

A Network-Based Array Data Interpreter

Kotoko Nakata¹
nakata@nihs.go.jp

Kyoko Toda¹
toda@nihs.go.jp

Eiichiro Ichiishi²
ichiishi@koto.kpu-m.ac.jp

Tsuguchika Kaminuma¹
kaminuma@cbi.or.jp

¹ Division of Chem-Bio Informatics, National Institute of Health Sciences, 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan

² First Department of Medicine, Kyoto Prefectural University of Medicine, 465 Kajicho, Kamigyo-ku, Kyoto 602-8566, Japan

Keywords: gene expression, activation, depression

1 Introduction

DNA microarray analysis is a powerful method to analyze genome-wide gene expressions. The number of human genes is estimated to be approximately 35,000. The so called DNA chip or microarray can assemble more than ten thousand genes on a single tiny plate, and even entire genome can easily be assembled on a single plate in the near future. Thus one experiment using a single chip or microarray produces data comparable to thousands of Northern blotting experiments. However, it is not easy to extract useful meaning out of these voluminous data.

In order to interpret expression data one need to know the inter relationships among up regulated and down regulated gene products. Though a number of software packages have been developed for this purpose, they usually do not take the network and pathway relations of the gene products explicitly. We are developing a network-based array analyses data interpreter using a commercial software package, GeneSpring (Silicon Genetics) [4] and the CSNDB (Cell Signaling Networks Database) [2, 3], which is a signal transduction pathway database developed by our group. This method was applied to several DNA chip analyses.

2 Method and Results

In order to concentrate informatics and computational aspects of the data mining problem, we ignore any experimental problems of the array technology such as accuracy or stability. Therefore the linear relation between measured optical intensities and gene expression intensities is assumed, and the optical intensity of each spot on an array is the measure of the expression of the corresponding gene. The array analyses either by Affymetrix type DNA chip or cDNA microarray give the data that tell whether the gene is over expressed or depressed as compared to certain standard measure.

Thus data can be mapped onto a n-dimension gene expression space (where n corresponds to the number of genes assembled on the plate) to which any standard analysis methodology of statistical analysis or pattern recognition can be applied. Usually experiments are repeated by different conditions or in a time course, therefore t times n-dimensional data, where t is the repetition number, is produced and is subjected for data analysis such as clustering. Software packages developed until today are designed to carry out this type of analysis. In such studies of the role of genes clustered into one group is interpreted by researchers empirical knowledge.

Our basic approach is to interpret the array analysis data by mapping genes to their products and to their roles in protein-protein interaction network. There already exist such network or so called

pathway databases for our purpose, and we chose KEGG [1] for metabolic pathways and the CSNDB for signal transduction pathways. We chose GeneSpring as our entry system and tried to map data output from the latter to either KEGG or the CSNDB. Fortunately the new version of GeneSpring has such interface, we have tried only to develop an interface between GeneSpring and the CSNDB and to add resultant interpretation module.

3 Application to Pharmaceutical Research

Array analyses are often used to see the difference in gene expression of a biological system under different conditions. This technique is called as differential display. Measuring effects of drugs are good example of such method. Advanced of pharmaceutical research now revised many drug targets at molecular levels. Today many drug designers are working with lock and key images using computational models. However the structured-based lock and key theory is not enough, for the stimulus triggered by fitting a key to the lock is the real effect of the drug. We thus devised a concept called Post Binding Signal Transduction Pathways (PBSTPs) in that how the cell responds to the stimulus by pathways following the binding. Our CSNDB is a good step stone for developing such model systems.

This method will apply to the data acquired through absolute analysis of the Affymetrix GeneChip expression analysis algorithm, and others. The example will be shown and discussed in the poster.

References

- [1] Kanehisa, M. and Goto, S., KEGG: Kyoto Encyclopedis of Genes and Genomes, *Nucl. Acids. Res.*, 28:27–30, 2000. (<http://www.genome.ad.jp/kegg/pathway/map/map01100.html>)
- [2] Takai-Igarashi, T. and Kaminuma, T., A pathway finding system for the cell signaling networks database, *In Silico Biol.*, 1:129–146, 1999.
- [3] Takai-Igarashi, T., Nadaoka, Y., and Kaminuma, T., A database for cell signaling networks, *J. Comp. Biol.*, 5:747–754, 1998. (<http://geo.nihs.go.jp/csndb/>)
- [4] <http://www.sigenetics.com/cgi/SiG.cgi/index.smf>