# 國立交通大學

## 資訊科學與工程研究所

## 碩 士 論 文

離散化基因表現資料以重建基因調控網路

Reconstruct transcription networks by discretizing gene expression data

研 究 生：吳秉蔚

指導教授：胡毓志　博士

中 華 民 國　九 十 五　年 六 月

# 離散化基因表現資料以重建基因調控網路
# Reconstruct transcription networks by discretizing gene expression data


研 究 生：吳秉蔚　　Student：Ping-Wei Wu

指導教授：胡毓志　　Advisor：Yuh-Jyh Hu

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所


碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

# 離散化基因表現資料以重建基因調控網路

研究生：吳秉蔚　　指導教授：胡毓志博士

國立交通大學資訊工程研究所

# 摘要

正確地預測生命體中轉錄因子與基因之間的相互關係，也就是所謂的基因調控網路，是生物資訊研究者們正熱烈探討的問題之一。微分方程式或貝氏網路常被應用於此問題的解決，但有鑑於這些方法的繁瑣或計算複雜度較高，本研究基於"轉錄因子與受調控基因之間的關係，確實存在於基因表現資料"的前提，將基因表現資料轉換成字串的形式，並利用字串排比，找出潛在的調控模組。本方法除了所需計算時間非常短之外，亦可解決基因表現中"時間差(Time Shift)"之問題。另外，對於基因表現資料並非全局相似之基因組，我們的方法亦提供預測的可能性。最後，我們以 SCPD 資料庫中的 26 個酵母菌的轉錄因子所組成之調控模組爲預測對象，並將實驗數據與前人之結果作比較，發現本研究方法提升其中 18 個轉錄調控模組的預測能力。

# Reconstruct transcription networks by discretizing gene expression data

Student: Ping-Wei Wu Advisor: Dr. Yuh-Jyh Hu

Institute of Computer Science and Engineering

National Chiao Tung University

Hsinchu, Taiwan, 300, Republic of China

## Abstract

To discover the relations betweens genes, e.g. genetic networks, is one of the prominent topics in bioinformatics. Bayesian networks and differential equations have been widely applied to solving the problems, and yet the complexity of these approaches also limits their performance. We propose a fast and effective method to predict transcription modules, and then to reconstruct genetic network. Our method considers not only the “time shift” issue but also the gene pairs whose expression data are partially correlated. We applied our methods to 26 known transcription modules of yeast, and compared the results with previous works based on SCPD. The results indicate that within 26 transcription modules, our method increased 18 transcription modules’ precision. In addition, we also tested our method for the capability of reconstructing genetic networks, using cell cycle-related genes and transcription factors.

# 致 謝

終於，論文完成了。

感謝我的指導老師，胡毓志，沒有他的指導與建議，我無法完成這本論文。跟在老師身邊的日子，除了專業的知識，老師的做事態度與研究精神，讓我真正見識到所謂學者的風範，而這些將讓我受用不盡。我認爲一位好老師對學生的影響，不會只有在他們相處的日子裡，而胡老師就是一位這樣的老師。海明威曾說：”巴黎，是一場流動的饗宴。”雖然我沒有到過巴黎，但在我接下來的生活裡，我也因此擁有流動的饗宴。

