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(3/3)

計畫類別：個別型計畫
計畫編號：NSC93-2118-M-009-001-
執行期間：93 年 08 月 01 日至 94 年 10 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：完整報告
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 94 年 10 月 31 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(3/3)

Statistical Pattern Discovery for Microarray Data

計畫編號：NSC 93-2118-M-009-001

執行期限：93 年 8 月 1 日至 94 年 10 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

本三年期計畫到目前為止已完成9篇論文如下。

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

2. “On Visualization, Screening, and Classification of Cell Cycle-Regulated Genes in Yeast,” *The 14th International Conference on Genome Informatics (GIW2003)*, 344-345, 2003.

3. “Statistical Analysis of the Gene Expression for Non-synchronized Cell Cycles of Human Glioma Cells after Gamma Irradiation by cDNA Microarray,” Technical Report.

4. "Evolution of the yeast protein interaction network," *PNAS (Proceedings of the National Academy of Sciences of the United States of America)*, 100, 22, 12820-12824, 2003.

5. "Gene Expression Analysis Refining System (GEARS) via Statistical Approach: A Preliminary Report," *The 14th International Conference on Genome Informatics (GIW2003)*, 316-317, 2003.

6. "Supervised Motion Segmentation by Spatial-Frequential Analysis and Dynamic Sliced Inverse Regression," *Statistica Sinica*, 14, 413-430, 2004.

7. “Patterns of Segmental Duplications in the Human Genome,” *Mol. Biol. Evol.*, 22, 1, 135-141, 2005.

8. “Explore Biological Pathways from Noisy Array Data by Directed Acyclic Boolean Networks,” *Journal of Computational Biology*, 12, 2, 170-185, 2005.

9. “Gridding Spot Centers of Smoothly Distorted Microarray Images,” *IEEE Transactions on Image Processing*, accepted.

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.

Rapid divergence in expression between duplicate genes inferred from microarray data

Zhenglong Gu, Dan Nicolae, Henry H-S. Lu and Wen-Hsiung Li

For more than 30 years, expression divergence has been considered as 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 $K_A \leq 0.3$. Thus, expression divergence increases with evolutionary time, and K_A is initially coupled with expression divergence. 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, more than 40% of gene pairs show expression divergence even when K_S is ≤ 0.10, and this proportion becomes >80% for $K_S > 1.5$. Only a small fraction of ancient gene pairs do not show expression divergence.

Published online: 01 November 2002

Expression divergence between duplicate genes has long been a subject of great interest to geneticists and evolutionists [1–4]. Indeed, Ohno [2] and others [3,4] had proposed expression divergence as the first step towards the retention of duplicate genes. In the past, however, studies of expression divergence were usually conducted for a limited number of gene families, providing no general picture of the rate of expression divergence between duplicate genes in a genome. Fortunately, a general picture can now be seen thanks to the advent of microarray gene expression technology (Box 1) and the complete sequences of many genomes. Indeed, using the microarray technology, Ferea *et al.* [5] showed that rapid change in gene expression can occur in experimental lineages of yeast.

These advances notwithstanding, there remains the difficulty of dating the divergence time between two duplicate genes, which is needed for inferring the rate of expression divergence. In a pioneering study using microarray data from *Saccharomyces cerevisiae*, Wagner [6] found no significant correlation (–0.30, $P = 0.18$) between expression divergence and protein sequence divergence (d) between duplicate genes, and concluded that expression divergence and sequence divergence are decoupled. This result, however, does not imply that expression divergence and evolutionary time are decoupled because d might not be a good proxy of divergence time. Because the rate of amino acid substitution varies tremendously among proteins [7,8], no single d value can be applied to date the divergence times of different protein or gene pairs. By comparison, the rate of synonymous substitution is more uniform among genes [7,8], and so K_S is a better proxy of divergence time. We shall therefore rely more on K_S than d.

To avoid using correlated data points, we selected independent pairs of duplicate genes in the yeast genome (Box 2).

For each gene family, we started with the pair with the smallest K_S and continued selecting pairs with increasing K_S, because gene pairs with a small K_S are fewer than those with a large K_S and because a smaller K_S can more accurately reflect the time course of expression divergence. Moreover, we selected gene pairs where neither duplicate shows strong codon usage bias, because this bias can retard the increase of K_S so as to make K_S a poor proxy of divergence time. Then we analysed the expression divergence for each gene pair using expression data from microarray analyses (see Box 2).

Figure 1a shows a significant negative correlation (–0.47, $P < 2 \times 10^{-5}$) between $\ell n[(1+R)/(1-R)]$ and K_S. We used the transformation $\ell n[(1+R)/(1-R)]$ instead of R to change the scale to a more appropriate one for a linear regression analysis (Box 2); actually, a similar correlation (–0.54) is obtained between R and K_S. A stronger correlation than this is not expected because K_S is only a crude

Box 1. Yeast microarray data

A total of 208 cDNA microarray experiment data points were compiled for this study. The dataset represents the gene expression under various developmental and physiological conditions in the yeast life history (Table I).

For some processes, more than one yeast strain or one time course were studied and we randomly selected only one of them for each process. Log_2-transformed ratios of gene expression in experimental populations to reference populations were used in the analysis.

Table I. Studied processes and number of data points in each process

Process	Data points	Ref.
Sporulation	9	[a]
Cell cycle	17	[b]
Zinc regulation	9	[c]
YPD growth	10	[d]
Diamide treatment	8	[d]
Nitrogen deletion	10	[d]
DTT treatment	8	[d]
H2O2 treatment	10	[d]
Menadione treatment	9	[d]
Diauxic shift	7	[e]
Heat shock	7	[d]
Hyper-osmotic shock	7	[d]
Different carbon resources	6	[d]
Amino acid starvation	5	[d]
Other experiments in response to environmental changes	86	[d]

References

a Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705

b Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297

c Lyons, T.J. *et al.* (2000) Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7957–7962

d Gasch, A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257

e DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686

http://tig.trends.com 0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)02837-8

Box 2. Duplicate gene selection and linear regression analysis

Open reading frames in the yeast genome (SGD, http://genome-www.stanford.edu/Saccharomyces/) were grouped into different gene families using a rigorous method [a]. Protein sequences of duplicate genes were aligned using ClustalW [b] and the corresponding coding regions were then aligned based on the protein alignment. The numbers of substitutions per synonymous site (K_S) and per nonsynonymous (K_A) site between duplicate genes were estimated using PAML [c] with default parameters. We selected only gene pairs with $K_S \le 1.5$ because when K_S becomes larger it is difficult to obtain a reliable estimate, owing to repeated substitutions at the same site. Similarly, we restricted K_A to ≤ 0.70. The computer program CodonW (ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z) was used to calculate the effective number of codons (ENC) for each gene studied.

Duplicate gene pairs were selected as follows: within each gene family, starting from the pair with the smallest K_S of greater than 0.01, we selected independent gene pairs; that is, pairs that share no genes in common with other pairs. To avoid gene pairs with strong codon usage bias, both genes in a selected pair must have an ENC > 35. Our study [a] suggests that K_S is substantially reduced by codon usage bias when ENC < 32, but is only mildly affected when ENC > 35. In total, 400 duplicate gene pairs were selected.

Because all of the duplicate gene pairs encoding ribosomal proteins have strong codon usage bias, we consider the divergence in the flanking sequences instead of K_S. For each gene pair, the 200 bp of both upstream and downstream flanking regions of both genes were extracted from gene annotation data. ClustalW was used to do the alignment, followed by minor manual adjustments. Genetic distances were calculated using Tamura and Nei's six-parameter method [d]. The average of the genetic distances in upstream and downstream flanking regions is denoted as D_{flank} (Supplementary Table 2 at http://download.bmn.com/supp/tig/decemberTable2.pdf).

The Pearson correlation coefficient (R) of gene expression over all data points in Table I in Box 1 was calculated for each selected gene pair if the expression data were available for more than half of the experiments studied for that pair (396 pairs were calculated, Supplementary Table 3 at http://download.bmn.com/supp/tig/decemberTable3.pdf). Linear regression analysis was used to investigate the relationship between R and K_S (K_A). Because R is bounded by –1 and 1, the transformation $\ell n((1+R)/(1-R))$ was used and the normal linear regression was then carried out between each pair of K_S (K_A) and the transformed R. The statistical package of S+ was used.

Each of the first 9 processes listed in Table I of Box 1, each of which has eight or more data points, was also treated separately; for each process the Pearson correlation coefficient was calculated for each selected gene pair (Supplementary Table 3 at http://download.bmn.com/supp/tig/decemberTable3.pdf).

References

a Gu, Z. *et al.* (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* 19, 256–262

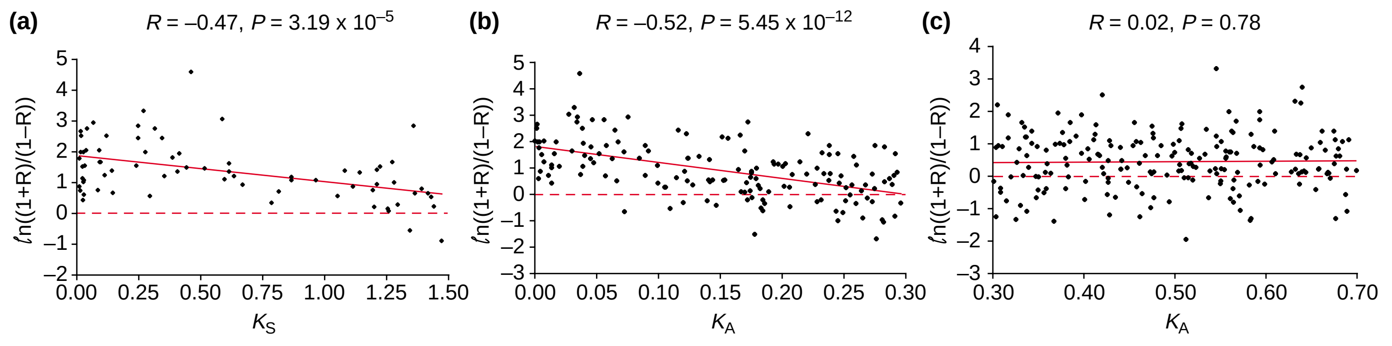
b Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

c Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43

d Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526

proxy of divergence time owing to the considerable variation in synonymous rate among genes [7,8]. As in [6], only a weak correlation (–0.30, $P = 4.57\times10^{-9}$) is found between $\ell n[(1+R)/(1-R)]$ and K_A ($K_A \le 0.70$); the correlation is significant because the dataset used is much larger than that in [6]. The weak correlation is not surprising because K_A is not a good proxy of divergence time, so that no correlation between R and K_A is expected when K_A becomes large. Indeed, Fig. 1c shows no correlation (0.02, $P = 0.77$) between $\ell n[(1+R)/(1-R)]$ and K_A for $K_A > 0.30$. However, a significant negative correlation (–0.52) between the two quantities is seen for $K_A \le 0.30$ (Fig. 1b). The range of $K_A \le 0.30$ is somewhat arbitrary, but the correlation coefficient varies only from –0.49 for $K_A \le 0.25$ to –0.48 for $K_A \le 0.35$. Thus, expression divergence and K_A are initially coupled to some extent. The same conclusions hold for Affymetrix microarray data, for which cross hybridization between duplicate genes is a less serious problem (see Supplementary Figure at http://download.bmn.com/supp/tig/decemberAffymetrix.pdf); the dataset is smaller than cDNA microarray data, so it was not used in the other analyses in this study.

In the above analysis, all experiments were considered together; that is, R was calculated over all data points. This pooling of data might obscure the relationship between expression divergence and sequence divergence because a pair of duplicate genes are not necessarily involved in all of the physiological processes tested. Note that if a gene pair is not involved in a process, it is unlikely to evolve expression divergence in that process. For this reason we now consider R separately for each of



TRENDS in Genetics

Fig. 1. Relationship between the correlation coefficient (R) of gene expression over all available data points and K_S (K_A) between duplicate genes. (a) A significant negative correlation between $\ell n[(1+R)/(1-R)]$ and K_S for gene pairs with $K_S < 1.5$. (b) A significant negative correlation between $\ell n[(1+R)/(1-R)]$ and K_A for gene pairs with $K_A \le 0.3$. (c) No correlation between $\ell n[(1+R)/(1-R)]$ and K_A for gene pairs with $K_A > 0.3$.

Box 3. Parametric bootstrap

For each process under study, denote the n pairs of observations on the expression levels of the two duplicate genes compared by $\boldsymbol{Z} = \{\mathbf{z}_i: i = 1, \ldots, n$, and $\mathbf{z}_I = (x_i, y_i)^t\}$. From the sample, the correlation coefficient (R) between x and y is calculated. We will assume that these n pairs of observations are independently, identically distributed as a bivariate normal distribution with a correlation coefficient (ρ) in the population. This assumption of normality has been checked by the Kolmogorov–Smirnov test on the Q–Q plot for $\tanh^{-1}(R) = \{\ell n[(1+R)/(1-R)]\}/2$ in every process (Supplementary Table 4 at http://download.bmn.com/supp/tig/decemberTable4.pdf).

With a large sample size n, the distribution of R can be approximated as follows. We transform R and ρ to $\tanh^{-1}(R) = \{\ell n[(1+R)/(1-R)]\}/2$ and $\tanh^{-1}(\rho) = \{\ell n[(1+\rho)/(1-\rho)]\}/2$. Then, the difference $\tanh^{-1}(R) - \tanh^{-1}(\rho)$ is approximately a normal variate with the following mean and variance (Ref. [a] p. 433):

$$\text{mean} = \mu = \frac{\rho}{2(n-1)},$$

$$\text{variance} = \sigma^2 = \frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2} \approx \frac{1}{n-3}$$

Using this normal approximation, we can evaluate various probabilities. For example, for $-1 \le c \le 1$, we can compute

$$P(c \mid \rho,n\} = P\{R \le c \mid \rho,n\} = P\{\tanh^{-1}(R) \le \tanh^{-1}(c) \mid \rho,\mathrm{n}\}$$

$$= P\{[\tanh^{-1}(R) - \tanh^{-1}(\rho) - u] / \sigma \le [\tanh^{-1}(c) - \tanh^{-1}(\rho) - u] / \sigma \mid \rho,n\}$$

$$\approx P\{Z \le [\tanh^{-1}(c) - \tanh^{-1}(\rho) - \mathrm{u}] / \sigma\}$$

where $\boldsymbol{Z}$ has a standard normal distribution, which can be easily evaluated.

For a small n, the parametric bootstrap can be used to find out the distribution of R [b]. The mean and variance in the population are estimated by the mean and variance in the sample, which are denoted as $\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$ and $\begin{pmatrix} S_x^2 & RS_xS_y \\ RS_xS_y & S_y^2 \end{pmatrix}$.

Given the population correlation coefficient ρ, a bootstrap sample, $\boldsymbol{Z}^* = \{\mathbf{z}^*_i: i = 1, \ldots, n\}$, is obtained by simulating a bivariate normal distribution with $\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$ and $\begin{pmatrix} S_x^2 & \rho S_xS_y \\ \rho S_xS_y & S_y^2 \end{pmatrix}$.

