Bioinformatics and Biosystems 2007, Vol. 2, No. 1, pp. 24-27

Inferring candidate regulatory networks in human breast cancer cells

Juhyun Jung, Doheon Lee

Dept. of BioSystems, KAIST

Abstract

Human cell regulatory mechanism is one of suspicious problems among biologists. Here we tried to uncover the human breast cancer cell regulatory mechanism from gene expression data (Marc J. Van de vijver, et.al.,2002) using a module network algorithm which is suggested by Segal, et. al.(2003) Finally, we derived a module network which consists of 50 modules and 10 tree depths. Moreover, to validate this candidate network, we applied a GO enrichment test and known transcription factor-target relationships from Transfac® (V. Matys, et.al, 2006) and HPRD database (Peri, S. et al. ,2003).

Keywords: Gene regulatory Network, Breast Cancer, Module network

서 론

Gene expression profiling has been applied in cancer research and is widely believed to reveal molecular mechanism underlying cellular functions. Computational analysis functions as a crucial bridge to suggest hypothesis from this gene expression profiling. For example, hierarchical clustering offered us opportunity to foresee future of cancer. (Sorlie T, et.al., 2001, Marc J. Van de vijver, et.al.2002) However, simply listing genes associated breast cancer metastasis is far from identifying the biological process in which these genes are involved. Therefore, it is a key challenge to develop an analysis method that can extract more biologically meaningful understanding of the processes giving rise to cancer. So far, regulatory mechanism has been studied based on yeast cells, not on human cells. (Lee TI, et.al, 2002, Segal, et. al.2003)

Here, we inferred a candidate module network in human breast cancer cell. Breast cancer alone is expected to account for 32% (211,240) of all new cancer cases among women in USA.(Cancer statistics,2005) The incidence rates of breast cancer have continued to increase in both Korea and USA. Study of Breast cancer is complicated, because breast tumors consist of many different cell types. Also breast cancer cells themselves are morphologically and genetically diverse.

Corresponding Author:
Doheon Lee(E-mail: doheon@kaist.ac.kr).
This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the National Research Lab. Program (No. 2005-01450), and the Korean Systems Biology Research Grant (2005-00343).

Several studies showed the relationships between gene expression patterns and breast cancer outcome. Unlike these previous studies, we wanted to find not only list of gene predictors but also regulatory program to make their prognosis different using an algorithm suggested by Segal, et. al. (2003)

Method

1. Microarray dataset

We utilized breast cancer cell cDNA microarray data set which consists of 295 samples from individual patients. (Marc J. Van de vijver, et.al.,2002) The previous study which utilized this data found 70 powerful predictors of the outcome of breast cancer. The patients had primary invasive carcinoma that was less than 5 cm in diameter. Their lymp nodes were tumor-negative, as determined by a biopsy. Also they were younger than 55 years. cDNA microarray probes were 9642 genes which have tolerant missing values and applicable derivatives between samples among genome scale. We normalized data by median.

2. Learning network

2.1 Genomica

We downloaded Genomica® from Segal's web site (http://genomica.weizmann.ac.il/). This software provides us a framework to make module networks. A module network is a probabilistic model for identifying regulatory modules from gene expression data. It is successfully applied on Segal's pa-

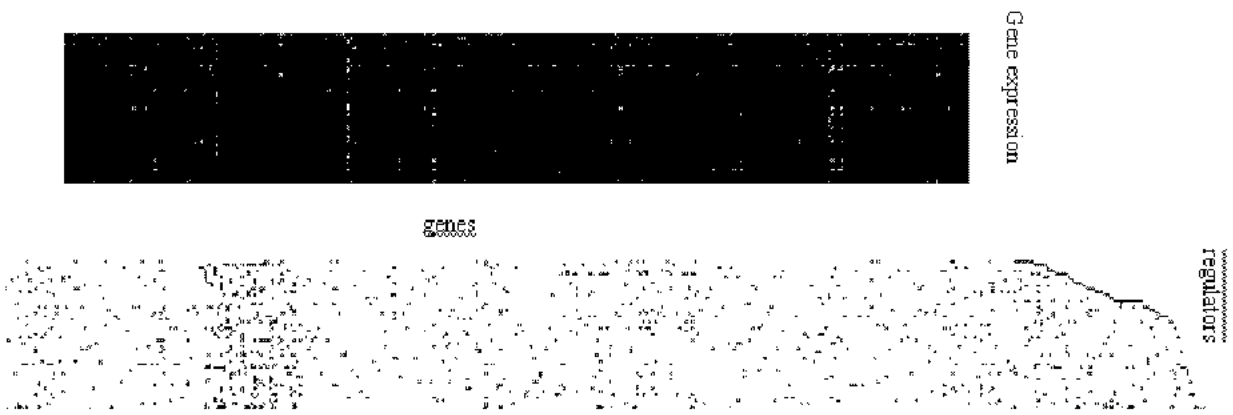

Figure 1. a global view of module network

per(2003) which constructed a regulatory network underlying the response of yeast to stress.