老師謝謝您。

還有實驗室的各位，子緯、鈞木、巽昌、勁伍、音璇、世彥、豐茂、登貴、貫中、繼養，還有陪我奮戰到最後的昀君，我的生活因爲你們而不單調。

感謝身邊的你們

2006 初夏

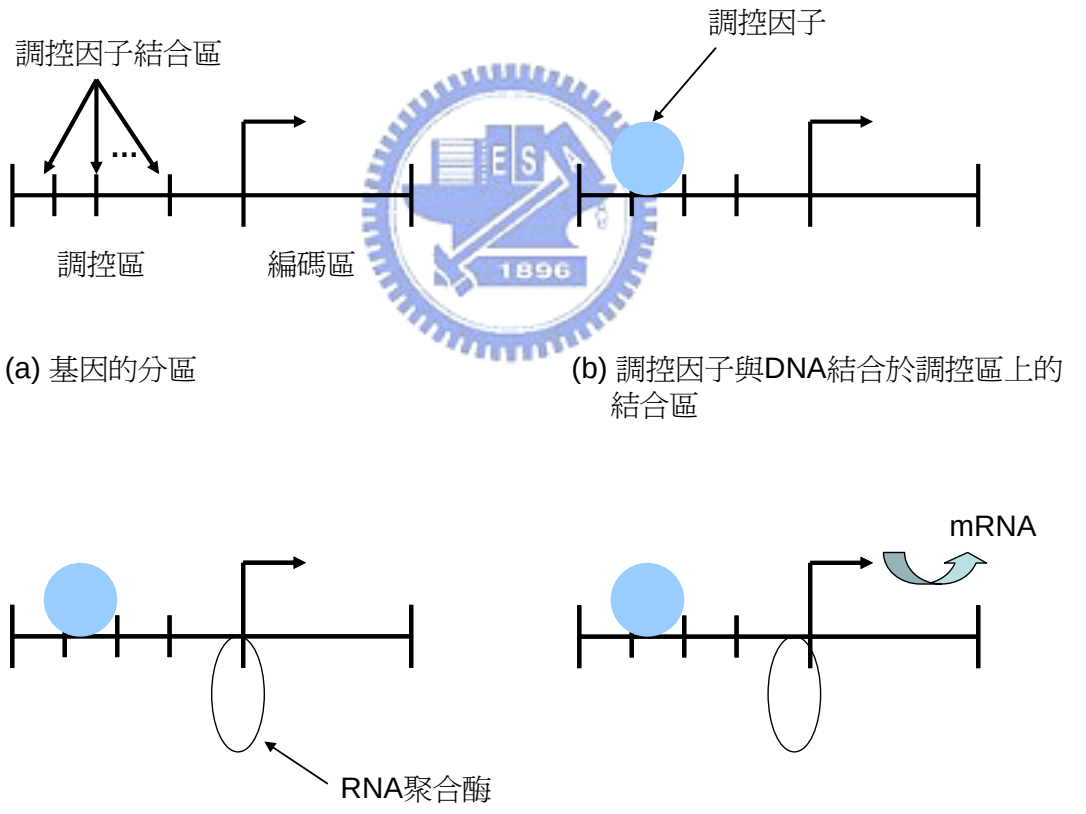
# 目錄

# 第1章　前言

## 1.1 簡介

分子生物學中，所謂的中心法則(central dogma)告訴我們，生命機制的藍圖都存於 DNA 上，經由轉錄成 mRNA 然後再轉譯成蛋白質，進而控制細胞的活動。而這些特定的 DNA 片段(DNA sequence)就是所謂的基因(gene)。

所謂的轉錄(Transcription)作用，指的是由一連串由 DNA 到 mRNA 的生化反應。下圖說明轉錄作用的過程，由基因、轉錄因子、RNA 聚合酶協力完成。一個轉錄因子(transcription factor)是由一個或多個蛋白質所組成，爲其他基因的產物。RNA 聚合酶(RNA Polymerase , RNAP)以 DNA 爲模板，催化合成 RNA。




(a) 基因的分區

(b) 調控因子與DNA結合於調控區上的結合區

(c) 調控因子使RNAP結合

(d) RNAP產生一調控聚合物合成mRNA

圖 1.1-1 轉錄作用的步驟

圖 1.1-1(a)中，是一個基因，也就是一段 DNA 序列。如圖所示，一個基因可以分爲調控區(regulatory region)與編碼區(coding region)。其中調控區含有轉錄因子

