行政院國家科學委員會專題研究計畫　期中進度報告

微陣列資料的統計型態發掘(1/3)

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-009-004-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學統計學研究所

計畫主持人：盧鴻興

報告類型：精簡報告
報告附件：國外研究心得報告
處理方式：本計畫可公開查詢

中　華　民　國 92 年 5 月 29 日

行政院國家科學委員會專題研究計畫成果報告

微陣列資料的統計型態發掘(1/3)

Statistical Pattern Discovery for Microarray Data

計畫編號：NSC 91-2118-M-009-004

執行期限：91 年 8 月 1 日至 92 年 7 月 31 日

主持人：國立交通大學統計學研究所盧鴻興教授

一、中文摘要

針對高效能的微陣列技術中的隨機性，我們需要應用統計方法來發展微陣列資料的的型態發掘，用以提供生物學及醫學研究的深入洞悉。我們這三年的長期計劃將發展一系列新的統計工具，來分析微陣列產生的基因表現資料。在進行微陣列資料探勘與知識發掘工作中，我們將面臨下列的主要議題。

第一步是選取特徵表示，用來定義基因微陣列資料的主要特徵型態。對於具有時間性的資料，多重解析法，例如小波，可以用來作特徵表示。對於不具時間性的資料，則可運用不同的距離測度和離散法來作特徵表示。接著，非線性維度減化的技巧可以進一步應用來調整距離測度和尋找資料適合的維度。這些技巧可以整合群集分析的方法，包含多元尺度法，階層式群聚法等等，作資料的探索分析。

下一步是進行監督式分類。由於現在的微陣列資料只有數百個陣列，但卻有超過數千個基因表現概廓，因此有需要篩檢或濾除出管家基因或無資訊基因。為了解決這些分類問題的困難，我們可以整合先驗訊息和其它相關的資訊到進階的分類法中，來分類樣本和基因的功能。交叉認証法可以用來評量這些方法的預測誤差。

最後，在這後基因體世代中，我們還需要發展新的統計工具，從微陣列資料中推論基因的反應路徑。針對微陣列資料的特性，我們將提出新的統計觀點。這些新的方法將會與現有的方法相比較，找出優缺點。我們也將同時應用國際和國內研究團隊所產生的互補 DNA 晶片和寡核甘酸晶片資料，進行我們的統計研究。

關鍵詞：資料探勘，知識發掘，特徵表示，多重解析分析，小波，群集分析，非線性維度減化，分類法，先驗訊息，預測誤差，交叉認証，反應路徑分析。

Abstract

Due to the inherent randomness in the high throughput technique of microarray, pattern discovery by statistical methods are important to provide insights for biological and medical studies. This three-year project is hence aimed at exploring a series of new statistical tools for analyzing gene expression data generated by microarray. We will focus on the following major issues involved in the tasks of data mining and knowledge discovery for microarray data.

The first is regarding the feature representation to define main patterns for microarray data. For data with time courses,

multiresolution analysis, like wavelets, will be investigated. For data without time course, different distance measures and discretization methods will be studied. Then, nonlinear dimension reduction techniques can be further applied to adjust the distance measures and search for intrinsic dimension. These techniques can be integrated with cluster analysis for exploratory data analysis, including multidimensional scaling, hierarchical clustering, and so forth.

The next issue is about supervised classification. Because current microarray data only have hundred arrays with expression profiles of more than thousands genes, it is important to filter or screen housekeeping or noninformative genes. In order to solve the ill-posedness of these classification problems, prior knowledge and other related information will be incorporated with advanced classification methods to classify samples and the function of genes. Prediction errors by cross-validation will be studied to evaluate the performance.

Finally, it is intended to develop new statistical tools for inferring the pathways from microarray data in this post-genome era. New perspectives that emphasis the particular properties of microarray data will be addressed. Comparisons of all these new methods with existing methods will be performed as well. Both cDNA and oligonucleotide chips produced in international and local research groups will be studied for these methods.

Keywords: Data Mining, Knowledge Discovery, Feature Representation, Multiresolution Analysis, Wavelets, Cluster Analysis, Nonlinear Dimension Reduction, Classification, Prior Knowledge, Prediction Error, Cross-Validation, and Pathway Analysis.

