行政院國家科學委員會專題研究計畫　期中進度報告

生物網路和調控路徑的統計分析(1/2)

計畫類別：個別型計畫
計畫編號：NSC94-2118-M-009-010-
執行期間：94 年 08 月 01 日至 95 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

處理方式：本計畫可公開查詢

中　華　民　國 95 年 5 月 15 日

行政院國家科學委員會專題研究計畫期中報告

生物網路和調控路徑的統計分析(1/2)

Statistical Analysis of Biological Networks and Pathways

計畫編號：NSC 94-2118-M-009-010
執行期限：94 年 8 月 1 日至 95 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

在後基因體時代，我們希望透過高通量的生物科技來探索生物的功能性研究。這些新技術提供的生物資料包括序列資料，蛋白質間交互作用，生物晶片資料如基因晶片、蛋白質晶片...等等。這些生物資料的快速累積對統計分析帶來巨大的挑戰。藉由這個長期計畫，我們打算從統計觀點針對生物網路和調控路徑的分析問題來發展方法和執行研究。

整體而言，我們將以基因體的大規模資料，結合布爾網路、貝氏網路、和微分方程模型的動態系統的方法，進行統計性質的研究。具體而論，我們將探討下列的問題。首先是針對不同類型的生物資訊發展出統計學模式以解釋生物的變異，實驗因子和隨機誤差。接著，統計量與估計方法將被應用來去除系統性和雜訊的干擾。我們因此能對生物元素之間的關聯與調控關係做更準確的估計。假設檢定或區間估計的方法可以進一步用來檢驗生物的假設，並且我們也將發展方法來控制偽錯誤和偽正確。我們最後將對生物的功能與結構關係進行更進一步的研究。

在這項計畫結束後，我們將應用不同類型的生物高通量資料進行生物網路和途徑的統計分析與重建。 並且結合不同的生物資料庫做完整的研究。這些新的統計方法將不僅能提供生物學數據方面的分析工具，而且將會為其他關於網路與途徑的科學研究提供一個新的觀點。

關鍵詞：生物網路, 調控路徑, 蛋白質間交互作用，基因晶片, 布爾網路, 貝氏網路, 微分方程模型的動態系統.

Abstract

In the era of post-genomics, it is highly important to explore the biological functions through high-throughput techniques. The biological data from these new techniques include sequence data, protein-protein interaction, biochip data of DNA microarray, protein array, and so forth. The fast accumulations of these biological data raise big challenges to statistical analysis. In this three-year project, we aim to develop methods and perform analysis of biological networks and pathways from statistical perspectives.

In particular, we will investigate the statistical properties of Boolean networks, Bayesian networks, and dynamic systems by differential equation models for the biological networks and pathways by biological data at the genomic scale. Specifically, we will investigate the following issues. The first issue is the development of statistical models for various biological data to account for biological variations, experimental factors, and random errors. Then, statistics and estimates will be proposed to remove the systematic and noise effects. We can therefore evaluate the association and regulation between biological elements more accurately. Hypothesis testing or confidence intervals will be constructed. The control of false negatives and positives will be implemented. The search of functional and structural relationships will be further studied.

At the end of this project, statistical methods for the analysis and reconstruction of biological networks and pathways will be developed for different types of biological data by high throughput techniques. The integration of biological databases for the

same and related techniques will be discussed as well. These new statistical methods will not only provide tools for analysis of biological data but also suggest new insights to other types of studies related to networks and pathways in scientific investigations.

Keywords: Biological Networks, Pathways, Protein-Protein Interaction, Microarray, Boolean Networks, Bayesian Networks, Dynamic Systems by Differential Equation Models.

二、緣由與目的

Currently, the studies in functional genomics and system biology related to biological networks and pathways are one of the most challenging research topics due to the fast accumulations of biological data by high throughput techniques, including sequence data, protein-protein interaction, biochip data of DNA microarray, protein array, and so on. There are three major types of approaches that could be applied for modeling biological networks and pathways (de Jong 2002): Boolean networks, Bayesian networks, and dynamic systems by differential equations. This project will investigate their statistical properties, improvements and integration.