(transcription factor, trans-regulatory elements)的結合區(transcription factor binding site, cis-regulatory elements)，提供轉錄因子與此基因結合，進而影響接下來的步驟；而編碼區將用於轉錄成 mRNA，最後轉譯成特定的蛋白質。當調控因子與 DNA 結合(圖 b)，將使得 RNA 聚合酶結合於編碼區的上游區(upstream)(圖 c)。接著，RNA 聚合酶產生一調控聚合物(transcriptional complex)將 DNA 的兩股分開，並沿著其中一股(編碼區)轉錄成 mRNA(圖 d)。

另外所以的啓動子(promoter)，指的就是 RNA 聚合酶結合區與轉錄的起始點；一個轉錄調控模組(Transcriptions module)是由被調控的基因與調控它的轉錄因子所組成；轉錄調控網路則是由多個轉錄調控模組所構成。

從前要得到數千個基因，在某一連續時間內，針對數個不同生理狀況(physiological condition)的表現量，可能需要花上數年的時間，因此想要建構完整的調控網路，雖非天方夜譚，但仍如移山填海。但隨著微矩陣(microarray)、DNA 晶片(DNA chip) …… 等大規模監測或篩選的科技問世後，以往曠日費時的工作，如今已省時不少，但隨之而來的即是一堆龐大的實驗資料。隨後生物資訊學者，開始想辦法利用統計、計算機 ……等方法，從這些資料中擷取出有用的資訊，建構出可能的基因網路，幫助生物學家能先針對較有可能的基因做實驗，加快研究的進度。

## 1.2 研究動機與目的

重建基因網路的研究，因爲微矩陣等技術的提出，而開始蓬勃發展。透過微矩陣等實驗，我們可以快速地得到大量的基因基於不同的生化條件之下的基因表現。我們相信這些資料必定隱含了生物學家們感興趣的問題之一 —— 基因之間的關係。
如何透過現有的資料，以及生物、計算機、統計 …… 等跨領域的技術，使重建基因網路成爲彈指間就能完成的工作，正是催生本研究重要的動機之一。

針對基因表現實驗之 time shift 問題，Min Zou 等人(Zou et al.2005)提出依簡單但合理之方法；Kwon 等人(Kwon et al.2003)等將基因表現資料離散化，用以判斷兩基因的關連性；Hsu 等人(Hsu et al. 2004)利用線性迴歸建構調控網路，並以 PF 值爲評分

標準；基於前人的研究方法(見第二章)以及本文的實驗假設(見 3.1)，我們提出一調控網路的建構流程，並以酵母菌(Saccharomyces cerevisiae)的基因表現資料爲研究對象，建構出已知的調控關係，進而提出尚未被證明，但存在可能性較高的調控模組。

## 1.3 論文架構

關於接下來的章節：

第二章爲文獻探討。針對目前幾個用於重建基因網路的方法做介紹。

第三章爲研究方法與實驗設計。爲本研究的核心，除了說明本實驗所基於的假設、基因表現資料的來源、前置處理，更詳細地說明本實驗的方法及流程。

第四章爲實驗結果。除了包含本研究所得到的結果之外，並將其與前人的結果交叉比對。

第五章則針對我們所以的方法做總結，提出關於重建基因網路之未來研究的方向。

# 第2章　文獻探討

本章由重建調控網路相關的研究方向與方法做切入。2.1 說明分群的目的，以及對重建基因網路的幫助，並針對幾個常用的分群法加以描述；2.2 介紹幾個重建調控網路常用的方法與模型；2.3 探討相關的研究爲了提高精確度所採取的策略；2.4 針對重建調控網路研究中，幾個其他待解決的課題，進行討論。

## 2.1 在重建基因調控網路之前

基因相關的研究中，針對一個基因來說，人們想知道的是該基因有什麼功能？哪些基因會調控它？這樣的基因(缺陷)會造成什麼疾病？有什麼藥物可以治療？……從前研究人員花了大量的時間，確定部分基因的功能及相互關係，但跟未知功能基因的數量比較起來，這些發現不過是滄海一粟。有了微矩陣所得之實驗資料，我們希望透過已知其功能之基因，快速找出相類似的基因，再進行其他生物上的研究分析。因此，如何做好分群(clustering)成爲生物資訊學者關心的方向之一。生物(資訊)研究人員希望透過分群：