We converted microarray data explained in the previous section to 'gxp' format to employ this program.

2.2 Regulation programs

A module network is defined by a set of contexts and the response of the module in each context. A context is a qualitative behavior of a small set of regulators that control the expression of the genes in the module. This set of rules is organized as a regression tree. A regression tree is composed of two building blocks: decision nodes and leaf nodes. Each decision nodes corresponds regulatory inputs and a query on its value. (e.g: "is E2F2 up-regulated?") Each decision node has two child nodes: the right child node is chosen when the query is true, otherwise the left child node is elected.

2.3 Setting parameters

We chose likelihood scoring scheme, called Bayesian score. Maximum tree depth was 10. Number of module was 50, same as Segal, et.al. (2003)

2.4 Candidate regulators

We compiled a set of 425 transcriptional factors and 548 signal transducers from "gene ontolgy" site. (http://www.godatabase.org). Regulators are annotated by a literature reference, another database or a computational analysis.

2.5 Learning a module network

A module network is learned by Expectation Maximization (EM) algorithm to maximize Bayesian score. After 30 iterations of EM steps we achieved got a module network which consists of 50 modules.

Result and validation

1. A global view of a module network

A global view of our module network is shown in fig.1.

2. Gene ontology enrichment scoring

To systematically validate this result, we utilized gene ontology enrichment score first. We manually curated clusters which are divided by 50 modules. Gene ontology enrichment is statistically tested using hypergeometric score. GO enriched clusters which have the lowest level of p-value are listed.

Table 1-3 showed very low p-value which means this result is not generated by randomly. Cluster 1424 has 49 genes related to cell cycle and 26 genes related to cell division, and has significantly low p-value. It means that module related to cluster 1424 has a function related to cell cycle and division, linked to cell growth. Also cluster 1424 has a nucleotide acid binding which gave us some rationale to suspect 1424 has regulatory function in cells. However, cluster 1569 is related to metabolism, and also related to nucleic acid binding.

To analysis more dedicatedly we collected a cluster which has the lowest averaged p-value to satisfy significances of biological process, cellular component and molecular function.

Table 1. GO enrichment clusters (biological process)

Function	Category genes	Total genes	Go depth	P-value	cluster
cell cycle	49	164	4	2.85E-31	1424
cell division	26	164	4	1.02E-29	1424
metabolism	217	408	3	2.25E-23	1984
response to stimulus	63	184	3	1.62E-21	1451
response to stimulus	77	260	3	1.87E-19	2200
organismal physiological process	57	184	3	1.33E-18	1451
metabolism	246	494	3	5.19E-17	1375
metabolism	158	323	3	2.93E-16	1569
metabolism	183	409	3	5.51E-16	2238
response to biotic stimulus	89	260	4	9.67E-16	2200

Table 2. GO enrichment clusters (Cellular component)

function	Category genes	Total genes	Go depth	P-value	cluster
nucleic acid binding	101	408	3	4.41E-14	1984
nucleotide binding	32	164	3	1.27E-10	1424
nucleotide binding	67	494	3	7.92E-10	1375
hydrolase activity, acting on ac	37	494	4	9.33E-10	1375
protein binding	58	184	3	1.25E-09	1451
protein binding	72	260	3	1.92E-09	2200
nucleic acid binding	54	203	3	3.20E-09	1815
nucleic acid binding	88	323	3	8.82E-09	1569
cytokine binding	11	260	4	3.59E-07	2200
GDP-dissociation inhibitor acti	4	494	5	6.16E-07	1375

Table 3. GO enrichment clusters (Molecular function)

function				p-value	cluster
intracellular	272	494	3	1.77E-39	1375
intracellular	217	408	3	3.03E-35	1984
intracellular	169	323	3	2.50E-19	1569
intracellular	190	409	3	7.05E-19	2238
intracellular	144	306	3	2.92E-14	2394
intracellular	97	203	3	3.83E-14	1815
intracellular	85	164	3	2.18E-13	1424
inner membrane	20	494	4	7.89E-13	1375
intracellular	129	311	3	8.71E-13	2328
plasma membrane	67	384	4	1.74E-12	2305

3. Transcription factor-target relationships and protein-protein interaction

To validate our results quantitatively, the most go-enriched cluster's regulatory program are drawn in fig.2.

