Diagnosis System of Drug Sensitivity of Cancer Using cDNA Microarray and Multivariate Statistical Analysis

Tatsuhiko Tsunoda[1] tatsu@ims.u-tokyo.ac.jp
Yoshiaki Hojo[2] hojo@ims.u-tokyo.ac.jp
Chikashi Kihara[1] chika@ims.u-tokyo.ac.jp
Norihiko Shiraishi[1] katoben@ims.u-tokyo.ac.jp
Osamu Kitahara[1] osamuk@ims.u-tokyo.ac.jp
Kenji Ono[1] onoken@ims.u-tokyo.ac.jp
Toshihiro Tanaka[1] toshitan@ims.u-tokyo.ac.jp
Toshihisa Takagi[1] takagi@ims.u-tokyo.ac.jp
Yusuke Nakamura[1] yusuke@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2] Hitachi, Ltd. Information Systems Group. Shinsuna Plaza 6-27, Shinsuna 1-chome, Koto-ku, Tokyo, 136-8632, Japan

Keywords: personalized medicine, drug sensitivity, cancer, microarray, multivariate statistical analysis

1 Introduction

One of the final goals of medicine is to apply adequate therapies to individuals according to the genetic background and environment of each. Inadequate therapy, e.g. applying ineffective drugs, might even cause side effects to patients. Practical diagnosis of complex diseases, e.g. cancer, and drug sensitivity of each requires high throughput profiling of gene expression in tissues. Currently, the most promising technology for it is microarrays [1]. In our method, each cDNA from a cDNA library is spotted (target) at one segment on glass-plates (4108*double/plate), with which mixture of cDNAs from normal cell mRNAs and cancer cell mRNAs (each is labeled with different fluorescence dyes; probes) are simultaneously hybridized; it finally identifies the amount-ratio of each mRNA from cancer cells versus normal cells. Such microarrays provide effective information for identifying cancers, and make us easy to discriminate drug-sensitive cancers from others. We propose the discriminating method using cDNA microarrays and multivariate statistical analysis, hypothesizing that the sensitivity can be detected by the mRNA expression pattern. It normalizes data, clusters and selects representative genes for the discrimination, and finally discriminates drug-sensitive cancer from non-sensitive ones using the quantification theory.

2 Method and Results

In this analysis, a set of 4108 cDNA clones are used for the targets although the size is getting large (currently > 20000). For probes, we prepared two types: one is a mixture of cDNAs (RT of aRNAs) from drug-sensitive human tumor cells (labeled with Cy3-dCTP), and cDNAs from normal cells from the same part of tissue (labeled with Cy5-dCTP), which are hybridized with targets within the spots on a glass plate (responder, R). The other is a mixture of cDNAs (RT of aRNAs) from drug non-sensitive tumor cells from a human tissue (labeled with Cy3), and cDNAs from normal cells from the same part of tissue (labeled with Cy5), which are hybridized with targets within the spots on another glass plate (non-responder, N). Each plate is scanned respectively for getting image data. Since they sometimes suffer from dusty spots or incomplete hybridization, we prepare a set of 4 (2/plate * 2

plates) under the same condition. Statistical analysis and decision by majority suppress the noise and finally provide the precise amount of expression. We also applied the median filter for suppressing the dusty marks.

First, adjusting the parameter so that the expression levels of 50 expression keeping genes are equal (average), we can manage to normalize the amount of mRNA between the three types of cells. Second, from 4 spots for the same condition, Cy3/Cy5 ratio is calculated using decision by majority. Third, cut-off for each expression level is automatically calculated using the relevance coefficient, since low amount of expression data disables reliable analysis. Finally, comparing the expression levels, relative gene expression is classified into 4 types of categories: up-regulated, down-regulated, not-change, and non-detectable (Figure 1).

Since simultaneous handling of many genes (hundreds of thousands in the future) requires too much computation, our method firstly selects significant genes. First, it calculates how each gene contributes to discriminating R from N. With a given cut-off, the subset of significant genes are selected. Finally, by calculating the dependency between every pair of genes in the subset, the method clusterizes the genes, and further selects the representative genes of which relevance to the discrimination are maximum in each group.

After the gene selection, we applied a multivariate statistical analysis to the behavior of the selected genes. Each gene is assigned a weight according to the expression level. Summing the gene expression levels multiplied by each weight, the cancer from each patient is diagnosed whether it is sensitive to the drug or not according to the total score. The quantification theory II provides the optimum weights so that the inter-class variance between responder group and non-responder group gets the maximum value, which discriminates R from N.

We applied our method to identification of CDDP + 5-fluorouracil drug sensitivity (postoperatively) of esophageal cancer: 4 drug responders, 4 drug non-responders, and normal cells. The gene expression is transformed into the 4 categories (relative to normal: up, down, no-change, nd). The gene selection algorithm automatically selected 5 genes (Table 1). Applying the multivariate analysis to these genes, we managed to discriminate R from N (Table 2). The weight for each gene shows how it contributes to the discrimination, and correlates well with the coefficients used in the gene clustering, which will provide information for estimating function of novel genes. Our algorithm will directly contribute to custom-made (order-made, personalized) therapies.

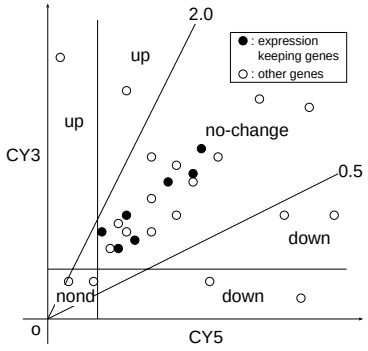

Figure 1. Categories for Cy3/Cy5.

Table 1. Representative genes, χ^2 to sensitivity, and weights.

#IMS	χ^2	up	down	no-c	nd
A206	6.0	0.10	-0.27	-0.70	0.15
A908	5.3	-0.37	-0.33	-0.32	0.55
A2553	4.8	-0.15	0.46	-0.76	-0.15
A5026	4.8	-0.35	-0.35	-0.35	0.21
A950	4.8	0.16	0.79	0.16	-0.47

Table 2. Exp. results.

	score
responder-1	-0.96
responder-2	-2.17
responder-3	-1.75
responder-4	-1.42
non-responder-1	1.23
non-responder-2	0.84
non-responder-3	2.15
non-responder-4	2.10

Acknowledgements

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, “Genome Science” from the Ministry of Education, Science, Sports, and Culture, Japan.

References

[1] *Nature Genetics Supplement*, 21(1), 1999.