三、結果與討論

本計畫到目前為止已完成 3 篇論文如下。

1. Statistical methods for identifying yeast cell cycle transcription factors. *PNAS (Proceedings of the National Academy of Sciences of the United States of America)*, 102, 38, 13532-13537 (http://www.pnas.org/cgi/reprint/102/38/13532).

Abstract:

Knowing transcription factors (TFs) involved in the yeast cell cycle is helpful for understanding the regulation of yeast cell cycle genes. We therefore developed two methods for predicting (*i*) individual cell cycle TFs and (*ii*) synergistic TF pairs. The essential idea is that genes regulated by a cell cycle TF should have higher (lower, if it is a repressor) expression levels than genes not regulated by it during one or more phases of the cell cycle. This idea can also be used to identify synergistic interactions of TFs. Applying our methods to chromatin immunoprecipitation data and microarray data, we predict 50 cell cycle TFs and 80 synergistic TF pairs, including most known cell cycle TFs and synergistic TF pairs. Using these and published results, we describe the behaviors of 50 known or inferred cell cycle TFs in each cell cycle phase in terms of activation/repression and potential positive/negative interactions between TFs. In addition to the cell cycle, our methods are also applicable to other functions.

2. Method for identifying transcription factor binding sites in yeast. *Bioinformatics,* Advance Access published online on April 27, 2006 (http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btl160v1).

Abstract:

Motivation: Identifying transcription factor binding sites (TFBSs) is helpful for understanding the mechanism of transcriptional regulation. The abundance and diversity of genomic data provide an excellent opportunity for identifying TFBSs. Developing methods to integrate various types of data has become a major trend in this pursuit.

Results: We develop a TFBS identification method, TFBSfinder, which utilizes several data sources, including DNA sequences, phylogenetic information, microarray data and ChIP-chip data. For a TF, TFBSfinder rigorously selects a set of reliable target genes and a set of non-target genes (as a background set) to find over-represented and

conserved motifs in target genes. A new metric for measuring the degree of conservation at a binding site across species and methods for clustering motifs and for inferring position weight matrices are proposed. For synthetic data and yeast cell cycle TFs, TFBSfinder identifies motifs that are highly similar to known consensuses. Moreover, TFBSfinder outperforms well-known methods.

3. Image Segmentation of cDNA Microarray Spots Using the Gaussian Mixture Model. Technical Report.

Abstract:

The segmentation of cDNA microarray spots is essential in analyzing the intensities of microarray images for biological and medical investigation. In this work, the Gaussian mixture model is applied to segment two-channel cDNA microarray images. This approach groups pixels into foreground and background. The segmentation performance of this model is tested and evaluated with reference to simulation and real microarray data. Additionally, the adaptive irregular segmentation approach employed in GenePix Pro 6.0 is compared. In particular, spike genes with various contents are spotted in a real microarray to examine and evaluate the accuracy of the segmentation results. Duplicated and swapped designs are also employed to elucidate the implementation and accuracy of the model. Results of this study demonstrate that this method can cluster pixels and estimate statistics regarding spots with high accuracy.