二、緣由與目的

The massive amonut of microarray data bring the big challenge of developing advanced data mining tools by statistical and computational methods, which motivate our great research interests in this three-year project. In particular, these data are high dimensional because the sample number is far smaller than the gene number, which causes the curse of dimensionality and stimulates the development of new data analysis methods (Donoho 2000). Therefore, this long-term project is aimed to develop new techniques to analyze microarray data generated by international and local research laboratories with state-of-art analysis tools and databases in the world for statistically pattern discovery.

Focusing on specific scientific problems, new data mining and knowledge discovery techniques will be developed and investigated. For example, filtering, screening, and exploratory data analysis of microarray data will be investigated. Dimension reduction and visualization techniques will be invented to extract the genuine feature in these data. Integration of related databases and biological knowledge would be performed to verify and confirm new findings. Systematical methods for unsupervised clustering and supervised classification will be developed.

三、結果與討論

本年期計畫到目前為止已完成 3 篇論文，其摘要如下。

1. “Rapid divergence in expression between duplicate genes inferred from microarray data,” *Trends in Genetics*, 18, 12, 609-613, 2002.

Abstract:
For over 30 years, expression divergence has been considered a major reason for retaining duplicated genes in a genome, but how often and how fast duplicate genes diverge in expression has not been studied at the genomic level. Using yeast microarray data, we show that expression divergence between duplicate genes is significantly correlated with their synonymous divergence (K_S) and also with their nonsynonymous divergence (K_A) if K_A 0.3. Thus, expression

divergence increases with evolutionary time, and expression divergence and K_A are initially coupled. More interestingly, a large proportion of duplicate genes have diverged quickly in expression and the vast majority of gene pairs eventually become divergent in expression. Indeed, over 40% of gene pairs show expression divergence even when K_S is 0.10 and the proportion becomes > 80% for K_S > 1.5. Only a small fraction of ancient gene pairs have not shown expression divergence.

2. "Visualization, Screening, and Classification of Cell Cycle-Regulated Genes in Yeast by Multidimensional Scaling, Nonlinear Dimension Reduction and Wavelet Transform".

Abstract:

We propose a new and integrated approach for visualization, screening, and prediction of gene functions using microarray data, based on multidimensional scaling (MDS), nonlinear dimension reduction and wavelet transform. This approached is applied to analyze the cell cycle of yeast in Spellman *et al.* (1998). The results show that this new approach indeed provides a visualization tool, which displays the functional relationships between genes, like the periodical pattern of a cell cycle. This representation can be further used to verify the functions of genes explored by experimental methods in literature. Based on this representation with biological knowledge, screening and prediction of the functions of all genes can be implemented. The performances of different methods in screening and prediction are evaluated by the Jackknife approach in this study. Hence, this integrated approach suggests a new perspective to discover and classify the functions of genes through their expression profiles by microarray. The findings for cell cycle-regulated genes in yeast are also reported and discussed.

3. "Statistical Analysis of the Gene Expression for Non-synchronized Cell Cycles of Human Glioma Cells after Gamma Irradiation by cDNA Microarray".

Abstract:

Microarray is a high throughput technique. We can observe a large number of gene expressions by this new technique in a short time. In order to understand the gene expression profiles in different phases of cell cycles of human glioma cells after gamma irradiation, microarray experiments are performed in Professor Ngo's laboratory at National Yang-Ming University (Ngo, Chan, Chang, 2001). The experiments are non-synchronized to imitate the *in vivo* expression pattern. Furthermore, these will provide fast and economic approaches to screen the gene functions. However, these data bring us challenges in analysis. This study is an attempt to separate the effect of non-synchronization from the genuine cell cycle function by statistical methods for microarray data with cell proportions measured by flow cytometry.

四、計畫成果自評

由上述的報告中，可以發現我們的研究內容與原計畫相符，達成預期的目標。我們將進一步將完成的技術報告投稿到學術期刊發表，並進一步將這些技術應用到實際的微陣列資料，提供更正確和有效的統計分析。因此，本計畫的研究除了在學術上分析方法的突破，也同時具備應用的價值。

五、參考文獻

[1] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci 2000 Aug 29;97(18):10101-6.