The correlation coefficient from the bootstrap sample $\boldsymbol{Z}^*$ is computed and denoted as R^*. Repeating the resampling procedure B times, we observe $R^*_1, \ldots, R^*_B$. The empirical distribution of $R^*_1, \ldots, R^*_B$ is used to approximate the distribution of R. In particular,

$$P(c \mid \rho,n) = P(R \le c \mid \rho,n) \approx \sum_{i=1}^{B} \mathrm{I}\{R_i^* \le c\} / \mathrm{B},$$

where $\mathbf{I}\{\cdot\}$ is a indicator function whose value is 1 when the event is true and 0 otherwise. Because the data contain small sample sizes, we will use this parametric bootstrap to estimate probabilities.

Now suppose that m processes are studied and there are n_j pairs of observations for each process, $j = 1, \ldots, m$. From the above approximation, we can evaluate the probability of $P_j(c) = P(c \mid \rho,n_j\}$. Then, we can find out the probability that there are κ R values observed among the m processes that are $\le c$:

$$P\{\text{no } R \le c \mid \rho,m\} = \prod_{j=1}^{m} [1 - P_j(c)],$$

$$P\{\text{only one } R \le c \mid \rho,m\} = \sum_{j=1}^{m} P_j(c) \prod_{\substack{k=1 \\ k \ne j}}^{m} [1 - P_k(c)] = \sum_{j=1}^{m} \frac{P_j(c)}{1 - P_j(c)} \prod_{k=1}^{m} [1 - P_k(c)],$$

$$P\{\text{at least two } R \text{ values} \le c \mid \rho,m\} = 1 - P\{\text{no } R \le c \mid \rho,m\} -$$

$$P\{\text{only one } R \le c \mid \rho,m\} = 1 - \prod_{j=1}^{m} [1 - P_j(c)] - \sum_{j=1}^{m} \frac{P_j(c)}{1 - P_j(c)} \prod_{k=1}^{m} [1 - P_k(c)] \quad \text{Eqn [1]}$$

and so forth.

Once we observe the sample correlation coefficients (R values) of one gene pair in the m processes, we can use this parametric bootstrap to evaluate the probability of observing the smallest R values given the population correlation coefficient (ρ). For example, let the smallest two R values be c_1 and c_2 with $c_1 \ge c_2$ Then, we can replace c by c_1 in Eqn [1]. Of course, by using the complete information of c_1 and c_2, we can obtain a more precise probability:

$$P\{\text{at least one } R \le c_1 \text{ and one } \mathrm{R} \le c_2 \mid \rho,m\}$$

$$= 1 - P\{\text{no } R \le c_2 \mid \rho,m\} - P\{\text{only one } R \le c_2 \text{ and all other } R \text{ values} > c_1 \mid \rho,m\}$$

$$= 1 - \prod_{j=1}^{m} [1 - P_j(c_2)] - \sum_{j=1}^{m} \frac{P_j(c_2)}{1 - P_j(c_1)} \prod_{k=1}^{m} [1 - P_k(c_1)] \quad \text{Eqn [2]}$$

Note that Eqn [2] is always smaller than or equal to Eqn [1] with $c = c_1$. All the probability computations in this paper were obtained using Eqn [2].

References

a Rao, C.R. (1973). *Linear Statistical Inference and Its Application* (2nd Edn), Wiley

b Efron, B. and Tibshirani, R.J. (1998). *An introduction to the Bootstrap*, Chapman & Hall/CRC

the first nine tests in Box 1, each of which has eight or more time points.

To define 'expression divergence', we note that the correlation coefficient between two duplicate genes is initially 1, so we consider a value of 0.5 as sufficiently low. Note that for $R = 0.5$, R^2 is only 0.25, so that knowing the pattern of expression of one gene provides little information for predicting the expression pattern of the other gene. More importantly, we actually define 'expression divergence' by requiring that the probability of observing the two smallest R values among the nine processes is <0.05, given that the population (true) correlation coefficient (ρ) is 0.5; see Box 3 for the test method. This definition is likely to underestimate the true degree of divergence because it uses only the information of two smallest R values in the observed R values and because it assumes that the gene pair is involved in all of the nine processes studied. Indeed, this definition is stringent because, in effect, it requires at least one or two negative R values among the nine processes (Table 1). For example, only 38% of the cases with one negative R show 'expression divergence'. Moreover, none of the 54 pairs of duplicated ribosomal protein genes in the yeast genome is 'divergent' under this criterion (data not shown).

Table 2 shows that over 40% of the non-ribosomal protein gene pairs studied show divergent expression even when $K_S \le 0.10$ and the proportion becomes >80% when K_S becomes larger than 1.5. The proportion of pairs with diverged expression increases even more rapidly with K_A (Table 2). Clearly, expression divergence has occurred quickly in many of the gene pairs studied.

If we relax the definition of 'divergent expression' by setting $\rho = 0.6$ instead of 0.5, the proportion of pairs with divergent expression increases with K_S at an even faster rate (Table 2). Indeed, more than 50% of the pairs studied show divergent expression even when K_S is ~0.10. The synonymous rate is not known in yeast but is probably higher than that in *Drosophila*, which has been commonly taken as 15.6×10^{-9} nucleotide substitutions per site per year [7]. Thus, $K_S = 0.1$ would correspond to less than 3.2 million years of divergence time, implying a rapid rate of

Table 1. Numbers and proportions of gene pairs with expression divergence (i.e. $P < 0.05$) for different numbers of negative R values in the nine processes studied.

Number of R values	Number of gene pairs	Gene pairs with $P < 0.05$[a]		% Gene pairs with $P < 0.05$[a]	
		$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.5$	$\rho = 0.6$
0	43	0	0	0	0
1	66	25	49	38%	74%
2	70	61	70	87%	100%
≥3	217	217	217	100%	100%

[a] The ρ value is the criterion for 'expression divergence'.

expression divergence between duplicate genes in yeast. A similar picture is seen for K_A (Table 2).

There are two factors that tend to underestimate the rate of expression divergence. First, the nine processes studied do not represent all the physiological processes in yeast, and a duplicate gene pair could have diverged in one or more of the processes that have not been studied, although it has not diverged in any of the nine processes tested. This factor is likely to have significantly reduced our estimate of the rate of expression divergence. Second, there is the possibility of cross-hybridization of cDNA probes when two duplicate genes are highly similar in their cDNA sequences. In view of the fact that many of the highly similar duplicate pairs ($K_S < 0.10$) have shown one or more small R values (data not shown), the extent of cross-hybridization was probably not serious. However, if it were not negligible, the initial rate of expression divergence would have been underestimated.

Alternatively, the noisiness of microarray data tends to reduce the true correlation (R) between the expression levels of duplicate genes and thus tends to overestimate the rate of expression divergence, especially in the early stage of divergence between duplicate genes. Thus, although our definition of expression divergence seems stringent for the case of $\rho = 0.5$, the conclusion should be taken with caution.

Table 2. Proportion of gene pairs with expression divergence[a] in different K_S and K_A intervals.

ρ			K_S Intervals		
	0.01–0.1	0.1–0.3	0.3–1.0	1.0–1.5	>1.5
0.5	0.43	0.55	0.50	0.77	0.81
0.6	0.52	0.55	0.70	0.86	0.89
			K_A Intervals		
	0–0.05	0.05–0.1	0.1–0.25	0.25–0.5	>0.5
0.5	0.45	0.53	0.81	0.85	0.76
0.6	0.55	0.71	0.89	0.92	0.85
	D_{flank} Intervals		(Ribosomal protein genes)		
	0–0.1	0.1–0.6	0.6–1.0	1.0–1.5	>1.5
0.5	NA[b]	NA	0	0	NA
0.6	NA	NA	0.02	0.25	NA

[a] The criterion for expression divergence is that the probability of observing the two smallest R values in the nine tests studied is less than 0.05, given the population correlation coefficient is ρ.
[b] NA = not applicable.

It is worth noting that a divergent duplicate pair that has a large K_S or K_A might already have gained expression divergence when its K_S or K_A was still small. Thus, a divergent pair with a large K_S or K_A does not imply a slow rate of expression divergence. It is also interesting to note from Table 2 that the proportion of divergent duplicate gene pairs eventually becomes more than 80% as K_S increases. As noted, we have considered only nine processes. If many more processes are considered, the vast majority of duplicate genes will probably eventually become diverged in expression.

There are, however, duplicate genes that do not show divergent expression even when K_S is large; for example, genes encoding proteasome components, aminopeptidases, aldo/keto reductases and ribosomal proteins. Ribosomal protein genes have not been included in Fig. 1 and Table 1, and have been treated separately in Table 2, because they have strong codon usage bias and their K_S does not reflect the divergence time well. We therefore consider instead the sequence divergence (D_{flank}) in their flanking regions (Box 2). Note that none of the ribosomal protein gene pairs shows expression divergence under the condition of $\rho = 0.5$ (Table 2). Even under the condition of $\rho = 0.6$, their rate of expression divergence is very slow, compared with that for genes encoding non-ribosomal proteins.

We have examined the functions of quickly diverged gene pairs, that is, those pairs that have a $K_S < 0.3$ but show expression divergence (Supplementary Table 1 at http://download.bmn.com/supp/tig/decemberTable1.pdf). The functions of many of these genes are still unknown or have not been well studied. However, we can see that these genes include many membrane proteins such as substrate transporters, and many enzymes such as aldehyde hydrogenase, aldo/keto reductase, helicase and phosphopyruvate hydratase.

In conclusion, because protein distance (or K_A) is not a good measure of divergence time, it was not surprising that no coupling of expression divergence and protein distance was found previously. However, an initial coupling of expression divergence and K_A does exist (Fig. 1b). K_S is a better measure of divergence time than K_A, and the significant correlation of expression divergence with K_S suggests that expression divergence increases with divergence time. Most interestingly, many duplicate genes in yeast have diverged quickly in expression and the vast majority of duplicate genes will eventually become diverged in expression. However, the rate of expression divergence varies among duplicate genes. The majority of duplicate genes such as many membrane proteins and many enzymes have diverged quickly in expression, whereas ribosomal proteins, proteasome components and some other proteins show a slow rate of expression divergence. Other duplicate genes show a moderate rate of expression divergence. Clearly, a proper analysis of microarray data can shed much light on the rate and mode of expression divergence of duplicate genes.

Acknowledgements

We thank Z. Zhu, T. Oakley, M. Long, C-C. Shih, H. Kaessmann, K. Makova, and L. Mets for help and comments. This study was supported by NIH grants.

References

1 Markert, C.L. (1964) Cellular differentiation – an expression of differential gene function. In *Congenital Malformations*, pp163–174, International Medical Congress

2 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag

3 Ferris, S.D. and Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* 12, 267–317

4 Force, A. *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545

5 Ferea, T.L. *et al.* (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9721–9726

6 Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6579–6584
7 Li, W-H. (1997) *Molecular Evolution*, Sinauer Associates
8 Makalowski, W. and Boguski, M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9407–9412

Zhenglong Gu
Wen-Hsiung Li*
Dept of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA.
*e-mail: whli@uchicago.edu

Dan Nicolae
Dept of Statistics, University of Chicago, 5734 S. University Ave, Chicago, IL 60637, USA.

Henry H-S. Lu
Insitute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Rd, Hsingchu, 30050 Taiwan.

Techniques & Applications

Web-based primer design for single nucleotide polymorphism analysis

Michael M. Neff, Edward Turk and Michael Kalishman

The detection of single nucleotide polymorphisms by PCR is necessary for many types of genetic analysis, from mapping genomes to tracking specific mutations. This technique is most commonly used when polymorphisms alter restriction endonuclease recognition sites. Here we describe a web-based program, dCAPS Finder 2.0, that facilitates the design of mismatched PCR primers to create or remove a restriction endonuclease recognition site relative to the polymorphism being analyzed.

Published online: 01 November 2002

Molecular genetic research relies heavily on the ability to detect polymorphisms in DNA. These molecular markers range from large deletions and rearrangements to single nucleotide polymorphisms (SNPs) [1]. Before the advent of polymerase chain reaction (PCR) technology [2], restriction fragment length polymorphism (RFLP) analysis required Southern blots of restricted genomic DNA [3]. PCR technology has led to a more rapid, less expensive version of RFLP analysis using cleaved amplified polymorphic sequence (CAPS) markers [4]. However, both RFLP and CAPS analysis require that the SNP creates or removes a restriction endonuclease recognition site. Because this is not always the case, a variety of techniques have been developed to genotype SNPs in an enzyme-independent manner [1]. Many of these techniques require specialized detection equipment and/or labeled PCR primers that cost more than standard primers. Derived cleaved amplified polymorphic sequence (dCAPS) analysis, widely used in the plant molecular genetics community, uses mismatches in one of the two PCR primers flanking the SNP to create or remove a restriction endonuclease recognition site in one of the two haplotypes being assayed [5,6] (Fig. 1). In this paper, we present a web-based program, dCAPS Finder 2.0, that facilitates the design of these dCAPS primers.

dCAPS Finder 2.0

The dCAPS marker technique was originally developed as a method for changing a SNP into an RFLP (see [5,6] and references within) (Fig. 1). The technique can also be used to modify an existing RFLP such that a less expensive restriction endonuclease can be used for SNP analysis. Because dCAPS primers use the same chemistry as regular PCR primers, there is also a cost advantage of this technique over more sophisticated, enzyme-independent methods of SNP analysis. The biggest difficulty for designing dCAPS primers lies in identifying restriction endonuclease recognition sites and accompanying primer mismatches. To facilitate this technique, a Macintosh-based computer

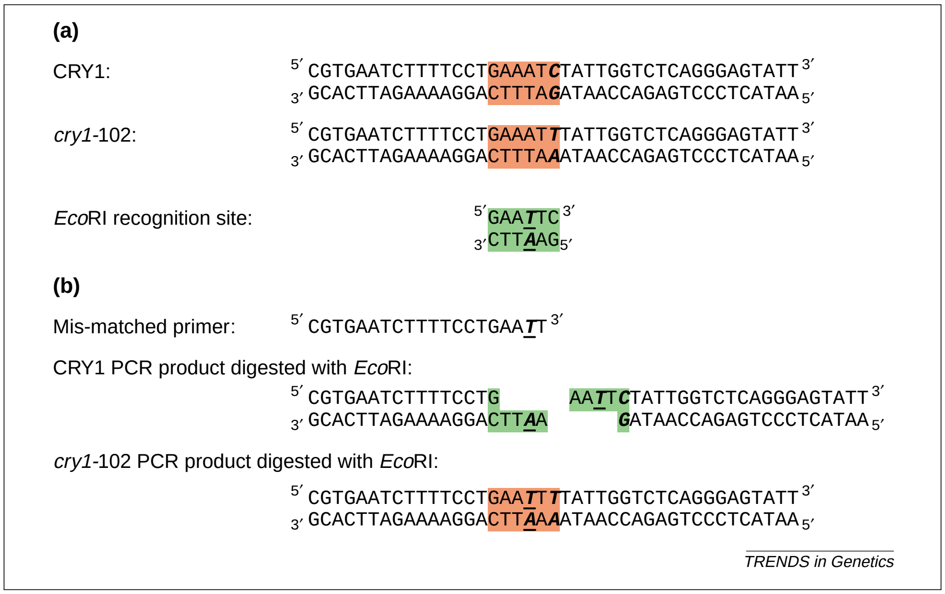

(a)
CRY1:
5′ CGTGAATCTTTTCCTGAAATCTATTGGTCTCAGGGAGTATT 3′
3′ GCACTTAGAAAAGGACTTTAGATAACCAGAGTCCCTCATAA 5′
cry1-102:
5′ CGTGAATCTTTTCCTGAAATTTATTGGTCTCAGGGAGTATT 3′
3′ GCACTTAGAAAAGGACTTTAAATAACCAGAGTCCCTCATAA 5′
EcoRI recognition site:
5′ GAATTC 3′
3′ CTTAAG 5′
(b)
Mis-matched primer:
5′ CGTGAATCTTTTCCTGAATT 3′
CRY1 PCR product digested with EcoRI:
5′ CGTGAATCTTTTCCTG AATTCTATTGGTCTCAGGGAGTATT 3′
3′ GCACTTAGAAAAGGACTTAA GATAACCAGAGTCCCTCATAA 5′
cry1-102 PCR product digested with EcoRI:
5′ CGTGAATCTTTTCCTGAATTTATTGGTCTCAGGGAGTATT 3′
3′ GCACTTAGAAAAGGACTTAAATAACCAGAGTCCCTCATAA 5′
TRENDS in Genetics

Fig. 1. Derived cleaved amplified polymorphic sequence (dCAPS) analysis uses a mismatched PCR primer to create a restriction fragment length polymorphism (RFLP) based on the single nucleotide polymorphism (SNP) being analyzed. (a) The *cry1-102* SNP (bold, italic letters) does not create an *Eco*RI-based RFLP because of one mismatch in the *Eco*RI recognition site (bold, underlined letters). (b) A primer containing this mismatch (bold, underlined letter) allows the amplification of PCR products that generate an *Eco*RI -based RFLP that is dependent on the *cry1-102* SNP. Red boxes show sequences that are not cleaved by *Eco*RI. Green boxes represent sequences that are cleaved by *Eco*RI.