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. Akutsu T, Kuhara S, Maruyama O, Miyano S. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a Boolean model. *Theoretical Computer Science.* 2003;298:235-251.
2. Akutsu T, Miyano S, Kuhara S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol.* 2000;7(3-4):331-43.
3. Akutsu T, Miyano S, Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics.* 2000 Aug;16(8):727-34.
4. Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell*. 2004 Apr 16;117(2):185-98.
5. Chen HC, Lee HC, Lin TY, Chen BS. System Stabilization mechanism of small heat shock proteins. *Bioinformatics.* 2004;to appear.
6. Chen HC, Lee HC, Lin TY, Li WH, Chen BS. Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics*. 2004 Aug 12;20(12):1914-27.
7. Chen T, He HL, Church GM. Modeling gene expression with differential equations. *Pac Symp Biocomput*. 1999;:29-40.
8. Datta A, Choudhary A, Bittner ML, Dougherty ER. External Control in Markovian Genetic Regulatory Networks. *Machine Learning*. 2003;52(1-2):169-191.
9. Datta A, Choudhary A, Bittner ML, Dougherty ER. External control in Markovian genetic regulatory networks: the imperfect information case. *Bioinformatics*. 2004;20(6):924-930.
10. de Hoon MJ, Imoto S, Kobayashi K, Ogasawara N, Miyano S. Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations. *Pac Symp Biocomput.* 2003;:17-28.
11. de Jong H. Modeling and simulation of

genetic regulatory systems: a literature review. *J Comput Biol.* 2002;9(1):67-103.
12. Friedman N. Inferring cellular networks using probabilistic graphical models. *Science*. 2004 Feb 6;303(5659):799-805.
13. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol.* 2000;7(3-4):601-20.
14. Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER. Growing genetic regulatory networks from seed genes. *Bioinformatics*. 2004 May 22;20(8):1241-7.
15. Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J Bioinform Comput Biol*. 2004 Mar;2(1):77-98.
16. Kim S, Imoto S, Miyano S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*. 2004 Jul;75(1-3):57-65.
17. Kim S, Li H, Dougherty ER, Cao N, Chen T, Bittner M, Suh EB. Can Markov Chain Models Mimic Biological Regulation? *Journal of Biological Systems*. 2002;10(4):431-445.
18. Li L and Lu HHS. Explore Biological Pathways from Noisy Array Data by Directed Acyclic Boolean Networks. *J Comput Biol*. 2004;to appear.
19. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*. 1998;18-29.
20. Lu HHS, Wu HM. On visualization, screening, and classification of cell cycle-regulated genes in yeast. *The 14th International Conference on Genome Informatics (GIW2003)*. 2003;:344-345.
21. Nachman I, Regev A, Friedman N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*. 2004 Aug 4;20 Suppl 1:I248-I256.
22. Nariai N, Kim S, Imoto S, Miyano S. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pac Symp Biocomput*. 2004;:336-47.
23. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*. 2001;17 Suppl 1:S215-24.
24. Qin H, Lu HH, Wu WB, Li WH. Evolution of the yeast protein interaction network. *Proc Natl Acad Sci U S A*. 2003 Oct 28;100(22):12820-4.
25. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002 Feb;18(2):261-74.
26. Shmulevich I, Dougherty ER, Zhang W. Gene perturbation and intervention in probabilistic Boolean networks. *Bioinformatics*. 2002 Oct;18(10):1319-31.
27. Shmulevich I, Dougherty ER, Zhang W. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*. 2002;90(11):1778-1792.
28. Shmulevich I, Gluhovsky I, Hashimoto R, Dougherty ER, Zhang W. Steady-State Analysis of Genetic Regulatory Networks Modeled by Probabilistic Boolean Networks. *Comparative and Functional Genomics*. 2003;4(6):601-608.
29. Shmulevich I, Kauffman SA. Activities and Sensitivities in Boolean Network Models. *Physical Review Letters*. 2004;93(4):048701(1-4).
30. Shmulevich I, Lahdesmaki H, Dougherty ER, Astola J, Zhang W. The role of certain Post classes in Boolean network models of genetic networks. *Proc Natl Acad Sci U S A*. 2003 Sep 16;100(19):10734-9.
31. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*. 2003 Oct;19 Suppl 2:II227-II236.

附件：封面格式

行政院國家科學委員會補助專題研究計畫期中報告

※※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　　生物網路和調控路徑的統計分析(1/2)　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：x 個別型計畫　　□整合型計畫
計畫編號：NSC94-2118-M-009-010
執行期間：94 年 8 月 1 日至 95 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：

- □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　95 年　5 月　15 日