Exploratory Data Analysis for microarray experiments with replicates

EUN-KYUNG LEE[1], SUNG-GON YI [2], TAESUNG PARK [3]

ABSTRACT

Exploratory data analysis(EDA) is the initial stage of data analysis and provides a useful overview about the whole microarray experiment. If the experiments are replicated, the analyst should check the quality and reliability of microarray data within same experimental condition before the deeper statistical analysis. We shows EDA method focusing on the quality and reproducibility for replicates.

Keywords. Exploratory Data Analysis, Microarray, Reproducibility.

1. INTRODUCTION

Microarray experiments generate huge data sets and most microarray experiment allows multiple slides under the same experimental condition via technical or biological replicates or multiple samples. If these experiments are reliable, they should produce the same results under the same experimental condition. Therefore it is important to check the reproducibility as well as the quality of microarray data.

Recently, Bunesset *al.*(2005) developed a software package arrayMagic for quality control of two channel cDNA microarray data. They considered several quality scores and provided the color map of similarities. However, they did not consider the reproducibility of replicates.

In this paper, we explain the method to explore microarray data with replicates in the reproducibility point of view, and propose a new method to check the reproducibility. We apply our new method to microarray experiment data with three replicates in two treatments.

[1]Department of Statistics, Seoul National University San 56-1, Sillim-dong Gwanak-gu, Seoul, 151-742, Korea (e-mail : gracesle@snu.ac.kr)

[2]Department of Statistics, Seoul National University San 56-1, Sillim-dong Gwanak-gu, Seoul, 151-742, Korea (e-mail : skon@kr.freebsd.org)

[3]Department of Statistics, Seoul National University San 56-1, Sillim-dong Gwanak-gu, Seoul, 151-742, Korea(e-mail : tspart@snu.ac.kr)

2. Data Description

Microarray data from mouse immune response study(Jain *et al.* (2003)) are used in this paper. Cytotoxic T-cells play an important role in mouse immune system. T-cells in the lungs suggests that the process is dependent on the activation status of the adoptively transferred T-cells. Triplicate microarrays of Affymetrix chip, containing 12488 genes were used to investigate each of the two populations of immune exposure : Naive(c1, c2 and c3) and Activated(t1, t2,and t3). Signal intensity values were obtained from the MAS 5.0.

3. Analysis With Graphs

The initial stage of the microarray data analysis is usually exploratory. Exploratory data analysis(EDA) provides the first contact with data. The techniques of EDA consist of a number of informal steps including checking the quality of the data, calculating simple summary statistics and constructing appropriate graphs. Also emphasis is on visual displays that have been a manor contribution of EDA. In this section, we explain a couple of most commonly used graphs.

Boxplots and scatter plots are simple to check overall quality of experiments. One boxplot shows the distribution of one experiment. With parallel boxplots, it is easy to figure out all the distributions of experiments. In figure 3.1, boxplots before normalization shows that medians of Naive groups are different from medians of Activated group. Also, IQRs of 6 experiments are different. After normalization, distributions of six experiments are quite similar and Naive group has more outliers than Activated group. From these plots, we can conclude that the preprocess and normalization works well. These plots are usually used to compare chip variations after normalization. However there is no comparison between experiments for each gene and it is difficult to check the reproducibility.

In scatterplots, we can check the relationship of all pairs of experiments. All pairs in the same condition should show stronger linear relationship than the relationship between two experiments from different conditions. Therefore, we can check the reproducibility for pairs at the first glance. In Figure 3.1, replicates within treatment(Naive or Activated) are quite similar, especially when the expression values are high. On the other hands, scatter plots of between treatments are quite different. Therefore we can conclude that this data looks reliable and reproducible.

