BMC Systems Biology

Poster presentation

A global statistical test for improved detection of gene activity

David JG Bakewell*[1] and Ernst Wit[2]

Address: [1]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3GJ, UK and [2]Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK

Email: David JG Bakewell* - d.bakewell@liv.ac.uk

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):P10 doi:10.1186/1752-0509-1-S1-P10

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Introduction

Interpreting gene transcription for understanding cellular processes using DNA microarray measurements presents a number of challenges to the investigator. One of these is the relatively large number of genes (mRNA transcript identities) that are measured with relatively few *independent* biological samples. The small number of independent samples limits the statistical inference that can be made about the behaviour of individual genes. Consequently, the microarray end-user is faced with bewildering lists of differentially expressed genes that typically contain false positives – a direct result of the well-known 'few samples – many genes' dimensionality problem. At the same time investigators are using microarrays to understand processes involving the collective action of a number of genes, often organized as a complex or pathway [1]. To address these needs and remedy the problem of dimensionality, we consider a global, likelihood ratio (LR) test for examining the behaviour of a pre-assigned *group* of genes – rather than individual genes.

This poster describes the development of the LR test for detecting low levels of activity in a group of related genes or 'pathway'. The underlying principle of the global LR test is to share variance information across related genes so that small concordant changes in expression in a group of genes can be detected more sensitively than current multiple univariate tests. The LR test takes into account only the magnitude and uncertainty of the changes, that is, it is not affected by whether the genes within the group are up- or down-regulated. Compared with other exploratory methods [2-4] that find patterns of differentially expressed genes, our method is more a confirmatory method. It is particularly well suited to investigating the involvement of a particular functional group of genes in explaining the difference between two or more phenotypes.

Methods

The LR method uses a two-layer hierarchy that models variation between genes within a pathway and experimental variation due to effects of sample preparation, microarray measurements and other sources. Hierarchical maximum likelihood estimation is used to determine the likelihood ratios, which are in-turn used to determine P-values (i.e. how likely the observed data would be if in fact none of the pathway genes are differentially expressed). The global maximum of the likelihood is evaluated using standard differential calculus techniques and numerical methods along similar lines described in previous work [5]. Further details are given in the poster. Comparisons are made with P-values evaluated using standard procedures, such as, the univariate regularised *t*-test [6] with decision rules that take into account the effect of multiple or 'sequential' testing. In the examples described in the poster, the Family Wise Error Rate is controlled by adjusted P-values calculated using the Bonferroni correction or Hochberg multiple step-down procedure [7-9].

Results

The P-values determined by LR and sequential *t*-test methods are compared using data from Monte Carlo (MC) simulations and a published two-channel spotted microarray experiment [10]. The MC simulations generated LR and T-statistics that enabled pathways to be identified as True Positive (Tp) and False Positive (Fp) for a particular cut-off value. Counts of Tp and Fp were combined into Receiver Operating Characteristics (ROCs) similar to earlier work [5]. The MC simulations show the hierarchical LR test yields lower P-values compared with the sequential *t*-test for low cutoff range, 0–5%, of interest. That is, the LR test leads to improved detection of differential gene-group expression compared to a sequential application of the *t*-test.

The LR and t-test methods were also applied to data about three transforming growth factor genes implicated in the regulation of breast cancer tumorigenesis [10,11]. The effect of sample size on the P-values for the two methods was investigated by randomly selecting samples (maximum 10) and taking combinatorial averages. Computed relationships of the P-values versus number of samples shows the LR test requires almost three times fewer independent biological samples to give the same gene-group P-value compared with a multiple comparison *t*-test.

Concluding remark

A global, likelihood ratio test is developed that shares variance information across a pre-assigned group of genes. The test is shown to be a more sensitive detector of differential gene expression than conventional techniques, such as, the multiple comparison *t*-test. This improvement is particularly advantageous for typical of microarray measurements where only a few independent biological samples are available.

Acknowledgements

D. B. would like to thank Cancer Research U.K. and The Wellcome Trust (Project 062511) for financial support and facilities.

References

1. Lewin B: *Genes VII Oxford University Press*; 2002.
2. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics* 2005, **21(13):**2988-2993.
3. Breitling R, Amtmann A, Herzyk P: **Iterative group analysis (iGA): a simple tool to enhance sensivity and faciltate interpretation of microarray experiments.** *BMC Bioinformatics* 2004, **5():**34.
4. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: **Gotoolbox: functional analysis of gene datasets based on gene ontology.** *Genome Biology* 2004, **5(12):**R101.
5. Bakewell DJ, Wit E: **Weighted analysis of microarray gene expression using maximum-likelihood.** *Bioinformatics* 2005, **21(6):**723-729.
6. Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4(4):**210-.
7. Dudoit S, Shaffer JP, Boldrick JC: **Multiple hypothesis testing in microarray experiments.** In *Working Paper Series 110 U. C. Berkeley Division of Biostatistics*; 2002.
8. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica Sinica* 2002, **7:**111-139.
9. Hochberg Y: **A sharper Bonferroni procedure for multiple tests of significance.** *Biometrika* 1988, **75(4):**800-802.
10. West RB, Nuyten DSA, Subramanian S, Nielsen TO, Corless CL, Rubin PR, Montgomery K, Zhu S, Patel R, Hernandez-Boussard T, Goldblum JR, Brown PO, van de Vijver M, van de Rijn M: **Determination of stromal signatures in breast carcinoma.** *PLoS Biology* 2005, **3(6):**e187.
11. Akhurst RJ, Derynck R: **Tgf-beta signalling in cancer – a double-edged sword.** *Trends in Cell Biology* 2001, **11(11):**S44-S51.