<1> 找出具類似功能、受某些調控因子共同調控知基因；

<2> 在重建基因調控網路的研究中，分群有助於縮小搜尋空間(search space)，加速調控網路的建立。

常用的分群演算法可分爲兩大類：階層式分群(hierarchical)與非階層式分群(non-hierarchical)。其中階層式分群所分出的各群，是由其他子群所組合而成。階層式分群根據建構方式的不同，又可分爲合併聚類(agglomerative clustering, bottom-up)以及分離聚類(divisive clustering, top-down)。合併聚類一開始將 N 個物件視爲 N 群，然後再逐步將這 N 群結合成 K 群；反之，分離聚類一開始將 N 個物件視爲同一群，然後再逐步將這一群分解成 K 群。

而非階層式分群是將 N 個物件分爲 K 群，群跟群之間沒有相乎依屬的關係。常見的演算法有 K-means(Tavazoie et al. 1999)、Self-Organized Map(SOM)(Tamayo et al. 1999)與 EM alfogrithm(expectation　maximization algorithm)。

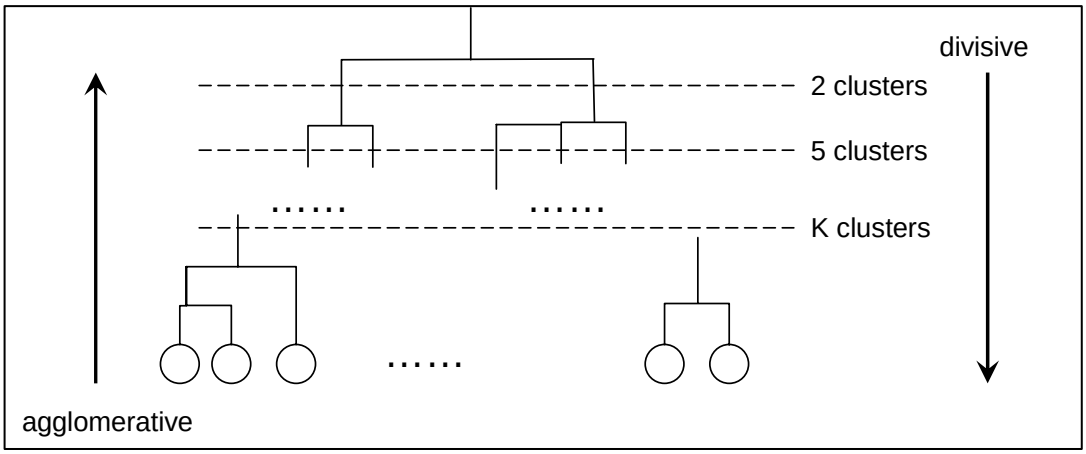



圖 2.1-1　階層式分群的建構方式。由上而下的建構方式稱爲分離聚類(divisive clustering)；由下而上的建構方式稱爲合併聚類(agglomerative clustering)；

### 2.1.1 階層式分群

#### 2.1.1.1 合併聚類(agglomerative clustering)

Eisen 等人(Eisen et al. 1998)利用可得的基因表現資料，以合併聚類將基因分群。步驟如下：

<1> 一開始，N 個基因分屬不同的 N 個群；

<2> 計算任兩群$C_p$,$C_q$之間的距離dist($C_p$,$C_q$)；$C_p$之群基因表現P，以其所有成員之基因表現平均值爲代表，其中

$$P=(p_1,p_2,\ldots,p_t)\ where\ p_j=\sum_{i=1}^{n}\frac{x_{ij}}{n}\ ,\forall j=1,\ldots t$$