Parallel coordinate plots (Wegman, 1990), or profile plots, are displays that

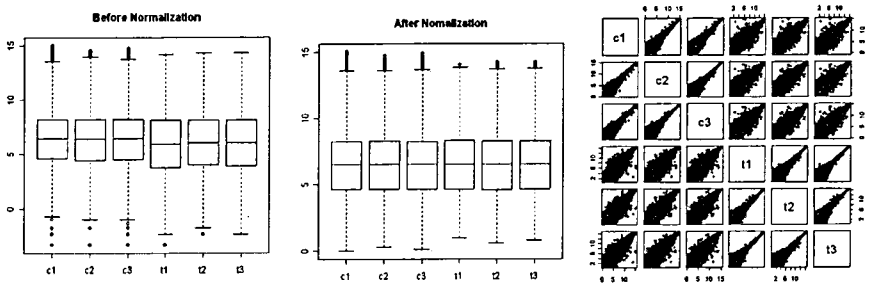


Figure 3.1: Boxplots and Scatter plot matrix

use line segments to represent each experiment and connect the same gene in each experiment with a line. These displays are quite useful when the experiments are ordered by time or when the replicates of one treatment are more than two. For exploratory analysis in the reproducibility point of view, parallel coordinate plots of all experiments in the same condition have straight horizontal lines. However, there is a limitation of scatter plot and parallel coordinate plot. For thousands of genes the points and lines are seriously overplotted as not more useful than a mess.

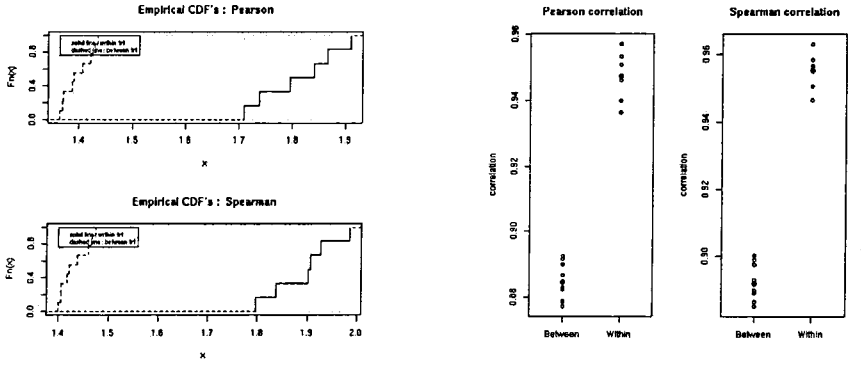
4. Reproducibility check

First, we consider the correlations within and between treatments.

The pearson correlation matrix for mouse data

	c1	c2	c3	t1	t2	t3
c1	1.000	0.938	0.936	0.863	0.867	0.865
c2	0.938	1.000	0.945	0.879	0.880	0.880
c3	0.936	0.945	1.000	0.869	0.878	0.872
t1	0.863	0.879	0.869	1.000	0.945	0.950
t2	0.867	0.880	0.878	0.945	1.000	0.944
t3	0.865	0.880	0.872	0.950	0.944	1.000

If the experiments are reproducible, the correlations within treatments should be close to one and larger than the correlations between treatments. To compare these two sets of correlations, we use Kolmogorov-Smirnov test and Wilcoxon rank sum test. In Figure 4.1(a), solid line shows the empirical cdf of the correlations within treatments and the dotted line shows the empirical cdf of the correlations



(a) The empirical CDF's of between and within correlation (b) plot of between and within correlations

Figure 4.1: Plots for correlation test

between treatments. The empirical cdf of the within correlations is positioned to the right side of the cdf of the between correlations. In Figure 4.1(b), the between correlations are quite smaller than the within correlations. Therefore we can conclude that the correlations within treatments are larger than the correlations between treatments.

5. Remarks

EDA provides a useful overview about the whole microarray experiments. If the experiment are replicated, the analyst should check the quality and reliability of microarray data within same experimental condition before the deeper statistical analysis. We introduce graphical methods focusing on the quality and reproducibility for replicates and procedures for checking the reproducibility. These methods can provide a guideline for analyst to determine the overall reliability and reproducibility for experiments.

Acknowledgements

The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation.

References

BUNESS, A., HUBER, W., STEINER, K., SULTMANN, H., AND POUSTKA, A. (2005)."arrayMagic : two-colour cDNA microarray quality control and preprocessing", *Bioinformatics.* **21(4)**, 554–556.

JAIN, N., THATTE, J., BRACIALE, T., LEY, K., O'CONNELL, M., AND LEE, J. K. (2003). "Local - pooled - error test for identifying differentially expressed genes with a small number of replicated microarrays", *Bioinformatics.* **19(15)**, 1945–1951.

PARK, T., YI, S-G., LEE, S. Y., AND LEE, J. K. (2005)."Diagnostic plots for detecting outlying slides in a cDNA microarray experiment", *BioTechniques*, **38**, 463–471.

WEGMAN, E. (1990). "Hyperdimensional Data Analysis Using Parallel Coordinates", *Journal of American Statistics Association*, **85**, 664–675.