

Mining estrogen microarray data: An approach using contrast data analysis

Haili Jiao*, Peixin Yang+, Zhengxin Chen**

*Department of Computer Science, University of Nebraska at Omaha

+Department of OB/GYN, University of Nebraska Medical Center

**Department of Computer Science, University of Nebraska at Omaha

hjiao@mail.unomaha.edu, pyang@unmc.edu, zchen@mail.unomaha.edu

1 Introduction

Estrogen is a steroid hormone, secreted by ovarian antral follicles, which regulates divergent functions of ovarian granulosa cells ranging from differentiation, apoptosis and proliferation. Cellular signals of estrogen are transduced by its cohort receptors. Estrogen receptors belong to nuclear receptor superfamily and comprise two subtypes, namely, estrogen receptor alpha (ER- α) and estrogen receptor beta (ER- β). Estrogens bound to ER- α or ER- β exert different cellular responses. In general, estrogen- ER- α induces cells to differentiation and estrogen- ER- β stimulates cells to proliferation. Estrogen has been related to cancers, such as breast cancer (ER- α action) and prostate cancer (ER- β action). The preantral follicular granulosa cells are estrogenic target tissues, and these cells express both estrogen receptors [3]. Many synthetic estrogen compounds are made for selectively activation ER- α or ER- β . DPN is a selective agonist for ER- β , while PPT is a selective agonist for ER- α . Preantral follicles were treated for five hours with estrogen, which will activate both estrogen receptors, DPN, PPT and plain culture medium as control. RNAs were extracted from those four samples and will be reversely transcribed to cDNA, which are hybridized to rat Affymetrix gene chips that contains 15866 genes. In this paper we describe the basic idea of contrast data analysis, and apply it to estrogen regulation.

2. Conducting contrast data analysis

Data mining methods [2] have proven to be very effective in microarray analysis and other applications in bioinformatics. However, these applications also have unique features and requirements which are not addressed by generic data mining methods, and so far no *single individual* data mining method can provide a satisfactory answer. For microarray analysis, data mining methods have been used in somewhat trial-and-error, *ad hoc* manner.

Contrast data analysis is intended to serve as a methodology for researchers to systematically handle issues related to analysis of data, and is thus complementary to existing approaches in data analysis. Technically, contrast data analysis can be conducted through granular computing (GrC) [1]. GrC refers to the study that makes use of granules in the process of problem solving.

The actual steps involved in contrast data analysis may vary due to the objective of the analysis. For our purpose, we first worked on data preparation. As for the analysis proper, we have carried out our tasks in two phases:

Phase I: Analysis of general aspects using data mining.

This gives us the big picture about the data.

Phase II: Based on the big picture obtained, find specific genes we are interested (using different criteria).

2.1 Data Preparation

There are four single arrays in this experiment as follows. The file format is Excel files storing experiment data generated by Affymetrix Microarray Suite system. There are 15866 rows in each array. Each row represents a gene.

Group1 -- The control group (baseline array).

Group2 -- Estrogen was added. Estrogen is a natural ligand for estrogen receptor.

Group3 -- PPT was added, which is an agonist of ER alpha.

Group4 -- DPN was added, which is an agonist of ER beta.

Attributes in each signal array are geneID, Signal and Detection. The contrast data is then prepared based on single array data. For this purpose, two samples, hybridized to two GeneChip probe arrays of the same type, are compared against each other in order to detect and quantify changes in gene expression. One array is designated as the baseline and the other as an experiment. The experiment file is analyzed in comparison to the baseline file. In this work, group1 is designated as the baseline array, and group2, group3 and group4 work as

the experiment arrays respectively. The attributes in each comparison group are GeneChip Position ID, and Change.

2.2 Phase I: Get the general picture

The contrast data obtained from data preparation are ready for various data mining techniques such as clustering, classification or association rule mining. For the objective of our application, this phase mainly consists of two major tasks: The first is to identify robust increase or decrease, while the second is to determine the interrelationship among attributes such as log ration, detection, and so on. The first task can be cast as a classification problem, whereas the second one can be cast as a problem of association rule mining.

