Estimating Confidence Limits for Differential Expression of Genes Based on Control Experiments

M. Singh[1], P. Guo[2] and M. Baum[2]

1 Computer and Biometric Services Unit 2 Integrated Gene Management Program

ICARDA International Center for Agricultural Research in the Dry Areas

Summary

In order to identify significantly differentially expressed genes in a two-channel microarray, one approach is to assign a threshold level. The gene is considered to have differentially expressed if the expression level falls outside the threshold. This paper describes an alternative method to estimate threshold levels with a given confidence coefficient, by analyzing the distribution of extremes of expression levels obtained from control hybridization experiments.

Introduction

In microarray experiments, a large number of genes are assayed for their expression levels and the data used to examine the patterns of gene expression. The objective is to identify genes which are significantly differentially expressed.

- cDNA fragments or oligos on the array are hybridized with genes labeled with two fluorescent dyes (red Cy5, green Cy3) which represent two different experimental conditions
- Ratio of fluorescence intensities between two channels for each gene could be considered as fold change in gene expression level between the two experimental conditions

These expressions are normalized and transformed for desirable statistical behavior (Ivana Yang *et al.*, 2002; Yee Yang *et al.*, 2002; Quackenbush, 2002).

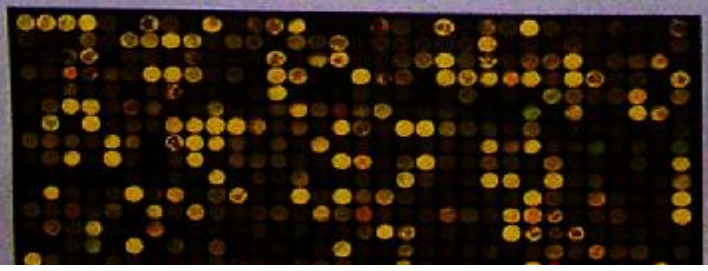
Differential expression (DE) of gene showing non-random variation: cut-off point is pre-assigned (i.e. using a fixed fold-change cut-off point), or determined in terms of mean and standard deviation of the expressions on a gene.

In our understanding, DE will occur on the extremes of the distribution of expression levels. Therefore, it would be worthwhile to estimate thresholds/limits (with a given confidence or coverage probability) for the extremes in a given random sample.

We suggest an alternative. If we could estimate the limits based on the control experiment – in which the same RNA was labeled with both Cy3 and Cy5 and hybridized to the same cDNA array – the limits so determined will serve as a DE threshold in other experiments as well. Our study aimed to develop methods to evaluate the limits of extremes of gene expressions, based on statistical distribution theory. These limits could be then be used on real data.

Statistical methods

Let $x_1, x_2, x_3, \ldots x_n$ be the observed expression levels obtained from a single channel microarray of n genes assayed on tissues from the same genotype, e.g. RNA from a given wheat genotype. The expression level likely to be DE will lie in the right or left tail of the distribution.



Let $x_{(1)}$ and $x_{(n)}$ be the minimum and maximum values of the sample (see Johnson and Kotz 1970, for properties of extreme values).

We assume that the maximum $x_{(n)}$ follows Type I Extreme Value distribution with following distribution:

$$\Pr[X \le x] = \exp(-\exp(-(x-\xi)/\theta))$$

Using this distribution function, we can obtain distribution of the minimum $x_{(1)}$ as well.

In order to estimate the parameters ξ and θ:

(i) We generated a bootstrap sample by resampling the observed sample $x_1, x_2, x_3, \ldots x_n$ with replacement, and computed its maximum and minimum values. (ii) Through independently repeated sampling B times, we generated B min/max values. Results can be tabulated for various values of B. (iii) Using the B values of maximums, ξ and θ can be estimated either by the maximum likelihood estimates, available in Genstat; or by method of moments, where the estimators of ξ and θ are:

$$\hat{\theta} = (\sqrt{6}/\pi) \times \sigma_{max} \quad \text{and} \quad \hat{\xi} = \bar{x}_{max} - \gamma\hat{\theta}$$

where σ_{max} and $\bar{x}_{max}$ are standard deviation and mean of B bootstrap values of maxima, and γ is Euler constant (0.57722).

Using the above distribution function for extreme values, upper threshold limit for maxima say, $x_{max,\,\alpha/2}$, can be computed as:

$$1-\alpha/2 = \exp(-\exp(-(x_{max,\,\alpha/2} - \hat{\xi})/\hat{\theta}))$$

Or, after simplification

$$x_{max,\,\alpha/2} = \hat{\xi} - \hat{\theta}\log(-\log(1-\alpha/2))$$

Similarly, minus of the bootstrapped minimum values could be used to estimate another set of extreme value distribution parameters, say ξ' and θ'. Let their estimates be denoted by $\hat{\xi}'$ and $\hat{\theta}'$.

Using the same approach, the lower $\alpha/2$ threshold limit for minima, say $x_{min,\,\alpha/2}$ would be

$$x_{min,\,\alpha/2} = (\hat{\xi}' - \hat{\theta}'\log(-\log(1-\alpha/2)))$$

- Thus, a gene with expression value above $x_{max,\,\alpha/2}$ or below $x_{min,\,\alpha/2}$ indicates DE at α probability level of significance.
- If the direction of the expression is known, then expression level exceeding $x_{max,\,\alpha}$ will indicate a significantly expressed up-regulated gene at α probability level. Expression level below $x_{min,\,\alpha}$ will indicate a significantly expressed down-regulated gene.

Remarks

These expressions measure the span of the extremes with a given confidence and can be applied on real data.

References

Quackenbush, J. 2002. Microarray data normalization and transformation. *Nature Genetics Supplement*, 32:496-501.

Yee Hwa Yang, Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematicvariation. *Nucleic Acids Research*, 30 (4), e15.

Ivana V Yang, Chen, E., Hasseman, J.P., Liang, W., Frank, B.C., Wang, S., Sharov, V., Saeed, A.I., White, J., Li, J., Lee, N.H., Yeatman, T.J., and Quackenbush, J. 2002. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology* 3 (11). (http://genomebiology.com/2002/3/11/research/0062.1)

Johnson, N.L. and Kotz, S. *Distributions in Statistics: Continuous Univariate*