Highlighted interactions mean that they are validated by HPRD or TRANSFAC. HPRD has protein-protein interaction information and TRANSFAC has transcription factor-target

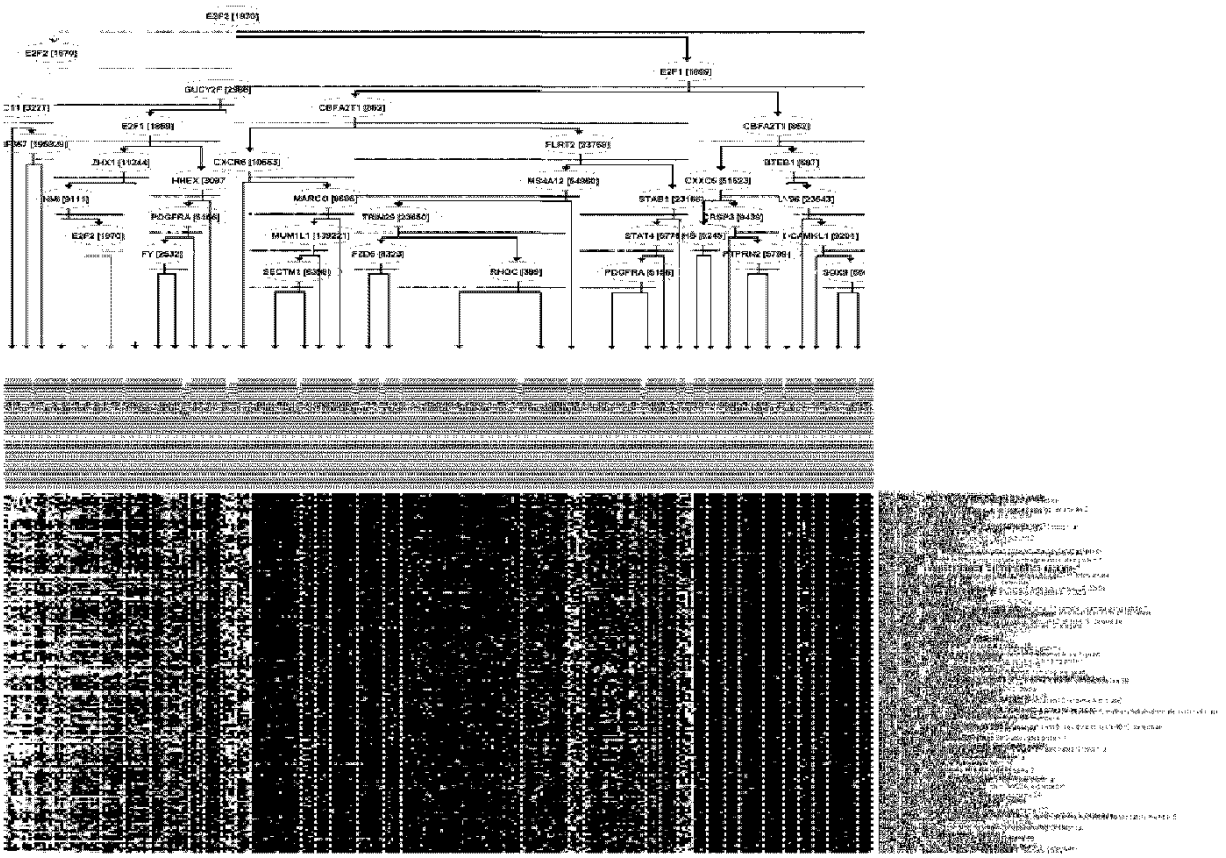

Figure 2. Genes in cluster 1424's regulatory program and their expression level

relationships. Because many of protein-protein interactions and transcription-target relationships are not annotated, it was hard to prove all of data is validated or not.

Discussion

Here, we represented a candidate regulatory network in human breast cancer cell. Most of modules that we retrieved have significant p-values and also they gave us regulatory program. From this result, we can say that regulatory program here is a plausible scenario in human breast cancer cell. But the new and crucial problem is which scenario is more plausible? That is why we suggested validation scheme to approach that issue. We calculated GO enrichment score which is based on statistical test, named hypergeomatric test. Also we validated the GO enriched cluster using TRANSFAC and HPRD database.

The study of human breast cancer cell is not that simple due to complexity of tumor itself. However, doing something is better than doing nothing. If we just stay in yeast cell, we will suffer lots of inherent problems, such as misfolding of protein, lack of human genes, and so on. So far, the advantage of study in yeast cell is only genome-scale availability of libraries of mutant strands. However siRNA opens the new window of genomics that allows to study of human gene function in vivo by knocking down genes. We think it is time to understand networks in human cells not only in yeast cells. Also we hope that our study accelerates this flow.

References

[1] Cancer statistics, (2005)

[2] Friedman,N. et al. (2000) Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620.

[3] Gao F, Foat BC, Bussemaker HJ, (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*.

[4] Lee TI, et.al (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae, *Science*

[5] Marc J. Van de vijver, et.al., (2002) A gene expression signature as a predictor of survival in breast cancer. *N Engl J Med*.

[6] Peri, S. et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*. 13:2363-2371.

[7] Segal, et. al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*.

[8] Sorlie T, et.al., (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc Natl Acad Sci*.

[9] V. Matys, et.al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*