CONSTRUCTING GENE REGULATORY NETWORK USING FREQUENT GENE EXPRESSION PATTERN MINING AND CHAIN RULES

Hong Kyu Park, Heon Gyu Lee, Kyung Hwan Cho and Keun Ho Ryu
Database and Bioinformatics Laboratory, Chungbuk National University, South Korea
{hkpark1980, hglee, khcho, khryu}@dblab.chungbuk.ac.kr

ABSTRACT: Group of genes controls the functioning of a cell by complex interactions. These interacting gene groups are called Gene Regulatory Networks (GRNs). Two previous data mining approaches, clustering and classification have been used to analyze gene expression data. While these mining tools are useful for determining membership of genes by homology, they don't identify the regulatory relationships among genes found in the same class of molecular actions. Furthermore, we need to understand the mechanism of how genes relate and how they regulate one another. In order to detect regulatory relationships among genes from time-series Microarray data, we propose a novel approach using frequent pattern mining and chain rule. In this approach, we propose a method for transforming gene expression data to make suitable for frequent pattern mining, and detect gene expression patterns applying FP-growth algorithm. And then, we construct gene regulatory network from frequent gene patterns using chain rule. Finally, we validated our proposed method by showing that our experimental results are consistent with published results.

Keyword: gene regulatory network, frequent pattern mining, chain rule, gene interaction.

1. INTRODUCTION

Recently, with the usage of Microarray method, techniques that can reason gene regulatory network from the revelation data of gene was used discovered in order to show the function and the process.

The function of a cell is regulated through complex interactions of gene groups and these interacting genes are called Gene Regulatory Networks, also known as GRN. Interactions of genes can be classified into two groups: Co-regulation and Control regulation (regulator). First, Co-regulation, as it is shown in Figure 1(a), means that revelation of both Gene A and B is being regulated by another Gene which is Gene X. Second, Control regulation, as it is shown in Figure 1(b), means that Gene B is regulated by Gene A from transcription.

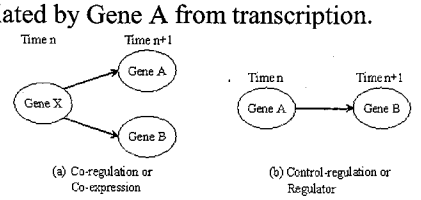

Figure 1. expression form of gene

Past Microarray analysis mechanism, such as Clustering and Classifying method [6, 7] merely predicted genes grouping by their similarity in revelation pattern and functions of new genes from a model which is instructed through genes that functions are known. With this method, it is impossible for us to understand how these genes relate to each other and how they interact. Consequently, frequent patterns of genes can be discovered through Microarray data and with these patterns we are able to explore regulatory relationships among genes from a chain rule, which is a statistical method that adapts series of conditional possibilities. In addition, proposed techniques can reveal regulatory relationship among genes that are not known and eventually, we can receive biological information that are known yet.

1.1 Related Work

There are two types of existing research of constructing Gene Interaction Regulatory Networks. First, it is time-series approach method. This method aims at modelling patterns of gene revelation on a certain point of time. Furthermore, that modelling is made by applying functions of past gene revelation patterns. However, this experiment has an obstacle that regards to the characteristics of Microarray data. So to speak, the small amount of time-point and a large number of genes are such problems and they lead to complex question that is hard to calculate. Therefore, do so that may except genes that did not show similar difference in expression in [2] and proposed Linear modeling predicative to shorten dimension. Enforcing SVD (singular value decomposition) in [3], leave gene of few number and solve interaction array and find easily interaction of gene. Second, it is Bayesian network that use machine learning [1, 3, 4]. Among them, proposed approach statistical in [1]. Method to reason in Boolean network form in [5] proposed. But, it is predictable to find the rate of high false positive among networks that are found.

1.2 Organization of the paper

The structure of this paper is composed of as following, aimed at effective comprehension. It describes discrete method of gene expression data by preprocessor phase for frequent pattern mining application in chapter 2 and chapter 3 and explain frequent pattern find processes of gene that apply FP-growth techniques. It also describes chain rule application algorithm and adjustment network construction method from gene expression patterns that is

created in chapter 4. The proposed method for constructing gene regulatory network is explored applied by Yeast data. Then the result of that experiment is explained in Chapter 5. Finally, chapter 6 states the conclusion of this paper.

2. PREPROCESSING FOR MICROARRAY GENE EXPRESSION DATA

This section, discrete method for serial gene expression data in gene expression Microarray data will be described. In addition, the process for transaction database of gene expression array will be discussed in order to make sequential pattern mining possible. Usually, value about expression amount of gene is an actual number and consist of "n" genes (or probe), "m" experiment samples in case of serial gene data.
The Sample of microarray data is same with Figure 2, and ORF (Open Reading Frame) that mean gene regards as item, each time point as transaction.