0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)02820-2

Evolution of the yeast protein interaction network

Hong Qin[†], Henry H. S. Lu[‡], Wei B. Wu[§], and Wen-Hsiung Li[†¶]

[†]Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637; [‡]Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30050, Republic of China; and [§]Department of Statistics, University of Chicago, 5734 South University Avenue, Chicago, IL 60637

Contributed by Wen-Hsiung Li, August 29, 2003

To study the evolution of the yeast protein interaction network, we first classified yeast proteins by their evolutionary histories into isotemporal categories, then analyzed the interaction tendencies within and between the categories, and finally reconstructed the main growth path. We found that two proteins tend to interact with each other if they are in the same or similar categories, but tended to avoid each other otherwise, and that network evolution mirrors the universal tree of life. These observations suggest synergistic selection during network evolution and provide insights into the hierarchical modularity of cellular networks.

Biological networks are the basis of cellular functions (1, 2). Understanding network evolution may shed light on the hierarchical modularity, scale-free property, and various uses of the building blocks of biological networks (3–12). The yeast protein interaction network is one of the best annotated complex networks to date (13–17). Previous studies on the evolution of this network focused either on gene duplication and molecular evolution at the protein level (9, 10) or on the global statistical properties (12). Neither approach can delineate the network evolutionary path, and there is no other comparable protein interaction data for the system-level comparison approach (5). Therefore, uncovering the growth patterns and the evolutionary path of the protein interaction network is a serious challenge (3, 4, 6, 7, 9, 12).

Parts of the present yeast protein interaction network would have been inherited from the last common ancestor of the three domains of life: Eubacteria, Archaea, and Eukaryotes. Thus, an analysis of the evolution of the yeast protein interaction network may provide new insights into the origin of eukaryotic cells (18–21), which has been a controversial issue.

A key question in the evolution of biological complexity (6, 7, 9, 12, 21, 22) is, how have integrated biological systems evolved? Darwinists (21, 23) proposed natural selection as the driving force of evolution. However, the striking similarities between biological and nonbiological complexities have led to the argument that a set of universal (or ahistorical) rules account for the formation of all complexities (22, 24, 25). The yeast protein interaction network is an example of a complex biological system and contributes to the complexity at the cellular level (26). By analyzing the growth pattern and reconstructing the evolutionary path of the yeast protein interaction network, we can address whether or not network growth is contingent on evolutionary history, which is the key disagreement between the Darwinian view and the universality view (22, 23, 27).

In this article, we studied how the yeast protein interaction network has evolved. We used graph theory to model the yeast protein interaction network. Each yeast protein is a node in the graph. Each pairwise interaction is a link between two nodes. Evolution of the yeast protein interaction network can then be inferred by analyzing the growth pattern of the graph. We classified all of the nodes (proteins) into isotemporal categories based on each protein's orthologous hits in several groups of genomes that are informative for yeast's evolutionary history. This scheme gives each protein a binary (b) value representing its evolutionary history. Proteins from the same isotemporal category share similar evolutionary histories. We then analyzed the interaction patterns within and between these isotemporal categories. Finally, we inferred the main path of the network evolution from six major isotemporal categories.

Materials and Methods

Data Collection. Genomic information of *Saccharomyces cerevisiae* was downloaded from the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/pub/yeast/data_download) on August 13, 2002. Protein interaction data were obtained from the Comprehensive Yeast Genome Database at the Munich Information Center for Protein Sequences (MIPS) (http://mips.gsf.de/proj/yeast/CYGD/db/index.html) (28, 29) on May 28, 2002, and from the reliable subsets of data from high-throughput screens (30). We excluded self-interactions and those involving mitochondrion proteins. The combined data set contains 6,633 interaction pairs. Orthologous analyses of the annotated ORFs in the yeast genome were parsed out from the clusters of orthologous groups (COGs) of proteins (ftp://ftp.ncbi.nih.gov/pub/COG) (31, 32) and the published orthologous analysis from the Bork group at the European Molecular Biology Laboratory (EMBL) (30). Mitochondrion genes and a few inconsistent orthologous assignments were removed from the analysis.

Data Analysis. Protein interaction networks were treated as undirected graphs in adjacency list format (33). Permutations of the networks were carried out in the Chiba City Linux cluster in the Mathematics and Computer Science Division of Argonne National Laboratory (www.mcs.anl.gov/chiba). Presentation of the network was performed by the program PAJEK (http://vlado.fmf.uni-lj.si/pub/networks/pajek) (34). Distance matrix-based analyses were conducted in the R environment for statistical computing and graphics (www.r-project.org) (35). The neighbor-joining (NJ) tree was generated by PAUP* (http://paup.csit.fsu.edu) and presented by the program TREEVIEW (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) (36).

Statistical Analysis of Interaction and Traversal Patterns. To evaluate the interaction tendencies within and between isotemporal categories, we measured the deviation of each observed interaction frequency from its random expectation (37). The observed interaction frequency between categories i and j, $F^{\text{obs}}_{(i,j)}$, is compared with the mean interaction frequency, $F^{\text{mean}}_{(i,j)}$, of a series of null models in which all proteins have the same connectivities, but their interaction partners are randomly chosen (37) [termed the Maslov–Sneppen 2002 (MS02) null models]. To describe the deviations of the observed interaction frequencies from the random expectations, we used Z scores, $Z_{(i,j)} = [F^{\text{obs}}_{(i,j)} - F^{\text{mean}}_{(i,j)}]/\sigma_{(i,j)}$, where $\sigma_{(i,j)}$ is the SD of the interaction frequency between categories i and j in the MS02 null models.

Similarly, we used the Z scores to measure the deviation of the average shortest path between two isotemporal categories from the mean of a series of isomorphic MS02 null models. This

Abbreviations: MS02 null model, Maslov-Sneppen 2002 null model; NJ, neighbor-joining; b, binary; d, decimal.

[¶]To whom correspondence should be addressed. E-mail: whli@uchicago.edu.

isomorphic MS02 null model retained the same topology with the original network. Network topology can greatly influence the average shortest path. The MS02 null model could change the total number of connected components in the original network and gave uninterpretable Z scores. The isomorphic null model was a simple method to exclude the topological influence on traversal path, and it enabled us to evaluate the association significance between two isotemporal categories.

Network Null Models. To generate an MS02 null model (37), the original network was first converted to pairwise-interacting nodes. These pairwise interacting nodes were then converted into an array of symbols. Permutation of this array of symbols was then used to generate a new list of pairwise-interacting nodes (self-pairing was prohibited during the permutation), which was then used to generate an MS02 null model in adjacency list format.

To generate an isomorphic MS02 null model, nodes with the same connectivity were concatenated into arrays of symbols. Permutation was then conducted on the arrays of symbols for each connectivity value. The original and the permutated arrays of symbols were then used to generate a lookup table in which each original node corresponded to a new node with the same connectivity. Based on this lookup table, all of the nodes in the original network were then replaced by the new nodes, resulting in a permutated network with the same topology.

Calculation of Average Shortest Path. We slightly modified the Dijkstra's algorithm to compute the shortest path (33). For a protein in isotemporal category i, its shortest path to isotemporal category j is defined as its traversal distance to the nearest neighbor in category j. The mean of the shortest paths to category j of all proteins in category i is taken as the distance from i to j, denoted as $d_{i \to j}$. Distance from j to i, $d_{j \to i}$, is calculated similarly. The average shortest path between categories i and j is the average of $d_{i \to j}$ and $d_{j \to i}$.

Results

Isotemporal Classification of Proteins. To study the growth of the yeast protein interaction network, we classified all yeast proteins into isotemporal categories, based on the presence or absence of their orthologous hits in each of the six groups of the universal tree of life (38), namely hyperthermophilic eubacteria, other eubacteria (excluding the hyperthermophiles), euryarchaeota, crenarchaeota, fungi, and other eukaryotes (excluding fungi) (Fig. 1). The first four groups are evolutionary pivotal groups (19). The hyperthermophilic eubacteria and other eubacteria may reflect one of the earliest splits in the eubacterial domain (38–41). Likewise, crenarcheota and euryarchaeota represent an early split in the archaeal domain (19, 38). We separated the fungal genomes from other eukaryotes because they may reveal recent evolutionary changes of yeast. For the purpose of orthologous analysis, the yeast genome is excluded from the groups of fungi and other eukaryotes. We parsed out the orthologous hits from the COGs (31) and another published orthologous analysis (30). Because the proteins in each category share the same or similar evolutionary histories, these categories might have been added to the yeast genome at various temporal intervals during evolution, and can be considered as isotemporal categories.

We designed a b coding scheme to represent the isotemporal categories (Fig. 1). The bits of the b coding scheme correspond to the six chosen evolutionary groups. For each yeast protein under study, the presence or absence of at least one orthologous hit in the genomes of each evolutionary group is represented by "1" or "0." Mathematically, this six-bit coding scheme gives 64 categories, but the yeast genome contains 42 categories with nonrandom distributions because of evolutionary constraints (see Fig. 4, which is published as supporting information on the PNAS web site, www.pnas.org). For presentation convenience, we used both b codes and their decimal (d) values. For example, category b000011 is equivalent to category d3, which contains proteins whose orthologs are found in the groups of fungi and other eukaryotes.

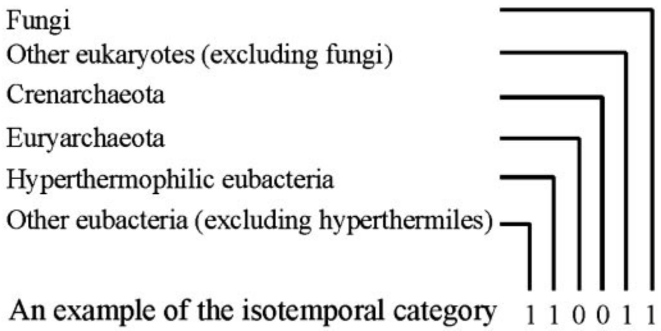

Fungi
Other eukaryotes (excluding fungi)
Crenarchaeota
Euryarchaeota
Hyperthermophilic eubacteria
Other eubacteria (excluding hyperthermiles)
An example of the isotemporal category
1 1 0 0 1 1

Fig. 1. Isotemporal classification of the yeast proteins. Isotemporal categories are designed through a binary (b) coding scheme. The b code represents the distribution of each yeast protein's orthologs in the universal tree of life. Bit value 1 indicates the presence of at least one orthologous hit for a yeast protein in a corresponding group of genomes, and bit value 0 indicates the absence of any orthologous hit. The presented example is 110011 in the b format and 51 in the d format. Orthologous identifications are based on COGs at the National Center for Biotechnology Information (31) and the results of the Bork group at the EMBL (30).

Interaction Patterns in the Network. We constructed a credible protein interaction network by using the manually curated protein interaction pairs maintained at MIPS (28) and the reliable subsets of data from high-throughput screens (30). The generated protein interaction networks are treated as undirected graphs. We excluded all self-interactions because we analyzed the network growth from the perspective of node additions. For simplicity, we also excluded the mitochondrion-coded proteins. The generated network contains only 39 isotemporal categories, with a biased coverage favoring the well conserved proteins in categories b000011 and b111111 (see Fig. 5, which is published as supporting information on the PNAS web site). This bias may reflect the assumption that conserved proteins are functionally more important than nonconserved ones, and the former deserve more experimental effort (37). In addition, interactions between well conserved proteins can be confirmed by their orthologs in other species (30).

