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

基因網路和表現型特徵間關聯之統計分析

Statistical analysis of association between gene networks and phenotypic patterns

計畫類別：V 個別型計畫　□整合型計畫

計畫編號：NSC 98-2118-M-009-004-MY3

執行期間：98 年 8 月 1 日至 101 年 7 月 31 日

執行機構及系所：國立交通大學統計學研究所

計畫主持人：盧鴻興教授

共同主持人：

計畫參與人員：

第一年計畫參與人員：蔡孟原、陳亮勳、吳啟豪、許乃文、王儷芬
第二年計畫參與人員：蔡孟原、鄭宇傑、黃偉恆、吳泰言、劉耿瑋
第三年計畫參與人員：蔡孟原、鄭宇傑、魏裕中、翁俊哲、吳宜靜、
周孟穎、陳麗安、李念錡、楊家瓏

成果報告類型(依經費核定清單規定繳交)：完整報告

本計畫除繳交成果報告外，另須繳交以下出國心得報告：

■赴國外出差或研習心得報告

□赴大陸地區出差或研習心得報告

□出席國際學術會議心得報告

□國際合作研究計畫國外研究報告

處理方式：除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

中　華　民　國 101 年 10 月 30 日

行政院國家科學委員會專題研究計畫期中成果報告

基因網路和表現型特徵間關聯之統計分析

Statistical analysis of association between gene networks and phenotypic patterns

計畫編號：NSC 98-2118-M-009-004-MY3

執行期限：98 年 8 月 1 日至 101 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

第一年計畫參與人員：蔡孟原、陳亮勳、吳啟豪、許乃文、王儷芬

第二年計畫參與人員：蔡孟原、鄭宇傑、黃偉恆、吳泰言、劉耿

第三年計畫參與人員：蔡孟原、鄭宇傑、魏裕中、翁俊哲、吳宜靜、周孟穎、陳麗安、李念錡、楊家瓏

一、中文摘要

高傳出技術產生大量異質性的生醫數據，可用來了解大規模基因網路與高維度表現型特徵之關聯。我們計劃藉系統生物學的觀點，結合布林網路、貝氏網路、核方法、維度縮減等技術去發展統計學習方法。

首先我們計劃透過基因型與表現型的相關結構，辨識哪些具共同表現型的基因。因愈多共同表現型的基因應愈傾向於影響彼此的生物功能，交互作用的強度也將幫助計算基因網路中表現型的相似值。我們將透過統計方法和生物資訊的整合，深入了解表現型相關性和基因網路特徵辨識。

另外具共同基因的表現型亦可被視為一種相似的特徵，可能是由相同的機制所造成。因此，計劃發展在基因網路結構下的量度工具，構造一個分析基因型及表現型之相依性的系統性架構。

依此方法，我們能執行任意疾病和組織型態所組成的特定網路分析。例如，能用此法以協助發現及開發高效與低副作用的藥物、治療。藉生醫研究中的基因網路和表現型相關性，我們將在長期計畫中進而深入發展統計方法，有助於疾病的治療和生物研究的分析。

關鍵詞：統計學習、布林網路、貝氏網路、核方法、維度縮減、高維度資料、表現型相關性、基因網路、機制分析、特徵辨識、系統生物、藥物研發、副作用。

Abstract

High throughput techniques have generated massive amount and heterogeneous types of biological and medical data that are in the immediate need of statistical analysis for association between structure in the large scale gene networks and the trait patterns of phenotypes in high dimensional space. In this long term project, we plan to develop statistical learning methods for the analysis and integration of different kinds of data by systematic approaches of network biology, including Boolean networks, Bayesian networks, kernel methods, dimension reduction and related techniques. Specifically, we will consider the following topics to investigate during this project.

Firstly, we plan to identify gene similarities based on their shared phenotypic features by the structure of association between genes and phenotypes. The reasoning behind this approach of analysis is that genes sharing more similar phenotypes shall have a stronger tendency for functional interactions in biology which will in turns provide the usefulness of phenotype similarity values in gene network analysis. Statistical measures of phenotypic association of genes will be investigated. The relationship of phenotypic association to the

pattern recognition of gene networks will be explored by statistical methods together with the integration of biological information.

Alternatively, phenotypes that are associated with common genes can be regarded as similar traits that are likely to be the outcomes of similar pathways consisted of these common genes from the perspective of biological systems. Hence, we plan to develop statistical measures for the similarity of phenotypes based on the association patterns of genotypes and phenotypes with the structure of gene networks behind the analysis steps. This will provide a systematic framework to analyze the interdependence of genotypes and phenotypes through biologic network structure from the perspective of systems biology with state-of-arts techniques in statistical analysis.

