行政院國家科學委員會補助專題研究計畫 □ 成 果 報 告
■期中進度報告

小樣本多變數下選取重要變數之研究

計畫類別：■ 個別型計畫 □ 整合型計畫
計畫編號：NSC 97－2118－M－009－001－MY2
執行期間： 2008 年 8 月 1 日至 2010 年 7 月 31 日

計畫主持人：洪慧念
共同主持人：
計畫參與人員： 王怡倫，莊育珊

成果報告類型(依經費核定清單規定繳交)：□精簡報告 □完整報告

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學統計學研究所

中 華 民 國 98 年 5 月 26 日

英文摘要

Microarray technology has been studied widely in multiple hypothesis testing, with thousands or even millions of test statistics z_i's to consider at the same time. These test statistics z_i's are correlated in some unknown form on multiple testing procedure. In this study, we will discuss three possible reasons for the density of histogram of the observed z_i's differs from theoretical standard normal distribution. The three reasons are correlation between genes, correlation among microarrays, and various marginal distribution assumptions. Then, we will use several models to simulate data and show that correlation among microarrays and various distribution assumptions are important reasons which will cause the density of the observed z_i's differs from theoretical standard normal distribution.

中文摘要

在傳統的雙樣本檢定問題中，我們常用的是 t-檢定量。在本年的研究中，我們探討的是基因晶片中多重雙樣本檢定的問題。根據過去統計學家的經驗，我們知道傳統的 t-檢定量的統計分佈並非是 t 分佈。因此，我們想要知道會造成這個現象的主要原因。這些可能原因包括基因之間的相關性，人與人之間的相關性還有基因表現的非常態性。在本年內，我們利用模擬的方式來探討此三種原因所造成的影響，並嘗試修改 t-檢定量的分佈。

報告內容

The analysis of microarray in biomedical research has been studied extensively in the past few years. Microarray is a technology to detect mRNA expression level. In general, detecting mRNA expression level can help identify genes that contribute to disease. That is, the goal of a

microarray experiment is to identify those genes that are differentially expressed within different samples. Besides, we observed the number of samples is much less than the number of genes in a microarray experiment, thus generating a large-scale multiple hypothesis testing problem (Gentleman et al., 2005; Efron, 2007).

A large-scale multiple hypothesis testing problem in a microarray experiment involves the simultaneous test of thousands, or even millions, of null hypotheses (Gentleman et al., 2005). Usually we use two-sample t-statistics, t_i, comparing expression levels under two different conditions for N genes. Then, the t_i's were transformed to z_i's (Efron, 2007). Efron (2007) displayed two histograms of z_i's from two microarray experiments and described the z_i's correlations can cause the histograms of z_i's differ from standard normal distribution. Furthermore, Efron (2007) assessed the size and effect of correlation in large-scale multiple hypothesis testing, particularly false discovery rate (FDR) techniques (Benjamini and Hochberg, 1995).

Since the earlier study did not focus on the reason of the histograms of z_i's differ from standard normal distribution in large-scale multiple hypothesis testing problem. Hence, in our study, we have two purposes: (a) to discuss the possible reasons for the density of histogram of the observed z_i's differs from the density of theoretical standard normal distribution under null distribution; (b) to do simulation experiment to exam the histograms of z_i's differing from standard normal distribution in large-scale multiple hypothesis testing problem.

In the first year of this project, we first reviews the multiple hypothesis testing problem in a microarray experiment and two microarray experiments: the breast cancer study and the HIV

study. And, then we discusses the possible reasons for over-diversion of the density of histogram of z_i's in the breast cancer study and over-converge of the density of histogram of z_i's in the HIV study. Also, we considers three possible models: (1) correlation between genes, (2) correlation among microarrays, and (3) models of the various distribution assumptions. Besides, we apply some simulation data and conclude some results from the simulation. Finally, we use the real data in multiple hypothesis testing problem and make some comments.

參考文獻

[1] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Ser. B, 57, 289-300.

[2] Chan N. H. (2001). Time series applications to finance.

[3] Dudoit, S., Shaffer J. and Boldrick J. (2003). Multiple hypothesis testing in microarray experiments.

[4] Dudoit, S., Yang Y., Callow M. J. and Speed T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.

[5] Efron B., Tibshirani R., Goss V. and Chu G. (2000). Microarrays and their use in a comparative experiment.

[6] Efron, B. (2001). Robbins, empirical bayes, and microarrays.

[7] Efron, B., Tibshirani R., Storey J., and Tusher V. (2001). Empirical bayes analysis of a microarray experiment.

[8] Efron, B. and Tibshirani R. (2002). Empirical bayes methods and false discovery rates for

microarrays.

[9] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99, 96-104. MR2054289

[10] Efron, B. (2005). Local false discovery rates.

[11] Efron, B. (2006). Size, power, and false discovery rates. *The Annals of Statistic*.

[12] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102, 93-103. MR2293302

[13] Ge, Y., Dudoit S., and Speed, T. (2003). Resampling-based multiple testing for microarray data analysis, *Test*, 12, 1-77.

[14] Gentleman R, Carey V., Huber W., Irizarry R., Dudoit S. (2005). Bioinformatics and computational biology solutions using R and bioconductor.

[15] Gold D., Wang J. and Coombes K. (2005). Inter-gene correlation on oligonucleotide arrays: how much does normalization matter?

[16] Gottardo R., Raftery A., Yeung K., and Bumgarner R. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples.

[17] Hedenfalk et al. (2001). Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344, 539-548.

[18] Lockhart, D. J., Dong, H.l., Byrne, M. C., Follettie, M.T., Gallo, M. V. Chee, M. S., Mittmann, M., Wang, C.,Kobayashi, M., Horton, H. & Brown, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology* 14: 1675-1680.

[19] Owen A.(2005). Variance of the number of false discoveries.

[20] Qiu, X., Brooks, A., Klebanov, L., and Yakovlev, A. (2005a). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, 6.

[21] Qiu, X., Klebanov, L., and Yakovlev, A. (2005b). Correlation between gene expression levels and limitations of the empirical bayes methodology in microarray data analysis. *Statistical applications in genetics and molecular biology*, 4, paper 34.k

[22] Shaffer J. P. (1995). Multiple hypothesis testing: a review.

[23] Shumway R., Stoffer D. (2005). Time series analysis and its applications.

[24] van't Wout, A., Lehrma, G., Mikheeva, S., O'Keeffe, G., Katze, M., Bumharner, R., Geiss, G., and Mullins, J. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4**[Trial mode]**-T-Cell lines. *Journal of Virology*, 77, 1392-1402.