We used Z scores to evaluate the interaction significance within and between isotemporal categories, based on the MS02 null models (Fig. 2*a*). Positive Z scores indicate that observed interactions are more frequent than random expectations; negative Z scores indicate the opposite. Therefore, large positive Z scores indicate strong interaction tendencies, whereas large negative Z scores indicate that proteins in the two categories tend to avoid each other in the network. Because the protein interaction network is treated as an undirected graph, the matrix presentation of the Z scores of all categories is symmetric. The diagonal distribution of large positive Z scores indicates that yeast proteins tend to interact with proteins from the same or closely related isotemporal categories. The observed intracategory association tendencies are consistent with the intuitive notion that a new function likely requires a group of new proteins, and that the growth of the protein interaction network is under functional constraints. For example, category b000011 (d3) contains the eukaryote-conserved nodes with intracategory interaction tendency, $Z_{(3,\ 3)} = 7.1$, indicating that nodes added

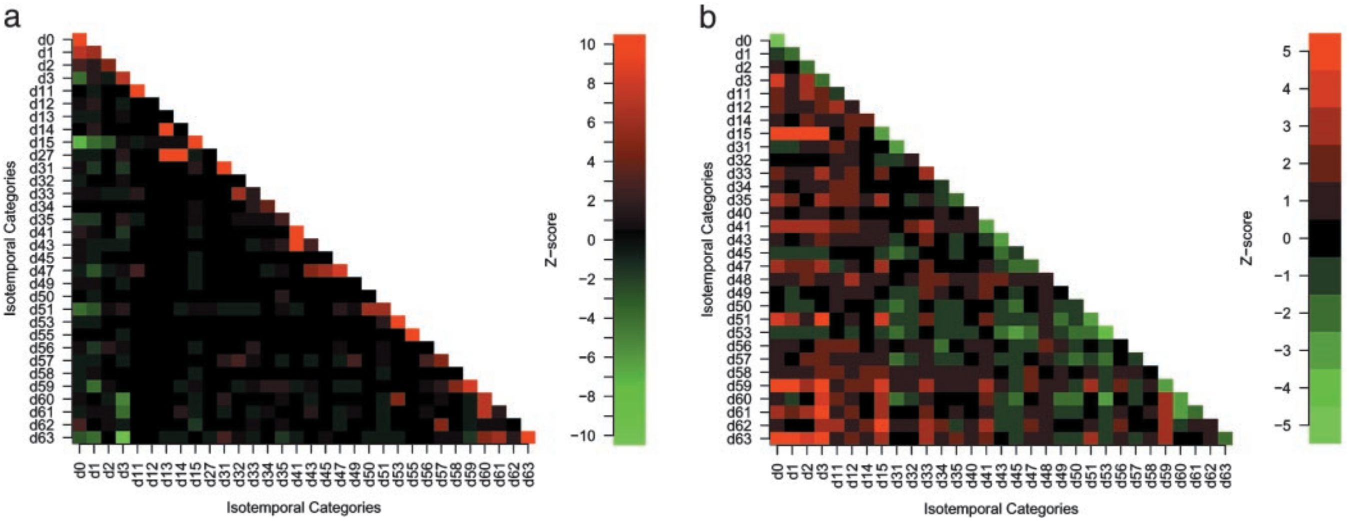


Fig. 2. Interaction patterns. (*a*) *Z* scores for all possible interactions of the isotemporal categories in the protein interaction network. For categories *i* and *j*, $Z_{(i,j)} = [F^{\text{obs}}_{(i,j)} - F^{\text{mean}}_{(i,j)}]/\sigma_{(i,j)}$, where $F^{\text{obs}}_{(i,j)}$ is the observed number of interactions, $F^{\text{mean}}_{(i,j)}$ and $\sigma_{(i,j)}$ are the average number of interactions and the SD, respectively, in 10,000 MS02 null models (37). A cutoff value of 10 is chosen in this presentation. The data matrix is in Table 2, which is published as supporting information on the PNAS web site. (*b*) *Z* scores for the average shortest paths of the isotemporal categories in the largest component of the analyzed protein interaction network. For categories *i* and *j*, $Z_{(i,j)} = [d^{\text{obs}}_{(i,j)} - d^{\text{mean}}_{(i,j)}]/\sigma_{(i,j)}$ where $d^{\text{obs}}_{(i,j)}$ is the observed average shortest path, $d^{\text{mean}}_{(i,j)}$and $\sigma_{(i,j)}$ are the averaged average shortest path and the SD, respectively, in 500 isomorphic MS02 null models. A cutoff value of 5 is chosen in this presentation. The data matrix is in Table 3, which is published as supporting information on the PNAS web site.

during the eukaryotic expansion tend to interact among themselves. In addition, the preexisting network may also contain clusters constrained by function, and many of these clusters have been preserved during the network evolution. For example, category b111111 (d63) may contain the most ancient nodes, and $Z_{(63,\,63)} = 13.6$, which indicates that these nodes still tend to interact among themselves. The result here suggests that evolution of the yeast protein interaction network has undergone additions of clusters of nodes, which we term isotemporal clusters (detailed below).

All observed negative *Z* scores are intercategorical. One of the most interesting ones is $Z_{(3,\,63)} = -9.1$, which indicates that the eukaryote-conserved proteins (b000011) tend to avoid the most conserved proteins (b111111).

To support the above conclusions, we also calculated the average shortest paths within and between the isotemporal categories in the largest connected component of the yeast protein interaction network. The above analysis considered only direct association, whereas the average shortest paths can measure indirect association. We used *Z* scores to evaluate traversal patterns within and between isotemporal categories, based on the isomorphic MS02 null models. Although this isomorphic null model is statistically overstringent, it is sufficient for evaluating the traversal profiles of the isotemporal categories. The *Z* score matrix shows that the intracategory traversal distances are usually significantly below random expectations (Fig. 2*b*). Thus, this analysis also shows that intracategory association tendencies are stronger than intercategory association tendencies.

Reconstruction of the Main Network Evolutionary Path. We reconstructed the main growth path of the network from the interaction patterns among the following six major isotemporal categories: b000000, b000001, b000011, b001111, b110011, and b111111. In our designed isotemporal categories, there are two groups of genomes for each domain of life (Eubacteria, Archaea, and Eukaryotes) (38). Categories b000011, b001111, b110011, and b111111 contain identical orthologous hits in both groups of genomes in each domain of life, and they are informative about the root of the universal tree of life (19, 38). Categories b000001 and b000000 may reveal the recent evolutionary history of the yeast. Furthermore, these six categories have large sample sizes.

We converted the *Z* score of intercategory interaction tendency into distance (d_z) through a logit-like transformation, $d_z = 1/(1 + e^Z)$, which transforms the *Z* scores into the range (0, 1). Positive *Z* scores correspond to small d_z values because they indicate that the observed intercategory interactions are above random expectations. Conversely, negative *Z* scores correspond to large d_z values. From the d_z distance matrix, we inferred an NJ tree (42) that describes the intercategory interaction tendencies of the major isotemporal categories (Fig. 3*a*). This tree is essentially the blueprint that accounts for the expansion of the protein interaction network, by means of the addition of groups of proteins to the network at various periods during evolution. The main assembling order of the major groups is represented by the path from the ancient proteins (b111111) to eukaryote-conserved proteins (b000011) and then to recent proteins (b000001 and b000000). Assuming that there existed an ancestral protein interaction network represented by the b111111 nodes, and assuming that network evolution can be described by node additions, the path from the ancient proteins to the recent ones in the NJ tree would thereby describe the major path of the network growth.

The positioning of b001111 (conserved between Archaea and Eukaryotes) and b110011 (conserved between Eubacteria and Eukaryotes) is consistent with the symbiotic hypothesis of the eukaryotic origin that argues for an archaeal host and a eubacterial symbiont (43).

Likewise, through the transformation, $d'_z = 1/(1+ e^{-Z})$, of the *Z* scores of the average shortest paths, we inferred an NJ tree with the same branching pattern (Fig. 3*b*). Therefore, by using two independent measurements, we observed that network evolution mirrors the universal tree of life.

Isotemporal Clusters in the Network. By using a single-linkage clustering method (44), we isolated the isotemporal clusters in the yeast protein interaction network by merging interacting proteins from the same isotemporal category into one node (see Fig. 6, which is published as supporting information on the PNAS

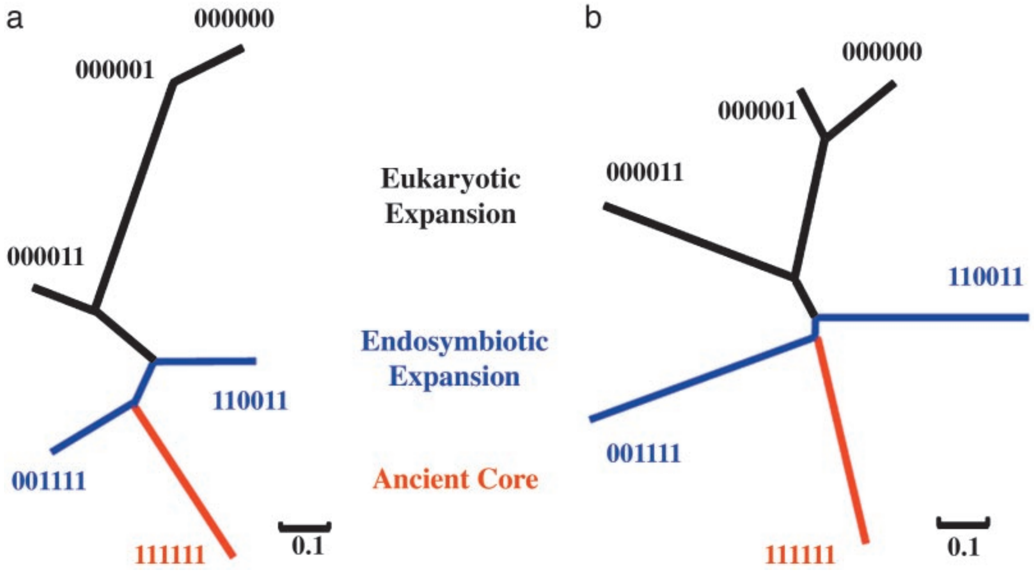


Fig. 3. The main path of network growth. (*a*) An NJ tree based on $d_z = 1/(1 + e^Z)$, where *Z* is the *Z* score for interaction tendencies from Fig. 2*a*. (*b*) An NJ tree based on $d'_z = 1/(1 + e^{-Z})$, where *Z* is the *Z* score for the average shortest path from Fig. 2*b*. Both methods give the same branching pattern.

web site). To estimate the clustering significance, we again used the isomorphic MS02 null model. For most isotemporal categories with relatively large populations, the numbers of their isotemporal clusters are significantly lower than the random expectations (Table 1). This result further supports the role of synergistic selection during network evolution. It is possible that new proteins are randomly added to the network. A single new addition to the network is more likely to be functionally irrelevant or deleterious, and tends to be filtered out during evolution, whereas additions of several interacting new proteins are more likely to be functional relevant and preserved. The observed isotemporal clusters and the proposed synergistic selection are consistent with the observed modularity in biological networks (7, 45).

Discussion

Although we used the best annotated data available at the time of this study, the problems of false-positive and false-negative (14, 30, 46–50) data were not completely avoided. There is also the biased coverage toward conserved proteins (30). All these factors, however, likely affect the inter- and intracategory interactions randomly and so may not alter our main conclusions.

Our isotemporal classification of yeast proteins is limited by the sequence similarity search, the methods chosen to define orthologous groups, and the number of genomes available. These limitations, however, would largely affect the bits with 0 in the b coding scheme and would contribute to the large sample sizes of b000000 and b000001. Possibly, some b000011 proteins have been misclassified as b000001, and some b000001 proteins have been misclassified as b000000. As a result, some true b000011-b111111 associations may have been misclassified as b000001-b111111 or b000000-b111111. These misclassifications may affect both b000000 and b000001 to a similar extent and therefore may not drastically alter the inferred intercategory association tendencies among these categories. In addition, misclassification decreases intracategory *Z* scores, which means that the true intracategory association is actually more significant than estimated above.

Table 1. Numbers and sizes of major isotemporal clusters

Isotemporal categories	Cluster numbers			Average cluster sizes	
	No.	*Z* score	*P* value	Size	*Z* score
000000	357	−6.2	<0.001	1.31	7.1
000001	272	−2.6	0.007	1.42	2.8
000011	264	−1.9	0.018	2.6	2.1
001111	46	−7.4	<0.001	2.13	10.9
110011	66	−4.1	<0.001	1.39	4.7
111111	199	−4.2	<0.001	1.67	4.9

Z scores and *P* values are calculated based on 1,000 isomorphic MS02 null models. A three-dimensional presentation of the isotemporal clusters is provided in Fig 6.

The evolutionary origin of cellular life has been a controversial issue (18, 20, 51). The endosymbiotic hypothesis (19, 43) postulates an archaebaterium as the host and a eubacterium as the symbiont. From our observed significant intracategory association for all isotemporal categories of proteins, the significant separation tendency between b000011 (eucarya-conserved) and b111111 (ancient) proteins, and the inferred path of the network evolution, our result is strongly consistent with the endosymbiotic hypothesis. In addition, comparison of metabolic networks is also consistent with this hypothesis (5, 52).

The key disagreement between the Darwinian view and the universality view on the evolution of biological complexity is the role of historical contingency (22, 27). Undoubtedly, efforts to search for universal rules benefit our understanding on biological complexity. However, by using the yeast protein interaction network as an example, we observed a correlation between network evolution and the universal tree of life. This observation strongly argues that network evolution is not ahistorical, but is, in essence, a string of historical events.

Although the turnover rate of the protein interaction network is suggested to be very fast (9), our results suggest that many isotemporal clusters can still remain well preserved during evolution. The formation and conservation of isotemporal clusters during evolution may be the consequence of selection for the modular organization of the protein interaction network. The progressive nature of the network evolution and significant isotemporal clustering may have contributed to the hierarchical organization of modularity in biological networks in general (7). Because of the similarities between biological and nonbiological networks (1–3, 6, 7), isotemporal clustering and synergistic selection may be relevant in the evolution of many complex networks.

We thank Bill Martin, Kateryna Markova, Todd Oakley, Martin Feder, Xun Gu, Eugene Koonin, Natalia Maltsev, Maximino Aldana, Henrik Kaessmann, Leigh Van Valen, Thomas Nagylaki, Zhenglong Gu, Joel Bader, Anton Nekrutenko, Jianming Zhang, Geoffrey Morris, Tao Pan, Leo Kadanoff, Peter Bauman, Shinhan Shiu, Richard Blocker, and many others for assistance, suggestions, and discussions. This work was supported by National Institutes of Health Grants GM30998 and GM66104 (to W.-H.L.).

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999) *Nature* **402,** C47–C52.
2. Davidson, E. H., McClay, D. R. & Hood, L. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1475–1480.
3. Barabasi, A. L. (2002) *Linked: The New Science of Networks* (Perseus Publishing, Cambridge, MA).
4. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002) *Nat. Genet.* **31,** 64–68.
5. Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabasi, A. L. & Szathmary, E. (2001) *Nat. Genet.* **29,** 54–56.
6. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002) *Science* **298,** 824–827.
7. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A. L. (2002) *Science* **297,** 1551–1555.
8. Wolf, Y. I., Karev, G. & Koonin, E. V. (2002) *BioEssays* **24,** 105–109.
9. Wagner, A. (2001) *Mol. Biol. Evol.* **18,** 1283–1292.
10. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002) *Science* **296,** 750–752.
11. Rzhetsky, A. & Gomez, S. M. (2001) *Bioinformatics* **17,** 988–996.
12. Wagner, A. (2003) *Proc. R. Soc. London Ser. B* **270,** 457–466.
13. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000) *Nature* **403,** 623–627.
14. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 4569–4574.
15. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. & Sakaki, Y. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 1143–1147.
16. Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., *et al.* (2002) *Science* **295,** 321–324.
17. Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., *et al.* (2001) *Science* **294,** 2364–2368.
18. Woese, C. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8742–8747.
19. Brown, J. R. & Doolittle, W. F. (1997) *Microbiol. Mol. Biol. Rev.* **61,** 456–502.
20. Martin, W. & Russell, M. J. (2003) *Philos. Trans. R. Soc. London B* **358,** 59–85.
21. Maynard Smith, J. & Szathmáry, E. (1995) *The Major Transitions in Evolution* (W. H. Freeman Spektrum, Oxford).
22. Kauffman, S. (1993) *The Origins of Order: Self-organization and Selection in Evolution* (Oxford Univ. Press, New York).
23. Corning, P. A. (1995) *Syst. Res.* **12,** 89–121.
24. Thompson, D. W. (1917) *On Growth and Form* (Cambridge Univ. Press, Cambridge, U.K.).
25. Wolfram, S. (2002) *A New Kind of Science* (Wolfram Media, Champaign, IL).
26. Oltvai, Z. N. & Barabasi, A. L. (2002) *Science* **298,** 763–764.
27. Gould, S. J. (2002) *The Structure of Evolutionary Theory* (Harvard Univ. Press, Cambridge, MA).
28. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002) *Nucleic Acids Res.* **30,** 31–34.
29. Mewes, H. W., Albermann, K., Heumann, K., Liebl, S. & Pfeiffer, F. (1997) *Nucleic Acids Res.* **25,** 28–30.
30. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) *Nature* **417,** 399–403.
31. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278,** 631–637.
32. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. & Koonin, E. V. (2001) *Nucleic Acids Res.* **29,** 22–28.
33. Cormen, T. H., Leiserson, C. E. & Rivest, R. L. (1990) *Introduction to Algorithms* (MIT Press, Cambridge, MA).
34. Batagelj, A. & Mrvar, A. (1998) *Connections* **21,** 47–57.
35. Ripley, B. D. (2001) *MSOR Connections* **1,** 23–25.
36. Page, R. D. M. (1996) *Comput. Appl. Biosci.* **12,** 357–358.
37. Maslov, S. & Sneppen, K. (2002) *Science* **296,** 910–913.
38. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
39. Achenbach-Richter, L., Gupta, R., Stetter, K. O. & Woese, C. R. (1987) *Syst. Appl. Microbiol.* **9,** 34–39.
40. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., *et al.* (1998) *Nature* **392,** 353–358.
41. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., *et al.* (1999) *Nature* **399,** 323–329.
42. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
43. Martin, W., Hoffmeister, M., Rotte, C. & Henze, K. (2001) *Biol. Chem.* **382,** 1521–1539.
44. Anderberg, M. R. (1973) *Cluster Analysis for Applications* (Academic, New York).
45. Rives, A. W. & Galitski, T. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1128–1133.
46. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002) *Mol. Cell Proteomics* **1,** 349–356.
47. Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. & Gerstein, M. (2002) *Trends Genet.* **18,** 529–536.
48. Bader, G. D. & Hogue, C. W. (2002) *Nat. Biotechnol.* **20,** 991–997.
49. Aloy, P. & Russell, R. B. (2002) *FEBS Lett.* **530,** 253–254.
50. Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A. & Holstege, F. C. (2002) *Mol. Cell* **9,** 1133–1143.
51. Brenner, S. (1998) *Science* **282,** 1411–1412.
52. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000) *Nature* **407,** 651–654.