其中n爲$C_p$中的基因個數；$X_{ij}$爲基因i在時間點j之表現量；同理，$C_q$之群基因表現 $Q=(q_1,q_2,\ldots q_t)$ ；

<3> 計算 P,Q 之歐幾里得距離(Euclidean distance)作爲群之間的距離：

$$dist(C_p,C_q)=\sqrt{\sum_{i=1}^{t}(p_i-q_i)^2}\quad ,p_i\in P,q_i\in Q\ \ \forall i=1\ldots t$$

<4> 重複步驟<2>到<3>N-1 次，我們可以得到如圖 2-1 之階層式分群樹。

### 2.1.1.2 分離聚類(divisive clustering)

Alon 等人(Alon et al. 1999)利用分離聚類對結腸癌(colon cancer)的基因表現資料做分群。步驟如下：

<1> 一開始，所有的基因都歸於同一群；

<2> 利用下列方程式，定義出兩個新的群中心(質心基因)，Cj，j=1,2

$$P_j(V_i) = \exp(-\beta\left|V_i - C_j\right|^2) / \sum_j \exp(-\beta\left|V_i - C_j\right|^2)$$

$$C_j = \sum_{i=1}^{n} V_i P_j(V_i) / \sum_{i=1}^{n} P_j(V_i)$$

<3> 利用上述公式，對每一個基因i分別求出$P_1(V_i)$、$P_2(V_i)$，若$P_1(V_i)>P_2(V_i)$，則基因i將被分到第 1 群；反之，分到第 2 群；

<4> 對所得的各群，重複步驟<2>和<3>，直到達到停止條件。

無論使用階層式分群或是分離聚類，最後建構所得的分群樹都與圖 2.1-1 類似。至於究竟該如何決定最適當的分群法，就因不同的應用而取決於使用者。

## 2.1.2非階層式分群

### 2.1.2.1 K-means

K-means 由 MacQueen 等人提出。以 N 組基因表現資料來說明，K-means 的目的是將這 N 個基因分爲 k 群，使得 V 爲最小值：

$$V = \sum_{i=1}^{k} \sum_{j \in S_i} \left|x_j - \mu_i\right|^2$$

其中{ Si, … ,Sk }爲 k 個群；$x_i \in S_j$，爲某基因之表現量，μi 爲第 i 群之質量中心(centroid)基因。透過下列步驟，我們可以求得 V 與{ Si, … ,Sk }：

<1> 給定 k 個質心基因，或任取 k 個基因作爲質心基因；

<2> 對於每個基因，計算其與此 k 個質心基因的距離，並將此基因分配至與之距離最近之質心基因該群；

<3> 針對每一群，重新計算群內兩兩基因之距離，並重新選出一質心基因，使得該基因與群內其他基因之距離爲最小；

<4> 重複<1>到<3>，直到收斂，即各群內的基因與前一次相同。

由上述流程，我們可以發現 K-mean 的幾個缺點：

<1> 無法預知步驟<1>到步驟<3>需要重複多少次(需要多少時間)；

<2> 收斂時的 V 以及{ Si, … ,Sk }不一定是最佳解，因爲最佳解的品質跟一開始選定的 k 個質心基因有很大的關係；

<3> 使用者必須自行決定 k 的大小，但是在實際應用上，使用者往往也無法知道 k 應設爲多少以求得最佳解。

## 2.2 重建基因調控網路的方法

### 2.2.1 線型模型

#### 2.2.1.1 線性回歸

在眾多重建基因網路的方法中，線性模型的概念較爲直覺。其假設爲：”受調控基因於某時間點之表現，會受到調控其基因於該時間點之前之基因表現影響。而影響的方式可以以線性的方式表示。”基本的概念可用下列方程式表示：

$$X_j(t) = \sum_{i=1}^{N} \left( W_{i,j} \cdot X_i(t-1) \right), \quad X_j, W_{i,j} \in \Re \qquad (2.1)$$