ORF	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7
YHR007C	1.12	1.19	1.32	0.88	0.84	0.38	0.43
YBR218C	1.18	1.23	0.77	0.75	0.79	0.71	1.7
YAL051W	0.97	1.32	1.33	1.18	1.12	0.88	0.93
...	...	...	...	...	...	...	...
YAL055W	0.68	1	0.92	0.96	0.81	1.28	1.85

Figure 2. Time-series Microarray data sets

Convert Figure 2 gene expression data to binary if do discretization. It is possible that make binary array by this value, and extract expression pattern of gene such as frequent pattern. Data conversion does discretization to 3 sections about ratio of expression value relative gene. Gene (item) according to ratio of expression 3 value expresses by *gene-up, gene-down, unchange.* If is expressed to item about gene ratio of expression, change to binary array and appear by transaction. Specification time-point if ration of expression value of gene is bigger than 1 in transaction that have appeared do value of item by gene-up = 1 and establish gene-down and *unchange* by 0. If the ratio of expression is value between 1 and 0, gene-down by 1, change *gene-up* and *unchange* to 0. Also, if gene is same with ratio of 1 *unchange* by 1 remainder *gene-up* and *gene-down* by 0 establishments. If it is expressed to item about gene ratio of expression, change to binary array and appear by transaction.

3. FREQUENT PATTERN MINING FOR DISCOVERY OF GENE INTERACTION

In this chapter, introduction of gene pattern to find a way how frequently in gene expression microarray data to find genes that have similar expression pattern at certain visual point, and propose CP that is new measured value for usefulness measurement between made found patterns. Also, propose remove redundant pattern and CP-tree structure for efficient pattern save of memory in patterns of created large quantity [8].

First, To define for gene pattern for frequent pattern abstraction in gene expression data and define frequent pattern finding process problem step by step hereafter.

Definition 1 Gene pattern: Assume that gene be one of the items. If itemset that display pattern of gene is known as $p = \{i_1, i_2, ..., i_n\}$, $1 \le j \le n$, i_j then it is one gene.

Definition 2 sub pattern (or super pattern): Pattern, $p = \{i_1, i_2, ..., i_n\}$ is a sub pattern of another pattern $p' = \{i'_1, i'_2, ..., i'_n\}$ if there exist integers $k_1 < k_2 ... < k_n$ satisfy $i_1 = i'_{k_1}$, $i_2 = i'_{k_2}$, $i_n = i'_{k_n}$
For example, $p = \{p_2, p_4, p_5, p_6\}$ is a sub pattern of $p' = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$, since there exist integer $k_1 = 2 < k_1 = 4 < k_1 = 5 < k_1 = 6$ that are $(p_2, i_1 = i'_2)$, $(p_4, i_2 = i'_4)$, $(p_5, i_3 = i'_5)$, $(p_6, i_4 = i'_6)$ are existence

Definition 3 FPs (Frequent Patterns): Frequent pattern, It is sub pattern of each transaction that satisfies Minimum support (Min_{sup}) that critical value.

Consequently, exploring frequent pattern from gene revelation data is similar to exploring all sets of frequent genes that satisfy users' designated minimum support in advance from gene expression data.

FP-tree is a prefix tree structure that store support about frequent pattern and parent nodes compose tree by way that items that have high support worth are situated and are situated in child node is low support's node. All patterns that FP-tree has header table structure to hold count value of items of each node, and is inserted to tree include count value of items.

3.1 Moving redundant pattern and CP-tree algorithm for pattern storage

After all frequent patterns are created by algorithm, it defines PC (pattern Cohesion) that is new measure beside support for usefulness measurement of each pattern, this is used redundant pattern at CP-tree creation and delete of unnecessary pattern [9].

Definition 4 PC (Pattern Cohesion): For a pattern $p = \{p_1, ..., p_n\}$ of length n, PC is a ranking measure defined as

$$PC(p_1, ..., p_n) = \frac{Cnt(p_1, ..., p_n)}{\sqrt[n]{Cnt(p_1) \times ... \times Cnt(p_n)}} \quad (1)$$

(1) extends (2) that is interrelationship measured values between two words in text classification to problem of item set that have length of n.

$$Cohesion(w_i, w_j) = \frac{P(w_i, w_j)}{\sqrt{P(w_i) \times P(w_j)}} \quad (2)$$

All created patterns being pattern list of bulk, include much redundant patterns. Therefore, for redundancy pattern remove that do bases storage and PC of efficient

patterns, propose CP-tree (Compressed Pattern tree) data store structure that change tree that is introduced in [14].