Explore Biological Pathways from Noisy Array Data by Directed Acyclic Boolean Networks

Lei M. Li[*,‡] and Henry Horng-Shing Lu[†]

April 5, 2004

Abstract

We consider the structure of directed acyclic Boolean (DAB) networks as a tool of exploring biological pathways. In a DAB network, the basic objects are binary elements and their Boolean duals. A DAB is characterized by two kinds of pairwise relations: similarity and prerequisite. The latter is a partial order relation, namely, the on-status of one element is necessary for the on-status of another element. A DAB network is uniquely determined by the state space of its elements. We arrange samples from the state space of a DAB network in a binary array and introduce a random mechanism of measurement error. Our inference strategy consists of two stages. First, we consider each pair of elements, and try to identify their most likely relation. In the meantime, we assign a score, s-p-score, to this relation. Second, we rank the s-p-scores obtained from the first stage. We expect that relations with smaller s-p-scores are more likely to be true, and those with larger s-p-scores are more likely to be false. The key idea is the definition of s-scores (referring to similarity), p-scores (referring to prerequisite), and s-p-scores. Like classical statistical tests, control of false negatives and false positives are our primary concerns. We illustrate the method by a simulated example, the classical arginine biosynthetic pathway, and show some exploratory results on a published microarray expression dataset of yeast *Saccharomyces cerevisiae* obtained from experiments with activation and genetic perturbation of the pheromone response MAPK pathway.

Keywords: microarray, pathway, Boolean networks, measurement error, EM algorithm

Running head: Directed Acyclic Boolean Networks

[*] Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.

[†] Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan.

[‡] Email address of corresponding author: `lilei@hto.usc.edu`, phone: (213)740-2407, fax: (213)740-2437

1 Introduction

One great challenge of post-genomic research is to identify complex biological networks and pathways from genome-wide data such as DNA sequences and expression profiles. This includes metabolic pathways, protein-protein interaction networks, gene regulatory pathways, etc. Along with biological methods such as phylogenetic profile and Rosetta Stone, see Eisenberg *et al.* (2000), McGuire and Church (2000), computational methods have been developed as powerful data-mining tools in the study of genomics.

Clustering is one such important technique to group genes and samples from microarray data; see Eisen *et al.* (1998), Ben-Dor *et al.* (1999), Alon *et al.* (1999). A central component of a clustering algorithm is the definition of similarity scores, either from a biological perspective or from a statistical perspective. We note that the relation of similarity between two biological elements such as proteins or genes is symmetric in nature. On the other hand, a biological process may include a cascade of reactions to environmental factors and regulation of protein syntheses. Thus concepts other than similarity are necessary for a complete description of pathways.

Data type is another consideration in the modelling of networks. In this article, we consider binary variables because we can always discretize continuous variables. In the presence of noise, careful discretization can even denoise to some degree. One such example can be found in Xing and Karp (2001). The use of Boolean networks has a long history in literature. Kauffman (1977), Kauffman (1979) considered a dynamic version of Boolean networks. A review of models of genetic regulatory systems including Boolean networks can be found in De Jong (2002). Based on the structure of Boolean networks, we introduce a new model for measurement error and propose a simple technique to infer pairwise relations between elements from noisy array data.

We note that Bayesian networks is a much more sophisticated and complete model to describe biological pathways than the method proposed in this article. For example, variables in a Bayesian networks can be either discrete or continuous. Bayesian networks is a structure that contains directed relations among elements. It has been extensively studied in the last two decades; see Pearl (1988) and Jensen (1996). Its structure is characterized by two components. The first component is a directed acyclic graph whose vertices correspond to random variables. The second component describes a conditional distribution for each variable, given its parents in the graph. Murphy and Mian (1999) and Friedman *et al.* (2000) applied Bayesian network models to analyze microarray expression data. The family of Bayesian networks is fairly large and the number of DAGs is super-exponential. Although some algorithms searching for Bayesian networks have been developed, see Heckerman *et al.* (1995), Spirtes *et al.* (2000a), the learning of Bayesian networks is a challenging task without *a priori* knowledge. Besides, to achieve high accuracy of estimation, sample sizes of several hundred are required even for relatively sparse graphs, see Spirtes *et al.* (2000b). The simple model considered in this article takes some aspects of Bayesian networks and serves as a tool of exploratory data analysis for array data.

Specifically we consider the structure of directed acyclic Boolean (DAB) networks as a tool of exploring biological pathways. In a DAB networks, the basic objects are binary elements and their Boolean duals. A DAB is characterized by two kinds of pairwise relations: similarity and prerequisite. The former represents a pair of elements with identical on-off states. The latter is a partial order relation, namely, the on-status of one element is prerequisite for the on-status of another element. A DAB networks is uniquely determined by its state space: all possible on-off states subject to the pairwise relations. We arrange samples from the state space of a DAB network in a binary array, and then introduce a random mechanism of measurement error. This results in a noisy array. Our goal is to reconstruct the DAB networks from the noisy array data.

Our inference strategy consists of two stages. First, we consider each pair of elements, and try to identify their most likely relation. In the meantime, we assign a score, s-p-score, to this relation. Second, we rank the s-p-scores obtained from the first stage. We expect that those relations with smaller s-p-scores are more likely to be true, and those with larger s-p-scores are more likely to be false. The key idea is the definition of s-scores (referring to **similarity**), p-scores (referring to **prerequisite**), and s-p-scores (by model selection). Like classical statistical tests, control of false negatives and positives are our primary concerns.

The s-p-scoring method is one kind of exploratory data analysis, and focuses on pairwise relations. After the ranking of pairwise relations, experts' knowledge may be incorporated. Depending on data, we expect to reconstruct all or partial sub-structures of a network. If we set an upper bound to the number of E-M iterations involved, the computational complexity of the procedure is $O(m^2 \log m)$, where m is the number of elements in a network.

The rest of the paper is organized as follows. In Section 2 we describe the structure of the model. In Section 3 we explain the s-p-scoring method. In Section 4 we illustrate the method by a simulated example, the classical arginine biosynthetic pathway, and show some exploratory results on the yeast *Saccharomyces cerevisiae* pheromone response MAPK pathway using an expression dataset obtained from experiments with activation and genetic perturbation. In Section 5 we discuss some relevant issues.

2 The model

The structure of directed acyclic Boolean (DAB) networks Suppose we are concerned with m elements, G_1, G_2, $\cdots$, G_m, each taking two states: on and off. These elements are abstracts of biological objects such as genes, mRNAs, proteins, environmental conditions, or a mixture of them. If an element is measured on a continuous scale or has more than two expression levels, then we need to discretize it and encode it by binary variables. We will come back to this issue later. The theory of directed graphs is helpful for the description of our model; we refer readers to Brightwell (1997) for relevant results on this subject. We generate a graph with $2m$ vertices or nodes, G_1, G_2, $\cdots$, G_m, and their Boolean duals $\bar{G}_1$,

$\bar{G}_2$, $\cdots$, $\bar{G}_m$, representing on-and-off state of the m elements, and this is referred to as the ground-set. We refer to a node A and its dual $\bar{A}$ as a Boolean pair.

We define a prerequisite relation between a pairs of elements A and B as follows: A is prerequisite for B if the on-status of A is necessary for the on-status of B, and we denote it by $A \prec B$. The prerequisite relation is a partial order. It is transitive on the ground-set, namely, $A \prec C$ and $C \prec B$ implies $A \prec B$. Also it is irreflexive in the sense that we never have $A \prec \bar{A}$. In addition, we assume that the dual of each partial order relation is also true, i.e. $\bar{B} \prec \bar{A}$ is true if and only $A \prec B$ is true. Similarly, we have the following three pairs of dual relations: $\bar{A} \prec \bar{B}$ with $B \prec A$; $A \prec \bar{B}$ with $B \prec \bar{A}$; and $\bar{A} \prec B$ with $\bar{B} \prec A$. We graphically represent a partial relation $A \prec B$ by drawing an arrow from the vertex A to B. It is not economical to include all the arcs in the directed graph due to the transitive property of partial orders. An ordered pair (A, B) is called a covering pair if there exists no vertex C such that $A \prec C$ and $C \prec B$. Thus it suffices to represent all partial orders by arrows between covering pairs, and this is referred to as the diagram of the directed graph. It is well known that the diagram of a partial order is acyclic. In addition, no path exists to connect a Boolean pair in the diagram of a DAB because we never have $A \prec \bar{A}$.

Another relation between pairs of elements is similarity. Two elements A and B are 'similar" if they are on and off simultaneously, and this is denoted by $A \sim B$. They are negatively "similar" if they are on and off in the opposite way, and this is denoted by $A \sim \bar{B}$. In the absence of measurement error, it is a trivial relation. But in practice, the presence of measurement error complicates the situation and it needs to be inferred from the data.

We use "—" to connect two "similar" elements in the diagram. Figure 1 shows a directed acyclic Boolean network, which has seven elements with one similar and eleven prerequisite relations. Another way to identify a DAB is to consider the on-off states of its elements. There are in total $2^7 = 128$ states for a seven-element DAB. Only thirteen of these states are compatible with the twelve pairwise relations in the above example. We enumerate them in Table 1, where "0" and "1" represent "off" and "on" respectively. It is a subset of the 128 states. In general, a directed acyclic Boolean network consisting of m elements corresponds to a unique subset of all 2^m states. Even though not every subset of the 2^m states corresponds to a directed acyclic Boolean network, the number of DABs, like the number of DAGs, is super-exponential.

Consider n samples generated from a directed acyclic Boolean network, i.e. we sample with replacement from the state space compatible with the networks; Table 1 shows the compatible states for the above example. We arrange the data in a matrix (y_{ij}), where $i = 1, \cdots, n$, $j = 1, \cdots, m$, whose entries take values of either 0 or 1. Table 1 is the transpose of (y_{ij}), and each row corresponds to an element and each column corresponds to a sample.

Without measurement error, we can reconstruct the directed acyclic Boolean network in Figure 1 from Table 1 by identifying all the pairs with prerequisite or similar relations. This is carried out by the following procedure. For each pair of elements, say, A and B, we count the four incidences of (A, B) being $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ from the corresponding columns of (y_{ij}), and arrange them in a 2×2 table; see the left of Table 2. We mark a cell "+" if the count is positive and mark it "0" otherwise. Consequently, the six relations are characterized by the count patterns in Table 3.

Next we consider the issue of selection bias. In practice, we sample from all the possible states compatible with a directed acyclic Boolean network. In the above example, we have only 13 cases. When m is large, this number could be large, and possibly only a fraction of them are sampled. Then the issue of estimableness arises. If we cannot have an exhaustive sample, i.e., some compatible states are missed in observation, then the count strategy described above may lead to false positive pairwise relations, either prerequisite or similarity. For example, if case 3 in Table 1 is missed from observations, then the count strategy indicates $C \prec B$, which is not consistent with the truth. Nevertheless, this strategy will not lead to any false negatives in the absence of measurement error.

Measurement error Next we introduce a mechanism of measurement error to the data sampled from a directed acyclic Boolean network. This results in a more practical model for many biological data such as expression levels. We assume that each entry in (y_{ij}) is switched to its opposite value according to a misclassification probability p, independently with one another, i.e.

$$x_{ij} = \begin{cases} y_{ij} & \text{with probability } 1-p\,, \\ 1-y_{ij} & \text{with probability } p\,. \end{cases}$$

This creates the noisy array (x_{ij}), which are the observations.

Problem and pairwise structure Our goal is to reconstruct the directed acyclic Boolean network from the array of binary data (x_{ij}). It is clear that the problem is equivalent to identifying all the pairs of elements with estimable similarity or prerequisite relations.

3 Method

Our inference strategy consists of two stages. First, we consider each pair of elements, and try to find their most likely relation. In the meantime, we assign a score, s-p-score, to this relation. Second, we rank the s-p-scores obtained from the first stage. We expect that those relations with smaller s-p-scores are more likely to be true, and those with larger s-p-scores are more likely to be false.

Probabilistic models for 2 by 2 tables To deal with measurement error, we resort to probabilistic models. Instead of a full model including every element, we consider pairwise models in the first stage. The count data in the 2 × 2 table on the left of Table 2 can be thought as being generated from a multinomial distribution with four cells whose probabilities are q_{00}, q_{01}, q_{10}, q_{11} respectively, as shown on the right of Table 2, where $q_{00} + q_{01} + q_{10} + q_{11} = 1$. Then the six types of relations between elements A and B are reformulated as hypotheses on the probability patterns; see Table 4. Please notice that (q_{00}, q_{01}, q_{10}, q_{11}) depend on both the structure of the DAB network and the sampling scheme.

Similar to (y_{ij}), we extract the data in (x_{ij}) for each pair of elements, say A and B, and arrange them on the left of Table 5. Now the counts n_{00}, n_{01}, n_{10}, n_{11} are not generated from the multinomial (q_{00}, q_{01}, q_{10}, q_{11}), but from another multinomial (r_{00}, r_{01}, r_{10}, r_{11}) as shown on the right of Table 5, where $r_{00} + r_{01} + r_{10} + r_{11} = 1$.