其中$X_j(t)$爲基因$X_j$於時間點t之表現程度；N爲所有基因個數；W爲一N×N矩陣，記錄基因間相互的關係。利用此方程式來建構基因網路，其目標爲找出最適當的W，以決定基因$X_i$與基因$X_j$之間是否具有調控關係。

Hsu 利用此一方法，求出調控基因與被調控基因之間的估計回歸方程式 $\hat{Y} = WX + \varepsilon$ ，並以統計檢定量之 p-value(PF 值)來決定此方程式的可靠程度。

### 2.2.1.2線性皮爾森相關係數

上述線性模型中，我們可以觀察到幾個利用此方程式所會面臨到的問題：

**<1> 時間差固定，與生物現象不符；**

方程式(2.1)使用的時間差(time lag)固定爲 1，但根據生物實驗的觀察，這樣的假設並不完全成立。因此，爲了增加預測模型的彈性，方程式 2.1 可以改寫爲：

$$X_j(t+\Delta t)=\sum_{i=1}^{N}\left(W_{i,j}\cdot X_i(t)\right),\qquad X_j, W_{i,j}\in\Re \qquad (2.2)$$

$\Delta t$ 爲所謂的 time lag。在大部分的情況中，受調控基因之基因表現會較調控基因之基因表現爲晚，而其間的時間差即爲$\Delta t$。決定$\Delta t$，亦是重建基因網路研究中，一個重要的方向(Ji et al. 2005；Zou et al. 2005；Liu et al. 2004)]。2.3 節中，將會介紹幾個前人用於決定Δt 的方法。

**<2> 需要大量的計算時間；**

因此，van Someren 等人(Someren et al. 2000)將基因分群，將表現相似之基因視爲同一群，以減少 search space 的大小，進而降低計算時間。

**<3> 生物實驗中，基因間的關係是否以線性的方式相互影響？**

根據前人的實驗結果，單獨使用線性模型重建基因網路，的確可以找出部分已知的基因關係，但精確度仍過低。

D'Haeseleer(D'haeseleer et al. 1999)與 Kuruvilla(Kuruvilla et al. 2002)以皮爾森相關係數(Linear Pearson correlation)來重建調控網路；Hsu(Hsu et al. 2004)以線性迴歸(Linear regression)來決定兩基因的相關性。這些方法認爲，每段基因的表現程度可由其他基因的表現程度以線性方程式表達；另外這些方法偏好基因表現資料具全局相似(global similarity)，對於只具有區域相似(local similarity)的基因表現，其表現較差。根據我們所提出的方法(見第三章)，對於全局相似或區域相似，我們都能順利地判斷基因間的相關性。

### 2.2.2 布林網路

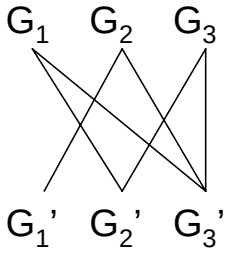
布林網路是以邏輯關係來描述基因調控網路。其假設如下：

<1> 每一段基因的表現程度可以分爲高(表現)、低(不表現)兩種層次；

<2> 每一段基因的表現程度可能由其他基因的表現程度以布林函式來決定。

圖 2.2-1 中，爲一將布林網路運用在重建基因網路之例子。其中基因$G_1$,$G_2$,$G_3$爲可能之調控基因；$G_1$',$G_2$',$G_3$'爲可能之被調控基因。(c)爲$G_1$,$G_2$,$G_3$ ,$G_1$',$G_2$',$G_3$'轉換過後的基因表現資料；$T_0$-$T_7$爲連續之時間序列；(a)爲所得的布林網路。其中基因$G_1$'之表現受$G_2$影響；基因$G_2$'之表現受$G_1$或$G_3$影響；基因$G_3$'之表現受$G_1$,$G_2$同時影響，或者受$G_2$,$G_3$同時影響，或者受$G_1$,$G_3$同時影響；(b)與(a)同義，爲其對應之邏輯關係。