With these analysis measures and methods, we can perform disease-specific and tissue-specific network analysis for any combination of disease and tissue types. For example, we can use this analysis framework to assist the discovery and development in designing high efficacy and low side-effect drugs and treatments. Other applications are also possible and will be investigated. The statistical methods developed in this long term project will be very useful for the treatment of human diseases and analysis of biologic studies through the association of gene networks and trait phenotypes in biologic and medical researches during collaborative studies and follow-up investigation.

Keywords: statistical learning, Boolean networks, Bayesian networks, kernel methods, dimension reduction, high dimensional data, phenotypic association, gene network, pathway analysis, pattern recognition, systems biology, drug discovery, side-effect.

二、緣由與目的

There exist interdependence relationships between genotypes and phenotypes from the perspectives of biological networks and biochemical pathways. Diseases are resulted from several genes mutation or abnormal expression. The exploring of these relationships requires tremendous biologic and medical experiments which cost lots of money and time. With growing scale of biological and medical data by high throughput techniques, we want to develop a highly efficient method through the statistical analysis based on the association networks of genotypes and phenotypes. It will provide a systematic framework for investigating the interdependence of genes and phenotypic features. With the research and development of analysis methods in this long term project, it can also speed up the research speed of drug discovery and enhance the reliability of new drugs and therapies. This method will be of great use in many biological and medical studies. In this long term project, we plan to develop statistical learning methods for the network analysis of biological studies in yeast and medical investigation of human diseases with collaborators.

三、結果與討論

本三年期計畫在計畫期間己發表論文如下。

1. Emerson, J. J., Hsieh, L.-C., Sung, H.-M., Wang, T.-Y., Huang, C.-J., Lu, H. H.-S., Lu, M.-Y. J., Wu, S.-H., and Li, W. H., "Natural selection on cis and trans regulation in yeasts", Genome Research, 20, 826-836. 2010.
2. Deng, L.-Y. , Lu, H. H.-S., and Chen, T.-B., "64-Bit and 128-bit DX random number generators", Computing, 89, 1, 27-43. 2010.
3. Cheng, J. H., Wang, Y., Chen, P. Y., Chen, T.-B., Chen, C.-J., Li, G.-C., and Lu, H. H.-S., "Mine Barcode of Life: Information Visualization and Fusion for the Environment and Society" , International Journal of Systems and Synthetic Biology, 1(1), 63-70. 2010.
4. Lu, H. H.-S., and Wu, H. M., "Visualization, Screening, and Classification of Cell Cycle-Regulated Genes in Yeast", International Journal of Systems and Synthetic Biology, 1(2), 185-198, 2010.
5. Wang, H., Lu, H. H.-S., Chueh, T.-H., "Constructing Biological Pathways by a

Two-Step Counting Approach", PLoS ONE 6(6): e20074, 2011.

6. Deng, L.-Y. , Shiau, J.-J. H., and Lu, H. H.-S., "Large-order multiple recursive generators with modulus $2^{31}-1$", INFORMS Journal on Computing, Published online before print, October 17, 2011.

7. Deng, L.-Y. , Shiau, J.-J. H., and Lu, H. H.-S., "Efficient computer search of large-order multiple recursive pseudo-random number generators", Journal of Computational and Applied Mathematics, 236, 3228– 3237, 2012.

8. Chiang, S, Swamy, K. B., Hsu, T. W., Tsai, Z. T., Lu, H. H.-S., Wang, D., Tsai, H. K., "Analysis of the association between transcription factor binding site variants and distinct accompanying regulatory motifs in yeast", Gene. 2012 Jan 10;491(2):237-45. Epub 2011 Sep 16.

9. Chueh, T.-H., and Lu, H. H.-S., "Inference of Biological Pathway from Gene Expression Profiles by Time Delay Boolean Networks", PLoS ONE 7(8): e42095, 2012.