For convenience, we used a software called Rosetta to conduct classification data mining using rough set approach. In order to further find relationships among the attributes, we have used DBMiner as an on-line analytical data mining system that runs on top of the Microsoft SQL Server 2000 Analysis Platform. As an example, a query has been written in DMQL (the query language in DBMiner) to analyze group2 versus group1 with group1 as the baseline concerning estrogen up-regulation genes. The result is shown in Figure 2 (Note that there may be a need for removing redundant rules).

	Body	Implics	Head	Support(%)	Confidence(%)
1	Ratio_D1 = [High]	==>	Change_D1 = [I] AND Detection_D1 = [P]	40.541	66.176
2	Ratio_D1 = [Middle]	==>	Change_D1 = [I] AND Detection_D1 = [P]	29.75	78.371
3	Detection_D1 = [A]	==>	Change_D1 = [I]	13.514	71.429
4	Change_D1 = [I]	==>	Detection_D1 = [P]	70.27	82.105
5	Detection_D1 = [P]	==>	Change_D1 = [I]	70.27	88.436
6	Change_D1 = [I]	==>	Ratio_D1 = [High]	52.252	81.933
7	Detection_D1 = [A]	==>	Ratio_D1 = [High]	13.514	71.429

Figure 2. Partial result of association rule mining

2.3 Phase II: Construct subspaces to identify individual genes

Data mining techniques applied in the first phase provides a general picture. In Phase II, comparative query analysis is conducted to sort the gene expression data and discover the gene-regulation patterns. The specific genes that are up/down-regulated by estrogen receptors, ER alpha or ER beta were retrieved respectively.

For convenience, we use two simple operators R_u and R_d , so that $x R_u y$ denotes up-regulated with y while $x R_d y$ denotes x is down-regulated with y . In addition, we

use $|x R_u y|$ and $|x R_d y|$ to denote the degree of up or down regulation respectively. Note that both $|x R_u y|$ and $|x R_d y|$ are represented by the corresponding contrast data items in the table representing contrast data. For example, after the first phase, in order to further find which genes are down-regulated by both ER alpha and ER beta and the responsiveness of ER beta is greater than that of ER alpha we can construct the following gene subpace:

$$\{g \mid (g R_d ER \text{ alpha}) \text{ and } (g R_d ER \text{ beta}) \text{ and } (|g R_d EP \text{ alpha}| > |g R_d EP \text{ beta}|)\}$$

This abstract query can be converted as an SQL in Oracle database. Upon identification of the relevant genes, more detailed database search can then be conducted on specific genes using various publicly-available bioinformatics databases.

3. Preliminary findings

New results have been obtained through contrast data analysis. Among other things, we have found that estrogen regulates many key components of several important protein kinase signaling pathways, mainly PI³ kinase and NF-kappaB, that involve cell survival. Furthermore, two estrogen responsive genes that are of importance of ovarian function were discovered and they are inhibin A and TGF beta. We provided experimental evidences that estrogen through ER alpha impacted genes related to cell differentiation, whereas ER beta regulated cell proliferation genes. Most significantly, regulated genes retrieved from the experimental data faithfully meet the criteria that Affimetrix system requires but does not provide any practical and analytical tools to resolve. It is convinible to retrieve regulated genes from large data sets and complicate experimental designs based on those multiple analysis.

Acknowledgement.

Z. Chen's work reported in this paper, is supported, in part, by a grant from NIH (P20 RR 16469).

References

- [1] Z. Chen, Bioinformatics as a study of structured granules, Int'l Conf. on Computational Intelligence and Natural Computing, *Proc. JCIS*, pp. 1629-1632, 2003.
- [2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.
- [3] P. Yang, A. Kritchko and S.K. Roy, Expression of ER-alpha and ER-beta in the Hamster Ovary: Differential Regulation by Gonadotropins and Ovarian Steroid Hormones, *Endocrinology* 143(6): 2385-2398.