Missing data structure With measurement error, a part of m_{00} may leak to the other three cells. We denote the re-distributed counts from m_{00} to the four cells by $m_{00,00}$, $m_{00,01}$, $m_{00,10}$, $m_{00,11}$. Analogous notation is defined for m_{01}, m_{10} and m_{11}. This splitting pattern is shown in Table 6. Correspondingly, their generating probabilities (q_{00}, q_{01}, q_{10}, q_{11}) are re-distributed as shown in Table 7, where we adopt the notation $q_{ij,kl}$ analogous to $m_{ij,kl}$. The two sets of counts and probabilities are linked as follows.

$$\left\{ \begin{array}{lcl} n_{ij} & = & \sum_{k,l=0,1} m_{kl,ij} \ , \\ r_{ij} & = & \sum_{k,l=0,1} q_{kl,ij} \ , \end{array} \right. \tag{1}$$

and

$$\left\{ \begin{array}{lcl} m_{kl} & = & \sum_{i,j=0,1} m_{kl,ij} \ , \\ q_{kl} & = & \sum_{i,j=0,1} q_{kl,ij} \ . \end{array} \right.$$

MLE and E-M algorithm The log-likelihood of the data is given, up to a constant, by the following

$$L = \sum_{i,j=0,1} n_{ij} \log r_{ij} \ , \tag{2}$$

where the probabilities r_{ij}'s are computed according to (1) and Table 7. Later we define s-scores and p-scores via maximum likelihood estimates (MLE). Except for a constant, the log-likelihood of the full data $\{m_{ij,kl}\}$ is given by

$$\sum_{i,j,k,l=0,1} m_{ij,kl} \log q_{ij,kl} \ , \tag{3}$$

where $q_{ij,kl}$ are those splitting probabilities in Table 7.

To estimate the MLE, the celebrated E-M algorithm maximizes the likelihood of full data (3) rather than that in (2); see Dempster *et al.* (1977), McLachlan and Krishnan (1997). In the E-step, we impute the splitting counts by their conditional expectations calculated at the current value of the parameter by the formula

$$E_{(p,q_{00},q_{01},q_{10},q_{11})}(m_{ij,kl}|n_{kl}) = \frac{n_{kl}\, q_{ij,kl}}{\sum_{i',j'=0,1} q_{i'j',kl}}\,, \tag{4}$$

where $i, j, k, l = 0, 1$. Under different hypotheses specified in Table 4, one or two probabilities of q_{00}, q_{01}, q_{10} and q_{11} are zero. In the M-step, we update the value of the parameter by maximizing the conditional expectation of the log-likelihood for the full data; See Li and Lu (2001) for details.

Pairwise scores We first consider a simpler problem than reconstructing a DAB network: what is the most likely relation for a pair of elements?

Definition 1 *For a pair of elements A and B,*

- *the* **s-scores** $s_{A\sim B}$ *and* $s_{A\sim\bar{B}}$ *are respectively the maximum likelihood estimates of p under the diagonal model: $q_{01} = q_{10} = 0$ and $q_{00} = q_{11} = 0$;*
- *the* **p-scores** $p_{A\prec B}$, $p_{\bar{A}\prec\bar{B}}$, $p_{A\prec\bar{B}}$, *and* $p_{\bar{A}\prec B}$ *are respectively the maximum likelihood estimates of p under the triangular model: $q_{01} = 0$, $q_{10} = 0$, $q_{00} = 0$, and $q_{11} = 0$; cf. Table 4.*

We compute s-scores and p-scores by the E-M algorithm described earlier. The heuristic of the definition is that we use the MLE $\hat{p}$ to measure the goodness of fit of each hypothesis: the smaller the score, the more support to the corresponding hypothesis.

Next we need to choose one score out of the two s-scores and four p-scores for a pair of elements. In other words, we need to select the hypothesis that is most consistent with the data. This is a problem of model selection; see Schwarz (1978).

Definition 2 *For a pair of elements A and B,*

- *Between the two diagonal models, select the one that achieves the smaller s-score;*

- *Among the four triangular models, select the one that achieves the smallest p-score;*

- *For the diagonal model corresponding to the smaller s-score and the triangular model corresponding to the smallest p-score, we compare their corresponding BIC values, namely, the penalized log-likelihoods as follows:*

  $$BIC = -\log likelihood + \frac{d \log n}{2},$$

  where n is the sample size and d is the number of parameters. This number is two for a diagonal model and is three for a triangular model. We choose the model with the smaller BIC value as the most likely relation for the pair A and B, and define their **s-p-score** *to be the score corresponding to the most likely relation.*

Please notice that s-p-score is one of the s-scores and p-scores, BIC values are only used to choose the hypothesis. It is easy to understand why we select the smallest s-score and p-score. Notice that each diagonal model is nested in two triangular models. To make the choice between a diagonal and a triangular model, we need to take into account model complexity. We here adopt the technique of BIC for model selection.

The basic idea of most powerful statistical tests is to minimize the chance of type II error (false positive) subject to a constraint on the chance of type I error (false negative); see Lehmann (1986). Even though the classical theory of hypothesis testing does not directly apply to our situation, its rationale remains our guide. For each hypothesis in Table 4, we expect that the s-score or p-score has the following property: it is a good estimate of the parameter p when the hypothesis is true; whereas it is considerably biased upward when the hypothesis is false.

Accuracy of estimation and control of false negative We next consider the statistical behavior of the s-scores and p-scores under the null hypothesis. Without loss of generality, we take the hypothesis: $q_{01} = 0$ for example. Notice that this is a composite hypothesis. In general, the maximum likelihood estimate in a regular setting is both consistent and efficient, see Bickel and Doksum (1977).

Proposition 1 *Suppose that the hypothesis $A \prec B$, i.e. $q_{01} = 0$ holds. Then except for the singular point at $q_{00} = q_{11} = 0$, the maximum likelihood estimate of p has the property of asymptotical normality, i.e.*

$$\sqrt{n}\,[\hat{p} - p, \hat{q}_{00} - q_{00}, \hat{q}_{10} - q_{10}, \hat{q}_{11} - q_{11}] \longrightarrow N(0, I^{-1})\,,$$

where I is the Fisher information matrix,

$$I = -\begin{pmatrix} E[\frac{\partial^2 logL}{\partial p^2}] & E[\frac{\partial^2 logL}{\partial p \partial q_{00}}] & E[\frac{\partial^2 logL}{\partial p \partial q_{10}}] & E[\frac{\partial^2 logL}{\partial p \partial q_{11}}] \\ E[\frac{\partial^2 logL}{\partial p \partial q_{00}}] & E[\frac{\partial^2 logL}{\partial q_{00}^2}] & E[\frac{\partial^2 logL}{\partial q_{00} \partial q_{10}}] & E[\frac{\partial^2 logL}{\partial q_{00} \partial q_{11}}] \\ E[\frac{\partial^2 logL}{\partial p \partial q_{10}}] & E[\frac{\partial^2 logL}{\partial q_{00} \partial q_{10}}] & E[\frac{\partial^2 logL}{\partial q_{10}^2}] & E[\frac{\partial^2 logL}{\partial q_{10} \partial q_{11}}] \\ E[\frac{\partial^2 logL}{\partial p \partial q_{11}}] & E[\frac{\partial^2 logL}{\partial q_{00} \partial q_{11}}] & E[\frac{\partial^2 logL}{\partial q_{00} \partial q_{10}}] & E[\frac{\partial^2 logL}{\partial q_{11}^2}] \end{pmatrix}.$$

It will take more than ten pages to write down the expression of I^{-1}. In fact, the computation was carried out by the symbolic calculation in MAPLE. Here we choose to only give the term corresponding to the parameter p as follows:

$$\frac{p(1-p)(3p^2 q_{00} + 3p^2 q_{11} - p^2 q_{10} - 3pq_{00} - 3pq_{11} + pq_{10} + q_{11} + q_{00})}{n(4p^2 q_{11}^2 + 4p^2 q_{00}^2 + 8p^2 q_{00} q_{11} - 4pq_{11}^2 - 4pq_{00}^2 - 8q_{00} p q_{11} + 2q_{00} q_{11} + q_{11}^2 + q_{00}^2)}\,. \tag{5}$$

In Figure 2, we plot the element of I^{-1} corresponding to p as a function of q_{00} and q_{01} in which p is fixed to be 0.05. The only singularity point occurs at $q_{10} = 1$ and $q_{00} = q_{11} = q_{01} = 0$. In this case, one element is house-keeping (on all the time), and the other one is silent (off all the time). By filtering out silent and house-keeping elements, we can eliminate this kind of singularity for the sake of inference. Consequently, we can find a bound on the inverse of the Fisher information matrix, and this means that the p-score will be around p within an order $1/\sqrt{n}$ radius asymptotically.

Control of false positive Next we look at how the p-score $p_{A \prec B}$ behaves under the alternatives: $q_{01} > 0$ versus the null $q_{01} = 0$. We study the asymptotic bias of the MLE.

Proposition 2 *Let the parameters in the true model be (p, q_{00}, q_{01}, q_{10}, q_{11}), where $q_{01} > 0$. As the sample size $n \to \infty$, the MLE ($\tilde{p}$, $\tilde{q}_{00}$, $\tilde{q}_{01}$, $\tilde{q}_{10}$, $\tilde{q}_{11}$) subject to $\tilde{q}_{01} = 0$ is given by the value that minimizes the Kullback-Leibler divergence between the null and alternative:*

$$D[\{p, q_{00}, q_{01}, q_{10}, q_{11}\}||\{\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11})\} = D[\{r_{ij}\}||\{\tilde{r}_{ij}\}] = \sum_{i,j=0,1} [-r_{ij} \log \tilde{r}_{ij} + r_{ij} \log r_{ij}],$$

where $\{r_{ij}\}$ and $\{\tilde{r}_{ij}\}$ are respectively defined by $\{p, q_{00}, q_{01}, q_{10}, q_{11}\}$ and $\{\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11}\}$ via (1) and Table 7.

The concept of Kullback-Leibler divergence can be found in Cover and Thomas (1991). The proof lies in the connection between likelihood and Kullback-Leibler divergence. When $n \longrightarrow \infty$, $n_{ij}/n \longrightarrow r_{ij}$, and maximizing the quantity in (2) becomes maximizing the following

$$\sum_{i,j=0,1} n \, r_{ij} \log \tilde{r}_{ij} \, ,$$

over $\{\tilde{r}_{ij}\}$. This is equivalent to minimizing

$$\sum_{i,j=0,1} [-r_{ij} \log \tilde{r}_{ij} + r_{ij} \log r_{ij}] \, ,$$

which is $D[\{r_{ij}\}||\{\tilde{r}_{ij}\}]$. Thus we complete the proof.

We expect that $\tilde{p} - p > 0$ when $q_{01} > 0$. We have confirmed this result numerically. In the range of $0 < p < 0.45$, $0 < q_{01} < 0.5$, we set up a mesh and calculate $\tilde{p} - p = p_{A \prec B} - p$. Figure 3 shows the result when $p = 0.05$ and $q_{01} = 0.1$.

Now we explain why we rather take $\hat{p}$ than the likelihood ratio as the statistics to test the hypothesis.

Proposition 3 *Suppose $(\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11})$ and $(p, q_{00}, q_{01} > 0, q_{10}, q_{11})$ are respectively the null and alternative hypotheses. Denote the significance level by α, and the chance of type II error of the optimal* *test by β_n, where n is the sample size. Then*

$$\lim_{\alpha \to 0} \lim_{n \to \infty} \frac{1}{n} \log \beta_n = -D[(\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11}) || (p, q_{00}, q_{01}, q_{10}, q_{11})] \, .$$

This result is a direct application of the Stein's lemma; see Chap 12 of Cover and Thomas (1991). It says that the chance of type II error (false positive) is characterized by the Kullback-Leibler divergence between the two hypotheses. We plot the Kullback-Leibler divergence for the case $p = 0.05$, $q_{00} = q_{11} = q_{10}$ in Figure 4. It remains zero till q_{01} reaches 0.25. This indicates that the likelihood ratio test cannot give good protection against false positives. In comparison, we plot $\tilde{p} - p = p_{A \prec B} - p$ against q_{01} for the case $p = 0.05$, $q_{00} = q_{11} = q_{10}$ in Figure 5. It can be seen that the score immediately goes up as q_{01} moves away from zero. Thus we rather adopt p-scores to play the role of test statistic.

Reconstruction of directed acyclic Boolean networks The s-p-scores are more meaningful if they are generated from a directed acyclic Boolean network because we may discover significant pairwise relations by ranking the scores in the ascending order. We collect those pairwise relations whose s-p-scores smaller than a threshold and put them in a watch list. Known biological results are helpful for the determination of threshold. For example, if we know the relation $A \prec B$ is true, then those s-p-scores smaller than $p_{A \prec B}$ should be in our watch list. Please notice that the more pairwise relations are included in the watch list, the more likely we observe incompatible ones. In this case, no DAB network exists to explain all the relations. We here mention one strategy, namely, **maximum compatibility criterion**: choose the maximum threshold value so that the selected pairwise relations contain no conflict. Next we illustrate the method by some examples.

4 Examples

Simulated example For the DAB example consisting of seven elements in Figure 1, we simulate a data set of 76 samples with misclassification probability $p = 0.05$. The data can be arranged in an array similar to that obtained from microarray. Namely, each row in this array corresponds to an element, and each column corresponds to a sample. We compute the 21 s-p-scores and sort them in Table 8. For each pair of elements, we show the counts of $n_{i,j}$ in the last four columns, two s-scores, and four p-scores in the middle. The sorted s-p-scores and their corresponding hypotheses are shown in the first two columns. The true relations and false relations (in parentheses) cross each other by only one case.

Arginine biosynthetic pathway Boolean logic is a useful tool for the study of pathways. We here revisit the analysis of the experiment concerning the biochemical pathway for the synthesis of the amino acid arginine in *Neurospora crassa*. It is a standard example to illustrate the one gene-one enzyme hypothesis, see Russell (1995). The pathway is shown in Figure 6. Using genetic crosses and complementation tests, we know the process involves four genes, which are designated $argE^+$, $argF^+$, $argG^+$, and $argH^+$ in a wild-type cell. The experiments generated growth pattern of the mutant strains on media supplemented with presumed arginine precursors. These intermediates are Ornithine, Citrulline, and Argininosuccinate.

Next we have another look at this example from the perspective of the boolean logic proposed in this paper. First we rearrange the data from the experiments in an array, see Table 9. Please notice that this state table is different from the one shown in Chapter 9, Page 275, in Russell (1995). The first four columns are definitions of the mutants. The next four columns show the presence state of the four arginine precursors when none of them is added externally. This can be deduced by the change of growth pattern after external controls. If we cannot determine the on-off status of an intermediate, we place a question mark.

The problem is to obtain the pathway in Figure 6 from Table 9. By checking with Table 3, we can easily infer that either $E^+ \sim Ornithine$ or $E^+ \prec Ornithine$, $F^+ \prec Citrulline$, $F^+ \prec Argininosuccinate$, $F^+ \prec Arginine$, $G^+ \prec Argininosuccinate$, $G^+ \prec Arginine$, and $H^+ \prec Arginine$. These pairwise relations are consistent with the sequence in Figure 6. Even though the heuristic arguments in Russell

(1995) can do the same job, the pairwise Boolean logic is more general. Also we note that measurement error has not been considered in the example. When measurement error is unavoidable, we still can make inference by s-p-scoring. This is its advantage over no-measurement-error logic.