(a)




(c)

| | $G_1$ | $G_2$ | $G_3$ | $G_1$' | $G_2$' | $G_3$' |
|---|---|---|---|---|---|---|
| $T_0$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $T_1$ | 0 | 0 | 1 | 0 | 1 | 0 |
| $T_2$ | 0 | 1 | 0 | 1 | 0 | 0 |
| $T_3$ | 0 | 1 | 1 | 1 | 1 | 1 |
| $T_4$ | 1 | 0 | 0 | 0 | 1 | 0 |
| $T_5$ | 1 | 0 | 1 | 0 | 1 | 1 |
| $T_6$ | 1 | 1 | 0 | 1 | 1 | 1 |
| $T_7$ | 1 | 1 | 1 | 1 | 1 | 1 |

(b)

$G_1' = G_2$

$G_2' = G_1$ or $G_3$

$G_3' = (G_1$ and $G_2)$ or $(G_2$ and $G_3)$ or $(G_1$ and $G_3)$

圖 2.2-1 一個布林網路的例子。(a)爲所得之網路架構；(b)爲(a)對應之邏輯關係；(c)爲建構網路所需的資料。

Liang 等人(Liang et al. 1998)提供一結合 Shannon Entropy 與 Mutual Information 建構布林網路之方法。其中 Shannon Entropy 用於描述一隨機變數或事件之亂度(Entropy)，常用於統計、資訊理論和熱動力學等領域。Fuhrman 等人(Fuhrman et al. 2000)將此觀念應用於新藥物的開發；Cunningham 等人(Cunningham et al. 2000)藉由基因的 Shannon Entropy 高低，來篩選可能的 toxicity target。

Shannon Entropy(H)定義爲：

$$H(X) = -\sum p(x)\log p(x)$$
$$H(X,Y) = -\sum p(x,y)\log p(x,y)$$

Mutual Information 用於描述兩隨機變數之間的相互依賴程度，Butte 等人(Butte et al. 2000)將此觀念應用於基因之分群。

Mutual Information(M)定義爲：

$$M(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$M(X,[Y,Z]) = H(X) + H(Y,Z) - H(X,Y,Z)$$

以前述例子爲例，此方法的目標是針對$G_1$',$G_2$',$G_3$'分別求出影響其基因表現之基因組合，$I_1$,$I_2$,$I_3$。條件爲：

$M(G_1',I_1)/H(G_1')=H(G_1')$、

$M(G_2',I_2)/H(G_2')=H(G_2')$、

$M(G_3',I_3)/H(G_3')=H(G_3')$

以$I_2$來說明：

第一步只考慮單一基因影響，即$I_2$={$G_1$,$G_2$,$G_3$}，根據公式，我們可以求得：

**$H(G_2') = 1.00$**

| | | |
|---|---|---|
| $H(G_2',G_1) = 1.50$ | $M(G_2',G_1) = 0.31$ | $M(G_2',G_1) / H(G_2') = 0.38 \neq 1.00$ |
| $H(G_2',G_2) = 1.81$ | $M(G_2',G_2) = 0.00$ | $M(G_2',G_2) / H(G_2') = 0.00 \neq 1.00$ |
| $H(G_2',G_3) = 1.50$ | $M(G_2',G_3) = 0.31$ | $M(G_2',G_3) / H(G_2') = 0.38 \neq 1.00$ |

發現當$I_2$={$G_1$,$G_2$,$G_3$}時，皆無法使$M(G_2',I_2)/H(G_2')=H(G_2')$。因此考慮

$I_2$={ ($G_1$|$G_2$),($G_2$|$G_3$),($G_1$|$G_3$) }：

| | | |
|---|---|---|
| $H(G_2',[G_1,G_2])=2.50$ | $M(G_2',[G_1,G_2])=0.31$ | $M(G_2',[G_1,G_2]) / H(G_2') = 0.38 \neq 1.00$ |
| $H(G_2',[G_2,G_3])=2.50$ | $M(G_2',[G_2,G_3])=0.31$ | $M(G_2',[G_2,G_3]) / H(G_2') = 0.38 \neq 1.00$ |
| $H(G_2',[G_1,G_3])=2.50$ | $M(G_2',[G_1,G_3])=0.81$ | $M(G_2',[G_1,G_3]) / H(G_2') =$ **1.00** |

