GENAW : A Genetic Network Analysis System Using Bayesian Network Modeling

Hwan-Gue Cho[1*], Kyung-Shin Lee[1], Seon-Hee Park[2], Miyoung Shin[2]
[1]Department of Computer Science and Engineering, Pusan National University,
[2]Bioinformatics Research Team, ETRI

This paper introduces our GENAW system, which is a GEnetic Network Analysis Workbench using Bayesian Network Model from the gene expression data of raw microarray experiments. This system consists of three parts, the first preprocessing and the second network analysis and the final visualization modules. Especially this GENAW enables the biologist to manipulate the network construction procedure by controlling the set of target genes, the condition set, and the time interval between two different condition time slots. In order to handle the heavy computation to get the optimal result, we use the Genetic Algorithm heuristics. Also GENAW users can control the specific computation step to investigate the details in Genetic Network to be constructed. Finally we show some result(snap shots of GENAW) obtained from real data set(yeast cell cycles).

1. Introduction

It is a crucial work to study the gene expression patterns to understand the functions of a cell or some organism. We know DNA, RNA and set interactions among proteins are responsible to control the genetic regulatory system. This kind of genetic regulatory system consists of positive feedback loops and negative loops, which is of very complicated structure. And there is a limit to discover the whole genetic regulatory structure by wet experiments only[1]. In order to investigate the genetic regulatory structure, it is a first and basic step to study the variation of gene expression patterns (in the form of time-series data) from a raw data, which could be a set of microarray data set[5].

Since the current microarray technology is useful and efficient to study the pattern of gene expression in detail, there were previous studies on the construction methodology of genetic network[1-4].

2. Genetic Network Analysis System Overview

2.1 Preprocessing Step

In this paper we assume the basic data set for this GENAW is the microarray data set. Since microarray has generically lots of noise (in laser scan image) and missing value, the preprocessing step should clear these problems. Simple image preprocessing techniques can eliminate the image noise. For missing value handling, KNN(K-Nearest Neighbors), SVD(Singular Value Decomposition), and Row average technique were introduced. KNN is very simple to use and implement, but does not

give bio-statistically reliable result. SVD is regarded one of the recommended one, but it is hard to implement[2]. Also lots of data normalization adjusting is required to compensate the experiment condition or microarray print-tip, the condition of scanned image states[6]. Finding robust microarray normalization algorithm is getting one of hot issues.

2.2 General Network Construction Algorithm

We can use the exhaustive searching to discover activator, inhibitor genes. However the exhaustive searching method is a very expensive computation, since the time complexity increases exponentially, with the number of genes in a network. So we need to develop a heuristic approach to construct the genetic network in a reasonable time and resources. Clustering genes of similar expression pattern is also applied and local network merging method can be used. In this paper we exploit the Bayesian Network Model to construct genetic network.

2.3 Characteristics of Bayesian Network Model

Bayesian network model is one of directed acyclic graph model to denote joint probability distribution obtained from raw data. If two variables are not connected in this graph model, that means each stochastic variable is independent [4].

Since Bayesian Network model is very flexible and reliable to show the causal influence relationships (dependence and independence) among several stochastic variables, this model is widely accepted in bioinformatics. Especially it is believed that Bayesian model with probabilistic learning theory could give the most valid and reliable result in Genetic Network design and prediction.

2.4 Issues in Network Evaluation

It is very important to evaluate the significance of the genetic network obtained by artificial computation models. Only wet experiment is the best way to see the result is correct or not, or the degree of correctness. Another simple alternative is to consult the Bio database by blind testing approach. However up to now there is few database which maintains the genetic network data, so the latter database consulting method is not practical and also does not guarantee the correctness[3,4]. Next we will introduce our GENAW system briefly.

3. Overview of GENAW System

GENAW is java-based workbench to construct genetic network from the microarray data set. As was assumed in other systems[3], we(GENAW) also give the constraints on the number of interacting regulator genes(the degree of a node) [3]. If we restrict the number of neighbor regulators(less than 3), we check all possible cases(exhaustive searching). Otherwise, if we allow the number of interacting genes is more than 3, then we apply GA(Genetic Algorithm) heuristics to get a sub-optimal result in a reasonable time(less than 5 hours in Pentium PC)

3.1 System Architecture

As was shown in Figure-1 GENAW consists of three parts: preprocessing, network construction, visualization.

