

Mega-Project 2: Integrated Gene Management

Estimating Confidence Limits for Differential Expression Genes Based on Control

Experiments

M. Singh¹, Peiguo Guo² and M. Baum²

¹Computer and Biometric Services Unit

²Integrated Gene Management Program



**International Center for Agricultural Research in the Dry Areas
P.O. Box 5466, Aleppo, Syria**

November 2005

Biometric Reports are informal communications or views of staff of Computer and Biometrics Services Unit, ICARDA and their collaborators about applications of statistical techniques in the field of agrobiological, environmental and socio-economic research. These reports are prepared to stimulate thinking and comments of professional colleagues within and outside the Center, and do not bear any endorsement of the Center as its formal publications.

Estimating Confidence Limits for Differential Expression Genes Based on Control Experiments

M. Singh, Peiguo Guo and M. Baum

International Center for Agricultural Research in the Dry Areas (ICARDA), PO Box
5466, Aleppo, Syria

Summary

In order to identify significantly differentially expressed genes in two-channel microarray, one approach is to assign a threshold level. If the expression level of a gene falls outside the threshold, the gene is interpreted to have differentially expressed. Using the distribution of extremes of expression levels obtained from control experiments through the self-versus-self hybridizations, this report presents estimation of threshold levels with a given confidence coefficient.

Introduction

In microarray experiments, a large number of genes are assayed for their expression levels and the data on the expression levels are used to examine the patterns of gene expression. In a microarray experiment, cDNA fragments or oligos on the array will be hybridized with genes labeled with two fluorescence dyes (red- Cy5 and green-Cy3) which represent two different experiment conditions, the ratio of fluorescence intensities between two channels for each gene could be considered as the fold change in gene

expression level between two experimental conditions. Further these expressions have been normalized and transformed for their statistical behavior (Yang *et al.*, 2002; Yang *et al.*, 2002a; Churchill, 2002; Quackenbush, 2002). The main interest in all these experiments is to determine genes which are significantly differentially expressed over a statistical population.

In several microarray experiments, one determines cut-off points for differential expression (DE) of gene showing non-random variation using a fixed fold-change cut-off point, or one determined in terms of mean and standard deviation of the expressions on a gene (References). These methods primarily address confidence estimation of mean expression levels, e.g. using a Z-score (Yang *et al.* 2002).

In our understanding, any DE gene expression level will be on the extremes of the distribution of the expression levels. Therefore, it would be essential to estimate thresholds/limits (with a given confidence or coverage probability) for the extreme measurements observed in given random samples. While in many microarrays, small replications only could be afforded when experimenting when accommodating a number of treatments, biological replicates and technical replicates. Here we believe that 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 threshold for DE in the other experiments as well. The purpose of this study is to develop evaluation method for limits of extremes of the gene

expressions based on statistical distribution theory. These limits could be applied on the data and results interpreted whenever the data becomes available.

Statistical methods

Let $x_1, x_2, x_3 \dots x_n$ denote observed expression levels obtained from a single channel microarray of n genes assayed on tissues from a same genotype, for example, RNA from a given wheat genotype. The expression level likely to be DE will lie in the right or left tail of the sample distribution based on $x_1, x_2, x_3 \dots x_n$. Let $x_{(1)}$ and $x_{(n)}$ denote extreme values, minimum and maximum, of the sample. Johnson and Kotz (1970) discuss in details main properties of the extreme value distribution. We assume that the maximum order statistic $X=x_{(n)}$ follow Type I Extreme Value distribution with following distribution:

$$\Pr[X \leq x] = \exp(-\exp(-(x-x)/q))$$

Using above distribution function, one can obtain distribution of the minimum order statistics $x_{(1)}$ as well.

In order to estimate the parameters, x and q , we first generated a bootstrap sample by resampling the observed sample $x_1, x_2, x_3 \dots x_n$ with replacement and computed its maximum and minimum values. Through independently repeated sampling B times, we generated B bootstrap values of maximum and minimum. Results were tabulated for three values of $B = 200, 500, 1000$. Using the B values of maximums, we used two

methods of estimation of the parameters, x and q . Using a simple method of moments, the estimators of x and q are:

$$\hat{\theta} = (\sqrt{6/\pi}) \times s_{\max} \quad \text{and} \quad \hat{x} = \bar{x}_{\max} - g \hat{q}$$

where s_{\max} and \bar{x}_{\max} are standard deviation and mean of B bootstrap values of maxima, and g is Euler constant (0.57722). The maximum likelihood estimates can be obtained using the Genstat.

