Peridocity, Change Detection and Prediction in Microarrays

Introduction

Microarray experiment is a promising technology to monitor the expression levels for thousands of genes simultaneously. This technology is relevant to almost all fields of life sciences. Microarrays provide a more complete understanding of the molecular variations among tumors, and hence direct to better diagnosis and treatment strategies for the disease. DNA microarray experiments, followed in defined time period, are highly suitable to gene expression levels during a biological process. Apart from monitoring transcript or messenger ribonuceic acid (mRNA) levels, DNA microarrays are used to detect single nucleotide changes, unbalanced chromosome aberrations by *Comparative Genomic Hybridization* (CGH) experiment (Nuber, 2005).

The analysis of microarrays demands solving a number of statistical problems ranging from normalization to different supervised and unsupervised studies. Growth and development of any organism requires appropriate regulation of cell division cycle (Whitefield *et al.*, 2002). In cancer cell, the molecular processes for duplication of cell are erratic. So, advent of treatment for cancer or some other diseases might get possible through proper understanding of cell division cycle. There are well established theory and application to test for periodicity in short time series but with Fourier frequencies. However, most of the microarray time series are short and there is no guarantee that the series will only have Fourier frequencies. Wichert *et al.* (2004) discussed the issue of investigating periodicity in the microarray cell cycle data using Fisher's g statistic. Our proposed method can lead to substantial improvement in power of the test

when non-Fourier frequencies are present in the series.

Due to the presence of large number of genes for each single array, the issue of multiple testing in a genome-wide data analysis plays a great rule in reaching the final conclusion. A significant p-value obtained from a given setting for a specific gene would very unlikely refer to randomness rather than true features of this gene. But the presence of large number of genes makes it possible to get false positive and false negatives for a defined hypothesis. Wichert *et al.* (2004) used a method of False Discovery Rate (FDR), first proposed by Benjamini and Hochberg (1995), as multiple testing procedure. False positive rate, which leads to p-value, differ conceptually from FDR. Storey and Tibshirani (2003) suggested working with positive FDR (pFDR) for multiple testing. Pounds *et al.* (2004) proposed a method, called the spacing LOESS histogram (SPLOSH) for estimating the conditional FDR (cFDR) and claimed that this approach is more stable than the $q$value. Simulation results and implementation to real data show the variation of selecting the number of periodically expressed genes through different multiple testing methods. SPLOSH revealed to be most conservative while $q$value approach seems to be liberal in detecting correct number of periodic genes.

Copy number changes, called as chromosome gains or losses, in the DNA content of a given subject's DNA often cause to tumorigenesis. Array CGH is a molecular-cytogenetic method that provides a way to do genomewide screening for such loss and gain regions referring to genetic alterations. To study and solve the challenge of efficiently identifying the regions with DNA copy number alterations, a number of methods have already been proposed. Pollack *et al.* (2002) proposed to use a moving average to the ratios and normal versus normal hybridization to compute the threshold. Maximum likelihood approach to fix mixture models corresponding to gain, loss and normal regions was used by Hodgson *et al.* (2001). An algorithm, proposed by Wang *et al.* (2005), builds hierarchical clustering-style trees along each chromosome, and then selects the clusters by controlling the FDR at a specific level. Wang (1995) develops a method for identifying the jumps in a time series by comparing wavelet co-

efficients of the data with a proposed threshold. In chapter 2, we propose a method using maximum overlapping discrete wavelet transform (MODWT) to detect the amplification or deletion points of DNA copy number. The region is defined to be gain or loss region using bootstrap procedure and thereafter some multiple test procedure.

A successful diagnosis and treatment of cancer depend on the classification of tumors through high-throughput microarray data analysis and this is one of the mostly studied issues in microarray experiment. Golub *et al.* (1999) worked with qualitative disease phenotypes, Brown *et al.* (2000) worked with classifying genes according to their functional role. Comparison of different classification methods was done by Dudoit *et al.* (2002). Traditional k-nearest neighbour method selects single nearest neighbor for the purpose of predicting future observations. In Chapter 3, we consider plausibility of taking first nearest neighbour covariates in the set of original inputs. The performance of four methods in four miroarray and one simulated data set was investigated. We found that this type of augmented covariate set can result in better prediction.

Bibliography

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Nat Acad Sci., USA*, **97**: 262-267.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289-300.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, **97**, 77-87.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21** , 1104-1111.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286 (5439)**, 531-537.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29**, 491.

Holmes, C. C. and Adams N. M. (2003). Likelihood inference in nearest-neighbour classification models. *Biometrika*, **90**, 1, 99-112.

Nuber U.A.(2005) *DNA Microarrays*. Taylor & Francis, New York.

Pollack J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffry, S.S., Lonning, P.E., Tibshirani, R., Botstein D., Borrsen-Dale, A. and Brown, P.O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, USA, **99**, 1296312968

Pounds,S. and Cheng C. (2004) Improving false discovery rate estimation, *Bioinformatics*, **20**, 1737-1745.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci., USA*, **100**, 9440-9445.

Wang, Y. (1995). Jump and Sharp Cusp Detection by Wavelets. *Biometrika*, **82**, 385-397.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., Tibshirani, R. (2005). Studies in crop variation. I. A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 1, 4558

Whitefield,M.L., Sherloc,G., Saldanha,A.J., Murrary,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977-2000.

Wichert,S., Fokianos K. and Strimmer K. (2004) Identifying periodically expressed transcrips in microarray time series data. *Bioinformatics*, **18**, 5-20.