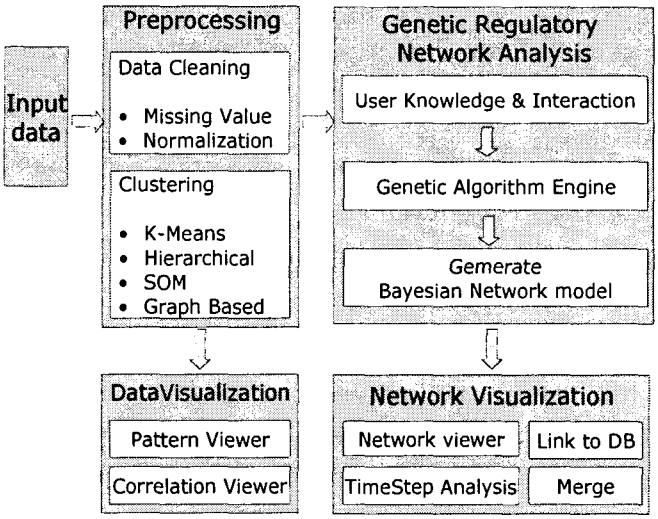

Figure 1. Software Architecture of GENAW

Input data value is R(ed)/G(reen) signal ratio of microarray spot image. Preprocessing step compensate the missing data and noise data entry and will normalize all data entry to [0,1] real number. Additionally user can select his/her own data point(we not select all experiment points) in a time-series data. After processing these steps(missing data handling, normalization, user controls), these preprocessed data set enters into GA(Genetic Algorithm) engine where the most probable genetic network model is constructed. And finally visualization module shows the concrete structure of the generated network and also it allows user interaction, thus we can move some nodes and edges to get a better network layout.

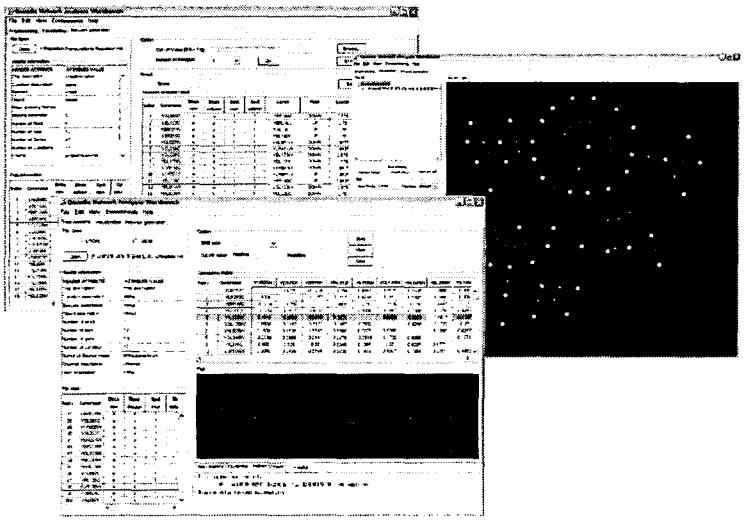

Figure 2. Snapshot of GENAW (top) genetic network data sheet(after preprocessing) (middle right) visualization of predicted network structure (bottom) correlation analysis of input microarray genes and visualization

3.2 Characteristics of GEANW

GENAW gives some useful functions for user to analyze the microarray data. Characteristic

functions of GENAW are as follows.

(a) User can control the data set and interesting parts selectively. User can add some "solid" interaction relation to get a realistic result. For example if we already know(or it is widely accepted in bio-informatics world) that gene Ga activates Gb, then we can initialize this relationship regardless of the experiment data to be analyzed. Adding this kind of *a priori* knowledge of gene regulation can affect the whole structure. This is a highly desirable feature.

(b) GENAW allows for user to merge different results obtained from same input data. It is common there are different result even if the system gets the same data set , since it depends on analysis methodology and the preference of user(biologist)

(c) Since the each gene works in different time step, it is very useful to watch the gene interaction in each time steps separately. GENAW allows user can select the time steps he wishes to analysis. Thus the working mechanism can be simulated along the time steps.

(d) GENAW has a set of very informative visualization tools and can be integrated to LIMS(Laboratory Information Management System). This enables more fast and systematic analysis and microarray data management.