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. Aha, D. W., Kibler, D., Albert, M. K., 1991. Instanced-based learning algorithms. Machine Learning 6 (1), 37-66.
2. Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S., 1998. Identification of gene regulatory networks by strategic gene disruptions and gene overexpression. Proc. 9th ACM-SIAM Symp. Discrete Algorithms, 695-702.
3. Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S., 2003. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a Boolean model. Theoretical computer science 298 (1), 235-251.
4. Akutsu, T., Miyano, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. Proc. Pacific Symposium on Biocomputing, 17-28.
5. Brem, R. B., Yvert, G., Clinton, R., Kruglyak, L., 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752-755.
6. Burnham, K. P., Anderson, D. R., 1998. Model selection and inference. Springer, New York.
7. DeRisi, J. L., Iyer, V. R., Brown, P. O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278 (24), 680-686.
8. Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95, 14863-14868.
9. Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using bayesian networks to analyze expression data. J. Comp. Biol. 7, 601-620.
10. Gaffney, S., 2004. Probabilistic curve-aligned clustering and prediction with mixture models. Ph.D. thesis, Department of Computer Science in University of California, Irvine.
11. Gaffney, S., Robertson, A. W., Smyth, P., Camargo, S. J., Ghil, M., 2007. Probabilistic clustering of extratropical cyclones using regression mixture models. Climate Dynamics 29 (4), 423-440.
12. Gaffney, S., Smyth, P., 2004. Joint probabilistic curve clustering and alignment. Advances in Neural Information Processing Systems 17, 473-480.
13. Galindo, C. L., Gadl, A. A., Sha, J., Chopra, A. K., 2004. Microarray analysis of aeromonas hydrophila cytotoxic enterotoxin-treated murine primary macrophages. Infection and Immunity 72 (9), 5439-5445.
14. Gancedo, J. M., 1998. Yeast carbon catabolite repression. Microbiology and Molecular Biology Reviews 62 (2),

334-361.

15. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O., 2000. Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell 11, 4241-4257.
16. Heckerman, D., Geiger, D., Chickering, D. M., 1995. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 20, 197-243.
17. Hunt, E., Marin, J., Stone, P., 1966. Experiments in induction. Academic Press, New York.
18. Jensen, F., 2001. Bayesian networks and decision graphs. Springer, New York.
19. Jensen, F. V., 1996. An introduction to Bayesian networks. University College London Press, London.
20. John, G. H., Langley, P., 1995. Estimating continuous distributions in Bayesian classifiers. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, 338-345.
21. Kauffman, S. A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of theoretical biology 22 (3), 437-467.
22. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K., 2001. Improvements to platt's smo algorithm for svm classifier design. Neural Computation 13 (3), 637-649.
23. Kerr, M. K., 2003. Design considerations for efficient and effective microarray studies. Biometrics 59 (4), 822-828.
24. Kerr, M. K., Churchill, G. A., 2001. Experimental design for gene expression micorarrays. Biostatistics 2 (2), 183-201.
25. Lauritzen, S., Spiegelhalt, D., 1988. Local computations with probabilities on graphical structures and their application to expert systems. Journal of Royal Statistical Society Series B 50, 157-224.
26. Le Cessie, S., Van Houwelingen, J., 1992. Ridge estimators in logistic regression. Applied Statistics 41 (1), 191-201.
27. Li, L. M., Lu, H. S., 2005. Explore biological pathways from noisy array data by directed acyclic Boolean networks. Journal of Computational Biology 12 (2), 170- 185.
28. Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo.
29. Quinlan, J., 1993. C4.5: Programs for machine learning.
30. Schuller, H.-J., 2003. Transcriptional control of nonfermentative metabolism in the yeast saccharomyces cerevisiae. Current Genetics 43 (3), 139-160.
31. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of Cell 9, 3273-3297.
32. Vapnik, V. N., 1998. Statistical learning theory. Wiley, New York.
33. Jensen FV. Bayesian Networks and Decision Graphs. Springer. 2001.
34. Heckerman D. A tractable inference algorithm for diagnosing multiple diseases. Proceedings of UAI. 1989:174–181.
35. Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Comput. 1986 Sep-Oct;3(5):34-48.
36. Middleton B, Shwe MA, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. II. Evaluation of diagnostic performance. Methods Inf Med. 1991 Oct;30(4):256-67.
37. Neapolitan RE. Learning Bayesian Networks. Prentice Hall. 2004.
38. Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. Methods Inf Med. 1991 Oct;30(4):241-55.

附件二

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

□ V 達成目標

□ 未達成目標（請說明，以 100 字為限）

□ 實驗失敗

□ 因故實驗中斷

□ 其他原因

說明：

The goal of this project is complete and the research results are profound.

2. 研究成果在學術期刊發表或申請專利等情形：

論文：V 已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 □無

技轉：□已技轉 □洽談中 □無

其他：（以 100 字為限）

The research results are published by international journals that are important journals in this research area.

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

The associations between gene networks and phenotypic patterns are very important to understand the effects of biological networks and their functions. We have proposed new methods to perform statistical analyses. The applications are demonstrated in many biomedical studies. These results reveal that we can utilize these new methods to explore more deep structures and their functionalities. Consequently, the research results build a very good foundation for further investigations and broad applications to both statistics and biomedical communities.