Yeast expression data To study the signaling and circuitry of multiple mitogen-activated protein kinase (MARK) pathways, Roberts *et al.* (2000) reported the expression data of yeast *Saccharomyces cerevisiae* for various knock-out cells under controlled experimental conditions. They particularly investigated four MARK pathways: pheromone, PKC, HOG, and filamentous growth. We mentioned earlier that it is important to sample as much as possible from the state space of a network to avoid selection bias. This view highlights why various kinds of activation and perturbation, as done in this experiment, are valuable and necessary for the study of pathways. After activating relevant environmental factors (α-factor in this study), a cascade of biological activities occur sequentially. We want to use DAB networks to describe some aspects of these biological processes. We apply the s-p scoring method to explore the expression profiles. Next we show some exploratory result on the pheromone pathway.

During mating of *S. cerevisiae*, haploid MATa and MATα cells communicate with each other through secretion of pheromones α- and a-factor, respectively. Pheromone stimulates yeast cells to increase the expression of mating genes and arrest cell division in the G1 phase of the cell cycle. The responses to pheromone are initiated by a cell surface receptor that couples to a G protein and downstream MAPK kinase cascade; see (Fig. 1A) in Roberts *et al.* (2000). In some experiments, MATa cells are exposed to α-factor concentrations ranging from 0.15 to 500 nM. Cells with various knock-out genes are also tested. The genome-wide expression levels are measured via the technique of cDNA microarrays. Namely, the abundance of each mRNA with respect to the reference is obtained in the form of expression ratios.

In our analysis, we exclude those experiments carried out under a different condition of 2% galactose for 3 hours, and two experiments measured at 0 and 15 minutes after the α-factor exposure. In total, we consider expression profiles from 45 experiments. We include the α-factor as an element, and discretize it by setting it on if the concentration is larger than 0.50 nM and off otherwise. Figure 7 shows a DAB network obtained from our analysis. The part of network close to the α-factor is well reconstructed. That is, the pheromone α-factor activates the receptor Ste2p. Then receptor stimulation releases free Gbg (Ste4p/Ste18p). The transcription factor Ste12p, which activates the promoters of mating, is also identified as one element downstream of the MAPK cascade. The positions of those genes in the middle of the pathway such as Ste20p, Ste11p, Ste7 are missed. FIG1 is a transcriptional reporter gene for activation of the MAPK. Our analysis indicates its position in the pathway as shown in Figure 7. We found that those genes whose expressions stay steady after some exposure to a concentration of α-factor are more easily identified.

5 Discussion

Discretization The data types in the DAB networks are binary. If elements such as expression levels are observed on a continuous scale, then we need to discretize them. In cDNA microarrays, a reference sample is also hybridized to probe. The ratios of expression levels (or differences in the logarithm scale) lead to a natural way of discretization. That is, an element is on if the log-ratio is larger than zero, and is off otherwise. If other information are available for some elements, we can exploit them to achieve better discretization. Consider expression levels of a gene A. Suppose the log-ratio of its expression is l_{-A} in a knock-out experiment $\triangle$A, and is l_{+A} in an experiment in which we know it is over-expressed. Then the threshold L must satisfy: $l_{-A} \leq L \leq l_{+A}$. Histograms of the expression levels are also helpful for discretization. In the case that discretization is not perfect, the error mechanism introduced in the model still allows us to run the s-p-scoring analysis. In Xing and Karp (2001), a mixture model is used as a quantizer for their clustering method and the result is quite good.

Coding issues Each element in a DAB network is a dichotomous variable. In practice, an element may have more than two levels. In this case, we introduce multiple pseudo elements to code for its values. For example, if an element A has four levels, then we code it by two pseudo elements as shown in Table 10. In general, the information in a binary element is equivalent to a bit, and n bits can encode up to 2^n values.

If samples are obtained from a time course, then it is possible to consider differences of expressions between two consecutive time points. In this way, the dynamics of the networks are included in the analysis. For networks with feedback, caution is necessary to apply the s-p-scoring analysis. One strategy is to consider data in a time window, and then examine how the pairwise relations evolve as the time window moves.

Computational complexity The key step of the procedure is the computation of s- and p-scores for each of the $\frac{m(m-1)}{2}$ pairs of elements, where m is the number of elements. The E-M procedure used to compute the MLE is an iterative algorithm. It converges at a linear rate that depends on the fraction of missing data; see McLachlan and Krishnan (1997). The number of iterations required for convergence varies depending on initial values of parameters. A common practice in numerical implementation is setting an upper bound for iterations. Consequently, this keeps the $O(m^2)$ complexity for the computation of MLE. According to our numerical experience, the convergence is quite fast for the 2 by 2 count data. The sorting algorithm such as heapsorting can rank the $\frac{m(m-1)}{2}$ s-p-scores in $O(m^2 \log m)$ time and in place. Thus the overall complexity is $O(m^2 \log m)$ in time and $O(m^2)$ in memory.

Software We have developed MATLAB code for the s-p-scoring method.

Acknowledgments

The authors would like to thank Prof. Michael Waterman, Wing Wong, and Simon Tavaré, for their help. The research is partially supported by the Functional Genomics Program, Institute for Pure and Applied

Mathematics, UCLA. Lu's research is partially supported by the National Science Council in Taiwan. Li's research is partially supported by the CEGS grant from NIH.

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine., A. J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA* 96, 6745–6750.

Ben-Dor, A., Shamir, R., and Yakhini, Z. 1999. Clustering gene expression patterns. *Journal of Computational Biology* 6, 281–297.

Bickel, P. J. and Doksum, K. A. 1977. *Mathematical Statistics : Basic Ideas and Selected Topics*. San Francisco : Holden-Day.

Brightwell, G. 1997. Partial orders. In L. W. Beineke and R. J. Wilson, editors, *Graph Connections: Relationships between Graph Theory and other Areas of Mathematics*. Clarendon Press, Oxford.

Cover, T. M. and Thomas, J. A. 1991. *Elements of Information Theory*. Wiley.

De Jong, H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology* 9, 67–103.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1–22.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. 1998. Clustering analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA* 95, 14863–14868.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. 2000. Protein function in the post-genomic era. *Nature* 405, 823–826.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. 2000. Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 7, 601–620.

Heckerman, D., Geiger, D., and Chickering, D. M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20, 197–243.

Jensen, F. V. 1996. *An Introduction to Bayesian Networks.* University College London Press: London.

Kauffman, S. 1977. Gene regulation networks: A theory for their global structure and behaviors. In *Current Topics in Developmental Biology*, volume 6, pages 145–182. Academic Press, New York.

Kauffman, S. 1979. Assessing the probable regulatory structures and dynamics of the metazoan genome. In R. Thomas, editor, *Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems*, volume 29 of *Lecture Notes in Biomathematics,*, pages 30–60. Springer-Verlag, Berlin.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses.* New York : Wiley.

Li, L. and Lu, H. S. 2001. "Span" directed acyclic boolean networks from array data. Technical report, Florida State University and University of Southern California.

McGuire, A. M. and Church, G. M. 2000. Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucleic Acids Research* 28, 4523–4530.

McLachlan, G. J. and Krishnan, T. 1997. *The EM Algorithm and Extensions.* John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, Weinheim.

Murphy, K. and Mian, S. 1999. Modeling gene expression data using dynamic Bayesian networks. Technical report, University of California at Berkeley, Department of Computer Science.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann: San Francisco.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., and Friend, S. H. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880.

Russell, P. J. 1995. *Genetics.* Harpercollins College Publisher.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Spirtes, P., Glymour, C., and Scheines, R. 2000a. *Causation, Prediction, and Search.* MIT Press, 2nd edition.

Spirtes, P., Glymour, G., Kauffman, S., Scheines, R., Aimalie, V., and Wimberly, F. 2000b. Constructing Bayesian network models of gene expression networks from microarray data. In *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology.*

Xing, E. P. and Karp, R. M. 2001. Cliff: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* 17, 306–315.

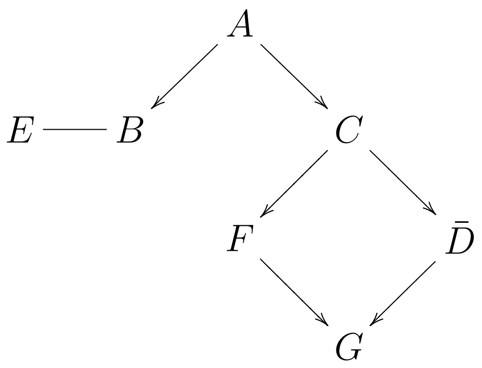


Figure 1: Diagram of a directed acyclic Boolean network with seven elements and twelve pair relations. Only arrows between covering pairs are shown.

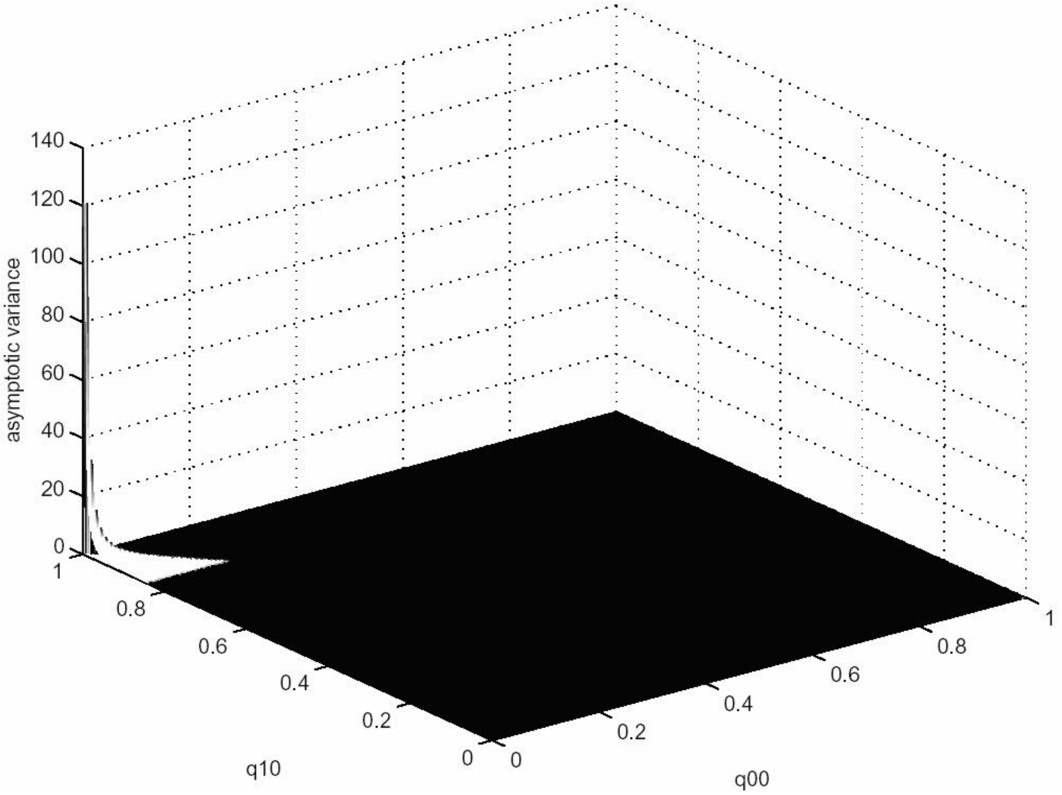


Figure 2: The asymptotic variance of the MLE of p when $p = 0.05$. One singularity point occurs at $q_{10} = 1$ and $q_{00} = q_{11} = q_{01} = 0$.

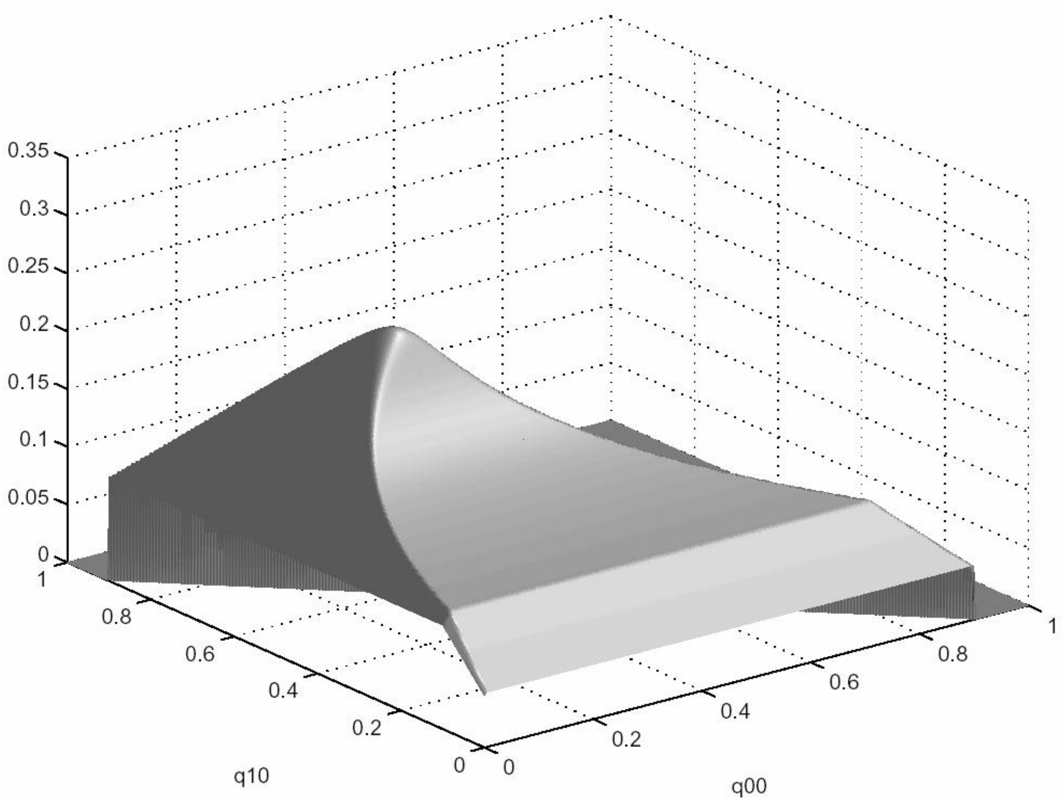


Figure 3: $p_{A \prec B} - p$, where $p = 0.05$ and $q_{01} = 0.1$. It confirms that $p_{A \prec B}$ is larger than p when $q_{01} > 0$.