4. Experiment and Conclusion

4.1 Data1: Yeast data set[7]

We applied S. Cerevisiae data of transcription regulatory network. Among these genes we selected 47 genes, which were experimentally approved on gene interaction.

According to the previous wet experiment, we know that RPS24B, RPL11A, RPS2 was controlled by FHL1. GENAW clearly and correctly discovered that the relationship among {RPS24B, RPL11A, RPS2}, {PRT2, CUP9}, {GCN4, MET4, GAT1}. Also we found that {CDC20, SWI5} and {GAL80, MTH1} functions in a similar way. In Figure-3, red(+) arrow edge means the activation, blue(-) edge means inhibition relation.

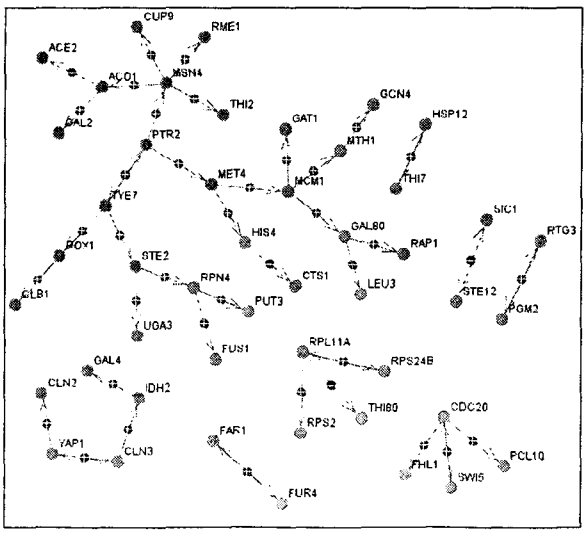

Figure 3. Gene regulation structure among the selected 47 yeast genes obtained by GENAW.

4.2 Time-series data for foreign currency

We can compare the foreign currency with the gene regulation in the whole world economy scale. We made a currency variation data on USA dollar, UK pound, Japan Yen over fiscal year of 2002, 2003(month average). Our final result correctly showed that USA dollar affects the currency rate of East Asia(activates) and Middle East highly. And US $ inhibits the European EURO currency. Interestingly SWISS Franc behaves oppositely comparing to USA dollars.

We conclude this paper. GENAW is very efficient, easy to use, reliable genetic construction system. One interacting feature of GENAW is that it allows user interaction(selecting time-step, adding a priori knowledge et al.). Also this is integrated to a LIMS(GELIM), user can easily manage and prepare dataset. In future, we have a plan to integrate GENAW with a protein-protein interaction analysis tool to help experiment biologist. Detail information and availability can be obtained by visiting the following Web Home site. Also constructive comments and advice are welcomed.

http://garnet.cs.pusan.ac.kr/~genaw

Acknowledgement

This work was supported Research Contract(2004) of ETRI Bioinformatics Research Team.

5. References

1. Hidde De Jong, Modeling and Simulation of Genetic Regulatory Systems: A Literature Review, *Journal of Computational Biology*, 9(1), pp.67-103, 2002.
2. Troyanskaya O. et al. Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17(6), pp.520-525, 2001
3. Hitoshi Iba et al. Inference of a gene regulatory network by means of interactive evolutionary computing, *Information Sciences*, 145, pp.225-236, 2002.
4. Nir Friedman et al. Using Bayesian Networks to Analyze Expression Data, *J. of Computational Biology*, 7(3/4), pp.601-620, 2000.
5. Patrik D'haeseleer et al. Gene expression Data Analysis and Modeling, *PSB* tutorial, 1999
6. Yee Hwa Yang et al. Normalization for cDNA microarray data: a robust composite method addressing and multiple slide systematic variation, *Nucleic Acids Research*, 30(4), 2002,
7. Tong Ihn Lee et al, Transcriptional Regulatory Networks in Saccharomyces cerevisiae, *Science*, pp.799-804, 2002.
8. Kyung-shin Lee, Gene Regulatory Network construction using Genetic Algorithm, Technical Report, ALGORIGENE Lab, PNU, 2003
9. Kyung-shin Lee, Genetic Network Construction with Continuous value data, Technical Report, ALGORIGEN Lab, PNU, 2004