Using the above distribution function form of the extreme value distribution, upper $a/2$ threshold limit for maxima say, $x_{\max, a/2}$, can be computed in terms of the estimates as follows:

$$1 - a/2 = \exp(-\exp(-(x_{\max, a/2} - \hat{x})/\hat{q}))$$

Or, after simplification, we get

$$x_{\max, a/2} = \hat{x} - \hat{q} \log(-\log(1 - a/2))$$

Similarly, minus of the bootstrapped minimum values could be used to estimate another extreme value distribution parameters say, x' and q' . Let their estimates be denoted by \hat{x}' and \hat{q}' . Following above approach, the lower $a/2$ threshold limit for minima say, $x_{\min, a/2}$ would be

$$x_{\min, a/2} = -(\hat{\mathbf{x}}' - \hat{\mathbf{q}}' \log(-\log(1-a/2))).$$

Thus, a gene with expression value exceeding $x_{\max, a/2}$ or falling below $x_{\min, a/2}$ will show an evidence of differential expression at a probability level of significance. If the direction of the expression is known, then a gene with expression level exceeding $x_{\max, a}$ will indicate a significantly expressed up-regulated gene at a probability level. Similarly a gene with expression level below $x_{\min, a}$ will indicate a significantly expressed down-regulated gene.

Remarks

These expressions measure the span of the extremes with a given confidence. In the absence of real data, their usefulness is of little value. Authors anticipate applying them once the data sets become available.

References

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

[Yang](#) Yee Hwa, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai and Terence P. Speed. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30, e15.

Deleted: Yang

[Yang](#) Ivana V, Emily Chen, Jeremy P Hasseman, Wei Liang, Bryan C Frank, Shuibang Wang, Vasily Sharov, Alexander I Saeed, Joseph White, Jerry Li, Norman H Lee, Timothy J Yeatman and John Quackenbush. 2002. within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology* 3:research0062.1-0062.12

Deleted: Yang

Churchill, G. A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* 32, 490 - 495

Johnson, N. L. and Kotz, S., (1970) *Distributions in Statistics: Continuous Univariate Distributions – I*. John Wiley & Sons, New York

ESTIMATING CONFIDENCE LIMITS FOR DIFFERENTIAL EXPRESSION OF GENES BASED ON CONTROL EXPERIMENTS



M. SINGH¹, P. GUO² AND M. BAUM²

¹ Computer and Biometric Services Unit ² Integrated Gene Management Program

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 (Jiang Yang *et al.*, 2002; Yue 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 threshold 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, \dots, 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:

$$P(X \leq x) = \exp\{-\exp[-(x-\xi)/\theta]\}$$

Using this distribution function, we can obtain distribution of the maximum $x_{(n)}$ as $n \rightarrow \infty$.

In order to estimate the parameters ξ and θ :

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

$$\hat{\xi} = (\sqrt{B})^{-1}(\bar{x}_{(n)} - \sigma_{(n)}^2) \quad \text{and} \quad \hat{\theta} = \bar{x}_{(n)} - \gamma \hat{\sigma}$$

where $\sigma_{(n)}$ and $\bar{x}_{(n)}$ 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 $x_{(n), \alpha/2}$, can be computed as:

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

Or, after simplification

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

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 $x_{(n), \alpha/2}'$ would be

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

- Thus, a gene with expression value above $x_{(n), \alpha/2}$ or below $x_{(n), \alpha/2}'$ indicates DE at α probability level of significance.
- If the direction of the expression is known, then expression level exceeding $x_{(n), \alpha/2}$ will indicate a significantly expressed up-regulated gene at α probability level. Expression level below $x_{(n), \alpha/2}'$ 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: A Primer*. *Genetics Supplement*, 37:495-501
- Yue Heng Yang, Dabao, S., Luo, F., Liu, G.M., Peng, Y., Nigam, I., and Speed, T.P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acid Research*, 30(14):e15
- Jiang Y Yang, Chen, Y., Huanren, J.P., Gang, W., Frank, M.C., Wang, S., Shao, S., Saad, A.J., White, J., Li, J., Liu, H.H., Younan, T.L., and Quackenbush, J. 2002. While the fold assessing differential expression measures are reproducibility in microarray assays. *Genome Biology*, 3(11): [http://www.genomebiology.com/2002/3\(11\)research0067](http://www.genomebiology.com/2002/3(11)research0067), 7
- Johnson, G.L. and Kotz, S. *Distribution in Systemic Continuous Distributions* (Distribution - 1. John Wiley & Sons, New York.