The CP-tree structure is available compressed pattern storage and reflects sub sequence pattern/super sequence pattern relations between patterns.

CP-tree structure is similar, but has next two other characteristic with FP-tree structure.

- About all patterns, CP-tree includes PC that is cohesion of each pattern as well as support.
- Only last leaf node has attribute information of all patterns, and single pattern is passing to route from leaf.

When all patterns are inserted to CP-tree, at the same time, delete of redundancy patterns happens.

4. GENE INTERACTION NETWORK CONSTRUCTION USING CHAIN RULE

Express network construction using set FP of all patterns and chain rule. This time, selected patterns become sub patterns of gene set that wish to construct. Composition of gene network that use frequent patterns becomes an issue by probable network model of chain rule and these chapter describe problem of approximation product that apply chain rule from FP pattern set with justice of probability connected with this.

Definition 6 conditional probability, joint probability: The conditional probabilities of an event X in relationship to an event Y is the probability that event X occurs given that event Y has already occurred. The notation for conditional probability is P(X|Y). The formula for conditional probability is (3):

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} \quad (3)$$

(3) can be express $P(X)P(Y \mid X) = P(X \cap Y)$ forms, and $P(Y)P(X \mid Y) = P(X \cap Y)$ forms. It is known that probability that X and Y happen at the same time is joint probability and it is known that this is multiplication rule to be derived to conditional probability's mathematical expression.

Definition 7 chain rule: Probability that each event a, b, c generate can express, and refers to this that is chain rule in expression of conditional probability product that is serial with (4) using conditional probability and joint probability.

$$P(A_1, A_2, \ldots, A_n) = P(A_1 \mid A_2, A_3, \ldots, A_n) \bullet P(A_2 \mid A_3, A_4, \ldots, A_n) \bullet \ldots \bullet P(A_{n-1} \mid A_n) \bullet P(A_n) \quad (4)$$

Network construction can presume by probabilistic model from gene data by [Definition 6 - 7]. First, if set $G = \{g_1, g_2, \ldots, g_n\}$ of set genes for network architecture is given, is expressed because applying chain rule as product of gene patterns that it does probability $P(G) = P(g_1, g_2, \ldots, g_n)$ by maximum.

Definition 8 product approximation: Probability $P(g_1, g_2, \ldots, g_n)$ of genes can be assumed by approximations that differ using chain rule of Definition 7. Each approximations express different condition independence family about attributes. $P(g_1, g_2, g_3, g_4)$ is calculated, and refers to this time two probability that is approximation product of by product $P(g_1, g_2) \cdot P(g_3, g_4 \mid g_1)$ or $P(g_1, g_2) \cdot P(g_4 \mid g_2) \cdot P(g_3 \mid g_1, g_4)$ of probability.

4.1 Composition algorithm of approximation product using chain rule

If set of genes for gene adjustment interaction network architecture is given, all frequent pattern lists that is included on the set are formed, and can choose fittings that do the gene network model's probability by maximum as approximation product of Definition 8.

Definition 9 Border, B: B exists inside frequent pattern set by set of genes for adjustment network architecture, $G = \{g_1, g_2, \ldots, g_n\}$'s sub patterns, and the relation is define as (5).

$$B = \{p \in FP \mid p \subset G\} \quad (5)$$

Definition 10 Pattern p selection rule for composition of approximation product:

rule 1: $|p - \mathrm{cov}| \geq 1$ rule 2: $PC(p) > PC(p')$

rule 3: $length(p) < length(p')$ rule 4: $|p - \mathrm{cov}| \leq |p' - \mathrm{cov}|$

In above rule 2 and rule 3 pattern p' instead p select.

Rule 1 means that include new item (gene) more than one that pattern p that is selected is not included in selected pattern necessarily before here and this guarantees effectiveness of chain rule and approximation product. Rule 2 does to select that length of pattern is short if is meaning that put priority p that have high cohesion, and pattern to have same cohesion. Rule 3 does amount of pattern that is used in composition of approximation product so that do maximize. Finally, rule 4 means that put priority p that number of unlisted item is smallest already of remaining patterns.

4.2 Gene adjustment interaction network construction from chain rule

If border list for approximation product is decided, we can discriminate regulators from list's all patterns. If gene set g are given, and appear after composition algorithm of approximation product and p selection rule procedure algorithm application with $G = \{g_1, g_2, g_6, g_9, g_{11}\}$, can assume as following (6).

$$P(G) = P(g_1, g_{11}) \cdot P(g_2 \mid \underline{g_{11}}) \cdot P(g_6 \mid \underline{g_2}) \cdot P(g_9 \mid \underline{g_6, g_{11}}) \quad (6)$$

- Regulators are g_{11}, g_2, g_6.

$$P(g_2 \mid g_{11}) = P(\frac{g_2 \wedge g_{11}}{g_{11}}) = P(g_{11} \rightarrow g_2) \quad (7)$$

Expression, $P(g_{11} \rightarrow g_2)$ means that g_2 does expression under state that gene g11 expression. We can yield

conditional probability for discriminate genes by whole regulator with Figure 3 by these methods and can construct adjustment network.