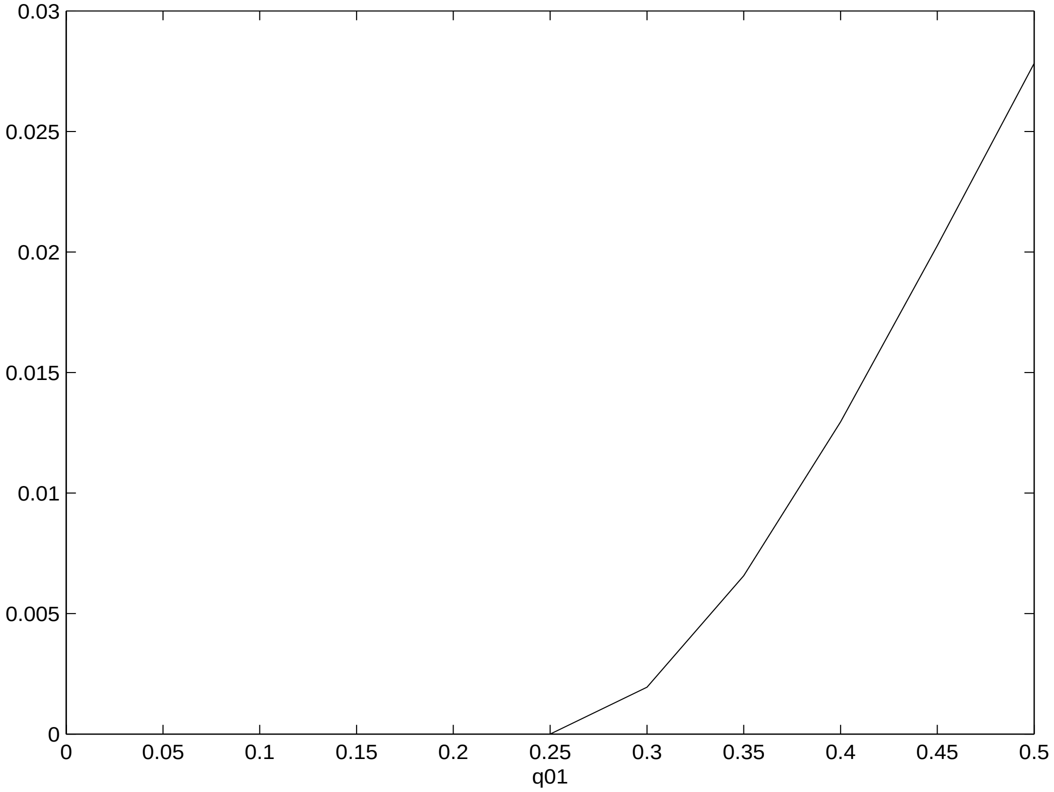


Figure 4: The Kullback-Leibler divergence between the full model $q_{01} > 0$ and the triangular model $q_{01} = 0$ against q_{01}, where $p = 0.05$, $q_{00} = q_{11} = q_{10}$.

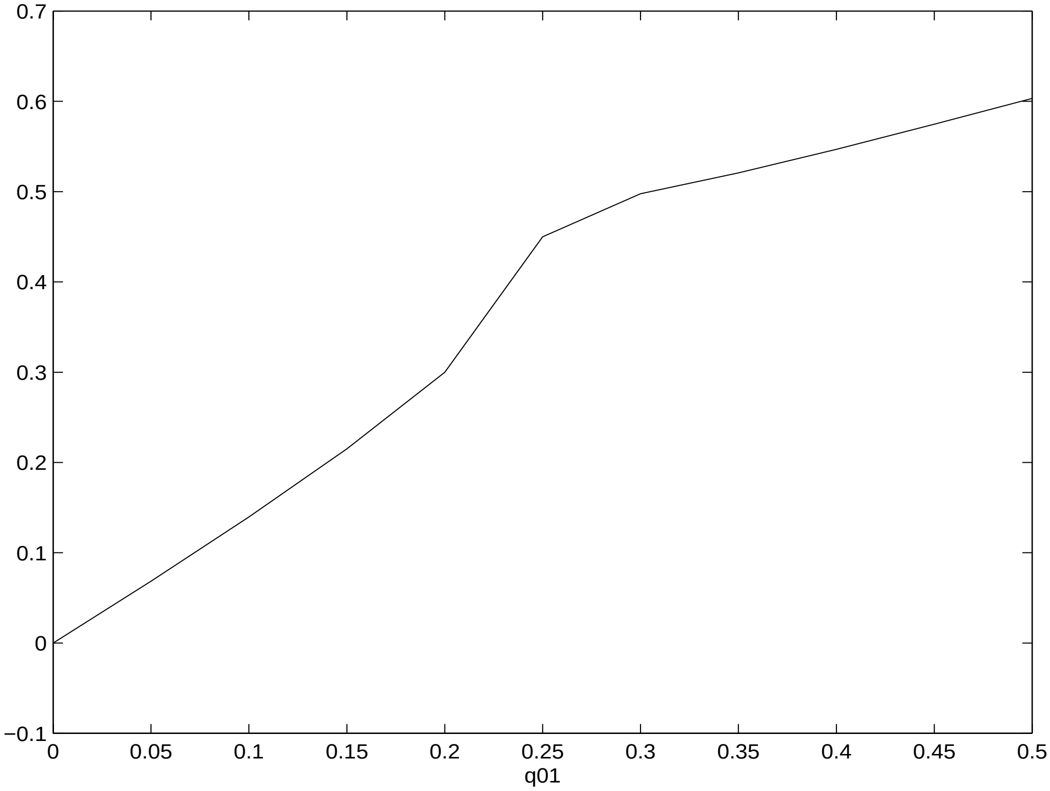


Figure 5: $p_{A \prec B} - p$ against q_{01}, where $p = 0.05$, $q_{00} = q_{11} = q_{10}$.

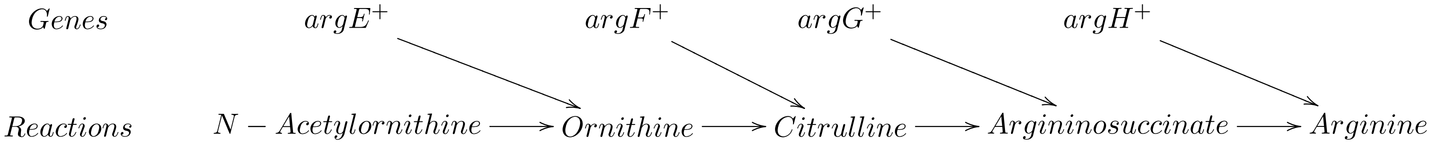


Figure 6: Arginine biosynthetic pathway. The four genes code for the enzymes (not shown) that catalyze each reaction.

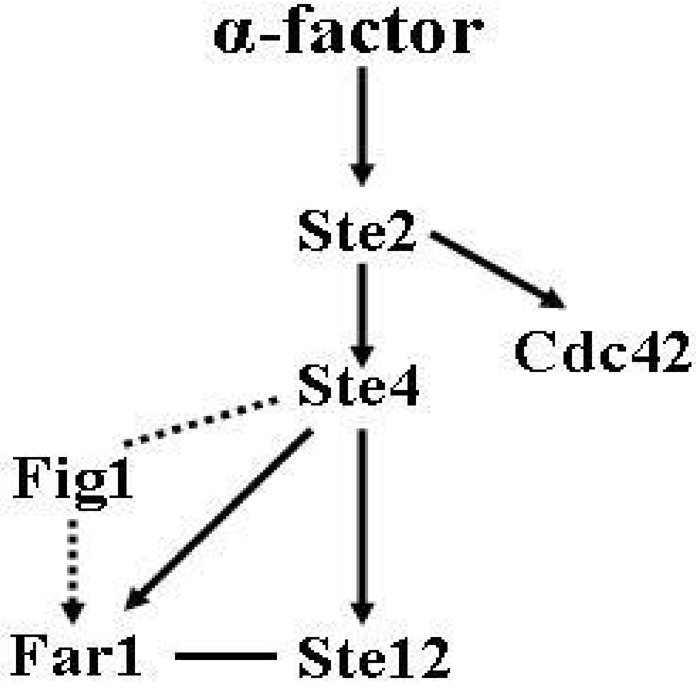


Figure 7: Some pairwise relations identified by s-p-scoring method from the expression data of yeast *Saccharomyces cerevisiae* with knock-out and activation; see Roberts *et al.* (2000).

Table 1: The table of states for directed acyclic Boolean network shown in Figure 1.

case	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0	1	1	1	1	1	1	1	1	1	1	1	1
B	0	0	1	1	1	1	1	1	0	0	0	0	0
C	0	0	0	1	1	1	1	1	1	1	1	1	1
D	1	1	1	1	0	1	0	0	1	0	1	0	0
E	0	0	1	1	1	1	1	1	0	0	0	0	0
F	0	0	0	0	0	1	1	1	0	0	1	1	1
G	0	0	0	0	0	0	0	1	0	0	0	0	1

Table 2: 2×2 tables for a pair of elements assuming no measurement error. The counts on the left are regarded as being generated from the multinomial distribution on the right.

A / B	0	1
0	m_{00}	m_{01}
1	m_{10}	m_{11}

A / B	0	1
0	q_{00}	q_{01}
1	q_{10}	q_{11}

Table 3: Count patterns for the six pairwise relations assuming exhaustive sampling and no measurement error.

$A \sim B$

A / B	0	1
0	+	0
1	0	+

$A \sim \bar{B}$

A / B	0	1
0	0	+
1	+	0

$A \prec B$, $\bar{B} \prec \bar{A}$

A / B	0	1
0	+	0
1	+	+

$\bar{A} \prec \bar{B}$, $B \prec A$

A / B	0	1
0	+	+
1	0	+

$A \prec \bar{B}$, $B \prec \bar{A}$

A / B	0	1
0	0	+
1	+	+

$\bar{A} \prec B$, $\bar{B} \prec A$

A / B	0	1
0	+	+
1	+	0

Table 4: The six pairwise relations, their corresponding probabilistic hypotheses and s-scores, p-scores.

	Relation	Hypothesis	scores
diagonal similarity	$A \sim B$ $\bar{A} \sim B$	$q_{01} = q_{10} = 0$ $q_{00} = q_{11} = 0$	$s_{A \sim B}$ $s_{\bar{A} \sim B}$
triangular prerequisite	$A \prec B$ $\bar{A} \prec \bar{B}$ $A \prec \bar{B}$ $\bar{A} \prec B$	$q_{01} = 0$ $q_{10} = 0$ $q_{00} = 0$ $q_{11} = 0$	$p_{A \prec B}$ $p_{\bar{A} \prec \bar{B}}$ $p_{A \prec \bar{B}}$ $p_{\bar{A} \prec B}$

Table 5: The 2×2 count table for a pair of elements and their generating probabilities in the presence of measurement error.

A / B	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

A / B	0	1
0	r_{00}	r_{01}
1	r_{10}	r_{11}

Table 6: Splitting counts caused by misclassification error.

A/B	0		1	
0	$m_{00,00}$	$m_{00,01}$	$m_{01,00}$	$m_{01,01}$
	$m_{00,10}$	$m_{00,11}$	$m_{01,10}$	$m_{01,11}$
1	$m_{10,00}$	$m_{10,01}$	$m_{11,00}$	$m_{11,01}$
	$m_{10,10}$	$m_{10,11}$	$m_{11,10}$	$m_{11,11}$

Table 7: Splitting probabilities caused by misclassification error.

A/B	0		1	
0	$q_{00,00} = (1-p)^2 q_{00}$	$q_{00,01} = p(1-p)\, q_{00}$	$q_{01,00} = p(1-p)\, q_{01}$	$q_{01,01} = (1-p)^2\, q_{01}$
	$q_{00,10} = p(1-p)\, q_{00}$	$q_{00,11} = p^2\, q_{00}$	$q_{01,10} = p^2\, q_{01}$	$q_{01,11} = p(1-p)\, q_{01}$
1	$q_{10,00} = p(1-p)\, q_{10}$	$q_{10,01} = p^2\, q_{10}$	$q_{11,00} = p^2\, q_{11}$	$q_{11,01} = p(1-p)\, q_{11}$
	$q_{10,10} = (1-p)^2\, q_{10}$	$q_{10,11} = p(1-p)\, q_{10}$	$q_{11,10} = p(1-p)\, q_{11}$	$q_{11,11} = (1-p)^2\, q_{11}$

Table 8: For the DAB in Figure 1, we generate 76 samples, and take $p = 0.05$. The true and false relations (in parentheses) cross each other by only one case.

ranking		hypotheses						counts in cells			
relation	s-p-score	$q_{01} = q_{10} = 0$	$q_{00} = q_{11} = 0$	$q_{01} = 0$	$q_{10} = 0$	$q_{00} = 0$	$q_{11} = 0$	n_{00}	n_{01}	n_{10}	n_{11}
$C \prec G$	0.000	0.441	0.250	0.000	0.441	0.250	0.197	23	0	38	15
$A \prec G$	0.000	0.441	0.138	0.000	0.441	0.079	0.138	6	0	55	15
$A \prec C$	0.017	0.146	0.388	0.017	0.146	0.079	0.388	5	1	18	52
$A \prec \bar{D}$	0.028	0.250	0.329	0.079	0.250	0.028	0.329	1	5	31	39
$A \prec E$	0.030	0.342	0.237	0.030	0.342	0.079	0.237	5	1	41	29
$B \sim E$	0.041	0.041	0.498	0.028	0.041	0.605	0.395	42	2	4	28
$A \prec F$	0.054	0.309	0.270	0.054	0.309	0.079	0.270	4	2	37	33
$F \prec G$	0.058	0.219	0.368	0.058	0.219	0.368	0.197	38	3	23	12
$C \prec \bar{D}$	0.059	0.362	0.231	0.303	0.362	0.059	0.231	3	20	29	24
$A \prec B$	0.060	0.329	0.250	0.060	0.329	0.079	0.250	4	2	40	30
$C \prec F$	0.099	0.244	0.382	0.099	0.244	0.303	0.382	18	5	23	30
$(C \prec E)$	0.112	0.319	0.349	0.112	0.319	0.303	0.349	18	5	28	25
$\bar{D} \prec G$	0.120	0.388	0.257	0.197	0.388	0.257	0.120	23	9	38	6
$(C \prec B)$	0.134	0.319	0.362	0.319	0.134	0.303	0.362	17	27	6	26
$(\bar{E} \prec G)$	0.148	0.296	0.401	0.197	0.296	0.401	0.148	36	10	25	5
$(\bar{B} \prec G)$	0.180	0.309	0.388	0.197	0.309	0.388	0.180	35	9	26	6
$(D \sim \bar{F})$	0.208	0.480	0.208	0.421	0.579	0.187	0.208	11	21	30	14
$(D \sim \bar{E})$	0.301	0.484	0.301	0.394	0.606	0.301	0.288	17	15	29	15
$(B \sim \bar{D})$	0.338	0.500	0.338	0.590	0.411	0.337	0.337	17	27	15	17
$(B \prec F)$	0.360	0.360	0.476	0.360	0.338	0.581	0.419	25	19	16	16
$(E \prec F)$	0.427	0.427	0.419	0.427	0.395	0.419	0.319	24	22	17	13

Table 9: The states of presence in the experiments of growth response.

	presence of elements							
Mutant strains	E	F	G	H	Ornithine	Citrulline	Arginino-succinate	Arginine
Wildtype	1	1	1	1	1	1	1	1
argE	0	1	1	1	0	0	0	0
argF	1	0	1	1	?	0	0	0
argG	1	1	0	1	?	?	0	0
argH	1	1	1	0	?	?	?	0

Table 10: Coding an element with 4 expression levels by two pseudo elements.

Level	0	1	2	3
element A_1	0		1	
element A_2	0	1	0	1