[2] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci 2000 Jan 4;97(1):262-7.

[3] Celis JE, Kruhoffer M, Gromova I, Frederiksen C, Ostergaard M, Thykjaer T, Gromov P, Yu J, Palsdottir H, Magnusson N, Orntoft TF. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. FEBS Lett 2000 Aug 25;480(1):2-16.

[4] Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. Making and reading microarrays. Nat Genet 1999 Jan;21(1 Suppl):15-9.

[5] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L,

Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998 Jul;2(1):65-73.
[6] Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, Sapinoso L, Hampton G, Elledge SJ, Davis RW, Lockhart DJ. Transcriptional regulation and function during the human cell cycle. Nat Genet 2001 Jan;27(1):48-54.
[7] Cox TF, Cox MAA. Multidimensional Scaling. 2000 CRC Press, London.
[8] Daubechies I. Ten Lectures on Wavelets. 1992 CBMS-NSF Series of Applied Mathematics, SIAM, Philadelphia.
[9] Donoho DL. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. American Math. Society conference: Math Challenges of the 21st Century, Los Angeles, California, August 6-11, 2000.
[10] Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999 Jan;21(1 Suppl):10-4.
[11] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 1998 Dec 8;95(25):14863-8.
[12] Eisen MB, Brown PO. DNA arrays for analysis of gene expression. Methods Enzymol 1999;303:179-205.
[13] Friedman N, Nachman I, Pe'er D. Using Bayesian Networks to Analyze Expression Data. ECOMB 2000. The Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan.
[14] Halgren RG, Fielden MR, Fong CJ, Zacharewski TR. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. Nucleic Acids Res 2001 Jan 15;29(2):582-8.
[15] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. Gene-expression profiles in hereditary breast cancer. N Engl J Med 2001 Feb 22;344(8):539-48.
[16] Kanehisa M. Post-genome Informatics. 1999 Oxford University Press, Oxford.
[17] Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998 Nov 15;58(22):5009-13.
[18] Knight J. When the chips are down. Nature 2001 Apr 19;410(6831):860-1.
[19] Kohonen T. Self-Organizing Maps. 1997 Second Extended Edition. Springer. New York.
[20] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999 Jan;21(1 Suppl):20-4.
[21] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996 Dec;14(13):1675-80.
[22] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000 Jun 15;405(6788):827-36.
[23] Mallat SG. A Wavelet Tour of Signal Processing. 1999 Second Edition, Academic Press, San Diego.
[24] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000;:455-66.
[25] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995 Oct 20;270(5235):467-70.
[26] Schulze A, Downward J. Analysis of gene expression by microarrays: cell biologist's gold mine or minefield? J Cell Sci 2000 Dec;113 Pt 23:4151-6.
[27] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. Nucleic Acids Res 2001 Jan 1;29(1):152-5.
[28] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998 Dec;9(12):3273-97.
[29] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci 1999 Mar 16;96(6):2907-12.
[30] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet 1999 Jul;22(3):281-5.
[31] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000 Dec 22;290(5500):2319-23.
[32] Wong WH. Computational Molecular Biology. J. Amer. Statist. Assoc. 2000;95(449):322-6.
[33] Zhang H, Yu CY, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci 2001 Jun 5;98(12):6730-5.

附件：封面格式

行政院國家科學委員會補助專題研究計畫成果

報告

※※※※※※※※※※※※※※※※※※※※※※※

※※※※

※

※ 微陣列資料的統計型態發掘(1/3)

※

※

※

※※※※※※※※※※※※※※※※※※※※※※※

※※※

計畫類別：x 個別型計畫 □整合型計畫

計畫編號：NSC91-2118-M-009-004

執行期間：91 年 8 月 1 日至 92 年 7 月 31 日

計畫主持人：國立交通大學統計學研究所盧鴻興教授

共同主持人：

計畫參與人員：

本成果報告包括以下應繳交之附件：

x 赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：

中　華　民　國　92 年　5 月 22 日