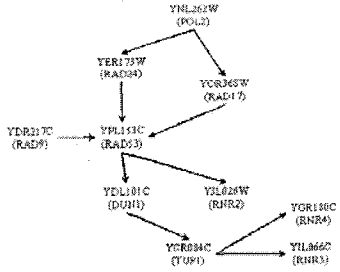

Figure 3. Adjustment network of genes connected with DNA damage repair reaction

5. EXPERIMENTS AND RESULTS

An experiment about gene control interaction network construction used control relation (activation, inhibition) in an experiment as genes that searched saccharomyces cerevisiae [6] and Yeast Protein Database [10].

Table 1. Control relation number at two growth steps that is created from YPD database and saccharomyces cerevisiae data

Data set	# of genes	# of activation	#of ingibition
Alpha-factor	332	343	96
Cdc28	365	469	155

Alpha-factor data set is included in activation relation, 96 inhibition relation that only 332 genes became match about 18time-point all and 343 function of this genes is informed. 365 gene is been matched all to cdc28 data and 121 genes excepted because is inconsistent. Cdc28 includes been matched 469,155 adjustment relation every moment. Discretization about expression ratio creates two dataset about positive and negative transformation. In the case of positive data conversion, do gene-up by 1 (gene-down = 0, unchanged = 0) and negative's case did gene-down by 1 (gene-up = 0, unchanged = 0). Genes that is found by regulator by chain rule discriminated by genes that have informed already biologic function. An experiment alpha-factor and cdc28 data set expression find out done all frequent gene expression patterns and apply chain rule and forecast regulators. Verification about prediction expresses by table7's confusion matrix, and estimation with prediction used Recall and Precision and F-Measure and MAE. Result about table5's alpha-factor and activator/inhibitor prediction of cdc28 data set is table8.

6. CONCLUSION

In this paper, we have predicted regulator that regulate level of gene expression using constructing Gene Regulatory Network. For that, we changed gene expression data to three type data by expression ratio and convert into transaction for applying frequent pattern mining. First step, the proposed constructing gene regulatory network process is, Find the frequent patterns of genes from each gene expression data by preprocessing. For applying the high efficiency of FP-growth, we introduced a novel algorithm by using PC measure and CP-tree for redundant pattern pruning. Last step, to predict a network by probabilistic model using chain rule from removing redundant patterns. Our experimental used to reflect well positive and negative more than binary conversion (up, down) about data sets alpha-factor and cdc28 among growth cycle of saccharomyces cerevisiae. Also, the experiment results verified through comparison with result that it is informed already with regulator who is predicted about each dataset.

6.1 References and/or Selected Bibliography

References from Journals:

[1] Friedman, N., Linial, M., Nachman, I. and Pe'er, D., *Using Bayesian networks to analyze expression data*, Journal of Computational Biology, 7:601-620, 2000.

References from Other Literature:

[2] Van Someren, E. P., Wessels, L. F. A., and Reinders, *Linear modeling of genetic networks from experimental data*, Proc., ISMB, 355-366, 2000.

[3] Holter, N. S., Maritan, A., Fedoroff, N. V. and Banavar, J. R., *Dynamic modeling of gene expression data*, Proc., Natl, Acad. Sci. 1693-1698, 2000.

[4] Rishi Khan, Yujing Zeng, Javier Garcia-Frias and Guang Gao, *A Bayesian Modeling Framework for Genetic Regulation*, Proc., CSB'02, 2002.

[5] Akutsu, T., Miyano, S., and kuhara, S., *Identification of genetic networks from a small number of gene expression patterns under the Boolean network model*, Pacific Symposium on Biocomputing 17-28, 1999.

[6] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., *Cluster Analysis and Display of Genome-Wide Expression Patterns*. Proc., National Academy of Science. 95:14863-14868, 1998.

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

[8] Han, J., Pei, J., Yin, Y., *Mining frequent patterns without candidate generation*. In SIGMOD'00, Dallas, TX, 2000.

[9] Lee, H. G., Noh, K. Y., Lee, B. J., Ryu, K. H.: Cardiovascular disease diagnosis method by emerging patterns. Lecture Notes in Volume 4093. Springer-Verlag, Berlin Heidelberg New York (2006) 819–826

[10] Yeast Protein Database (http://www.proteome.com)

6.2 Acknowledgements

This work was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.