Extraction of Informative Genes from Integrated Microarray Data

Dongwan Hong[1], Jongkeun Lee[1], Sangkyoon Hong[1],
Jeehee Yoon[1], and Sanghyun Park[2]

[1] Division of Information and Communication Engineering, Hallym University,
Okcheon-Dong, Chuncheon, 200-702, Korea
{dwhong,jeikei,kyoons,jhyoon}@hallym.ac.kr
[2] Department of Computer Science, Yonsei University,
Shinchon-Dong, Seodaemun-Gu, Seoul, 120-749, Korea
sanghyun@cs.yonsei.ac.kr

Abstract. We have recently proposed a rank-based approach as a new microarray data integration method. The rank-based approach, which converts the expression value of each sample into a rank value within the sample, enables us to directly integrate samples generated by different laboratories and microarray technologies. In this study, we show that a non-parametric scoring method can be efficiently employed for the rank-based data, and informative genes can be effectively extracted from the integrated rank-based data. To verify the statistical significance of the scoring results from the rank-based data, we compared the distribution of the score statistics to a set of distributions obtained from the randomly column-permuted data. We also validate our methods with experimental study using publicly available prostate microarray data. We compared the informative genes extracted from each individual data to the informative genes extracted from the integrated data. The results show that we can extract important prostate marker genes by directly integrating inter-study microarray data, which are missed in either single analysis.

Keywords: Informative genes selection, microarray data integration, prostate cancer, statistical significance verification.

1 Introduction

Microarray experiments enable scientists to obtain a tremendous amount of gene expression data at one time, so they are effectively used in identifying the phenotypes of diseases. In general, increasing sample size is quite desirable for more reliable and valid results. However, microarray experiments are still cost-expensive, so it is hard in reality to obtain experimental results based on a large number of samples. Thus, the experimental results from different investigations with the same research goals are somewhat different and usually contain many errors.

With the rapid accumulation of microarray data, it is of great interest and challenge to integrate inter-study microarray data to increase sample size, which leads to better experimental results. In our earlier work [1], we proposed a new

A. An et al. (Eds.): ISMIS 2008, LNAI 4994, pp. 618–627, 2008.

microarray integration method using a rank-based approach. The rank-based approach, which simply converts the expression value of each sample into a rank value within the sample, enables us directly integrate samples generated by different laboratories and microarray technologies.

In this study, we show that a non-parametric scoring method can be efficiently employed for the rank-based data, and informative genes can be effectively extracted from the integrated rank-based data. As a non-parametric scoring method, Park's method [2] is employed. However, as the scoring method compares the sample values of each gene to calculate a score, it may give slightly different score results when it is applied to the rank-based data and the actual expression value data, respectively. Here we verify the statistical significance of the scoring result from the rank-based data. We compared the distribution of the score statistics to a set of distributions which is obtained from the randomly column-permuted data. Golub's leukemia data [3] was tested, and its result was significant with the p-value of 0.0005 for the rank-based data. Then we compared the informative genes extracted from the rank-based data to the informative genes extracted from the actual expression value data. To exemplify the effectiveness of our integration method, we used three publicly available prostate microarray data. We compared the informative genes extracted from each individual data to the informative genes extracted from the integrated data. The results reveal that important marker genes are selected from the integrated data, which are missed from a single data.

2 Related Works

Experimental microarray data are organized as matrices where rows represent genes and columns represent samples. However, even when considering the microarray data with the same research goals, differences in platforms, protocols, set of genes, and scales of gene expression values lead to difficulties in integrating microarray data across experiments.

To integrate microarray data, the typical methods include *meta-analysis method* [4], *normalization and transformation method* [5, 6], and *rank-based approach* [1]. Instead of comparing microarray expression values from individual experiments, *meta-analysis method* combines the results of individual experiments by using statistical technique. However, there are many cases where the individual experimental results are not reliable due to the small sample size. So the integration of these results may bring an even worse analysis. *Normalization and transformation method* transforms the gene expression values of individual experimental data into a common scale, and then integrates inter-study data [5]. A classical method is the z-score transformation [6], which normalizes the expression values with the mean and standard deviation of each sample. Statistical tests, such as fold ratio, z ratio/test [6], and t statistical test, can be applied directly to the normalized data for predicting significant changes in gene expressions. However, there is still no consensus on the best method to perform data normalization [7]. *Rank-based approach* converts the expression value of each