prior knowledge, thus it should be given more weight. If there is an edge which has PPI, GI knowledge, then weight of edge is 0.5.

Due to GA which randomly generates edge between nodes, Bayesian network might have cycle. Bayesian network is DAG; therefore we remove cycle by performing depth first search.

III. EXPERIMENT & RESULT

To evaluate the performance of our approach, we consider KEGG pathway database [8] as "known" regulatory network. We assume that interactions represented in KEGG are true.

We reconstruct regulatory network using the nodes on map05210 in KEGG to prove effectiveness of our approach. Map05210 in KEGG is gene regulatory network about colorectal cancer. Therefore, we used GSE8671 with 32 colon adenoma samples and 32 normal samples [3]. The PPI, GI, Protein-DNA interaction dataset obtained from OPHID [4], Biogrid [5], HTRIdb [6].

GA parameters are listed below Table.1. We generate approximately 150 different Bayesian networks regardless of generation and population. Due to use almost un-directional data for construction of network, we consider generated network as un-directed network..

TABLE I. GA PARAMETER

Parameter	***Value***
Generation	100, 200, 300, 400, 500, 700, 1000
Population	20, 30, 40, 50, 70, 100
Mutation rate	0.5
Cross over rate	0.5

We calculate Pearson correlation coefficients (PCC) between fitness score and number of true edge to measure robustness of fitness function. In KEGG graph, there are indirect edge and direct edge. The overall average correlation coefficient that only counts direct edges is 0.75 which shows positive correlation regardless of generation and population. The overall average correlation coefficient that counts both edges is 0.71. It means that proposed fitness function is enough to find gene regulatory network only direct edge but also indirect edge. The reason correlation coefficients slightly lower when include both edges is that data for construction of network include only direct information between genes. We define p-value as the probability of obtaining k true edges and n-m false edges in n inferred edges when network are randomly generated using hypergeometric distribution.

$$p - value = \sum_{m=0}^{k} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \quad (2)$$

We consider indirect edge in inferred network which is represented direct edge in KEGG as true edge. The average p-value of inferred network is 0.015 and is far below the threshold α=0.05. This result supports that inferred network using Bayesian network is statistically significant.

IV. CONCLUDION & FUTURE WORK

Genetic interaction most reveal through biological experiment. In this paper, we propose a methodology to infer gene regulatory network using Bayesian network. We confirmed that microarray data with biological knowledge include PPI, GI, protein-DNA interaction can infer the structure of gene regulatory network.

Due to complexity of our approach, we limited node as nodes in KEGG. We can choose node through feature selection to reveal "novel" regulatory information. To infer regulatory information more accurately, we can use various prior knowledge.

REFERENCES

[1] Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." Journal of computational biology 7.3-4 (2000): 601-620

[2] Imoto, Seiya, et al. "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks." Journal of Bioinformatics and Computational Biology 2.01 (2004): 77-98

[3] Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E et al. Transcriptome profile of human colorectal adenomas. Mol Cancer Res 2007 Dec;5(12):1263-75. PMID: 18171984

[4] Brown, Kevin R., and Igor Jurisica. "Online predicted human interaction database." Bioinformatics 21.9 (2005): 2076-2082.

[5] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. Biogrid: A General Repository for Interaction Datasets. Nucleic Acids Res. Jan1; 34:D535-9

[6] Bovolenta, L. A.; Acencio, M. L.; Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genomics, v.13, p.405, 2012

[7] R. Gentleman, Elizabeth Whalen, W. Huber and S. Falcon (2006). graph: graph: A package to handle graph data structures. R package version 1.38.3.

[8] Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research 28.1 (2000): 27-30