發現當$I_2$=$G_1$|$G_3$時，$M(G_2',I_2)/H(G_2')=H(G_2')$。因此決定$I_2$=$G_1$|$G_3$，即$G_2$'=$G_1$ or $G_3$。

### 2.2.3 貝氏網路

Murphy 等人(Murphy et al. 1999)首先將貝氏網路應用於重建基因網路。貝氏網路爲 graphical models 之一。由結點(vertices, nodes)與邊(edges, links)所組成。其中每一個結點代表一個可觀察變數(如某基因之基因表現)，而邊則代表變數間的某種因果關係。若兩點之間有邊相連，則表示此兩點間具此種因果關係。貝氏網路中，結點上的機率，只會受和該相連的父結點影響。如圖 2.2-2。

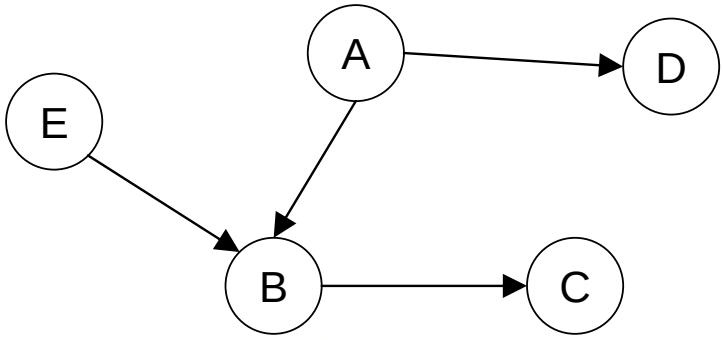



圖 2.2-2 爲一個貝氏網路的結構。

根據這樣的結構，我們可以得知：

***I(A;E), I(B;D\|A,E), I(C;A,D,E\|B), I(D;B,C,E\|A), I(E;A,D)***

上述幾個條件獨立的關係。其中 ***I(X;Y\|Z)***表示在給定 ***Z*** 的條件下，***X*** 獨立於 ***Y***。

圖形G中，一個結點$X_i$其條件機率爲$P(X_i|Pa_i^G)$，其中$Pa_i^G$爲$X_i$之父結點所形成的集合。因爲$X_i$的變化只會受其父結點影響。根據連鎖率，任何聯合機率，在滿足貝氏網路的假設下，可分解爲

$$P(X_1,...,X_n)=\prod_{i=1}^{n}P(X_i \mid Pa^G(X_i))$$

例如，其上圖的聯合機率爲：

***P(A, B, C, D, E) = P(A)P(B\|A,E)P(C\|B)P(D\|A)P(E)***

貝氏網路的建構流程是先決定網路結構，再以實驗資料評估該結構的事後機率。所以一個包含$K$個結點的網路，存在$2^K$種可能的結構，因此計算時間的需求往往爲人詬病，因此Min Zou(見 2.3.1)提出一前處理，減少候選的結點。

## 2.3 結合多種資訊以提高預測的精確度

在第一節所提的方法，線性模型、貝氏網路，都是建構在數學、統計理論基礎完備的方式上，但是關於生物領域的問題，有句名言：”例外永遠存在”。因此單純使用這些數學模型來重建基因網路，往往無法得到令人滿意的結果。許多的研究開始結合其他生物上的資訊或其他的研究方法與評分方式，試圖重建出在生物上而非統計上有意義的基因網路。

Imoto 等人將蛋白質—蛋白質、蛋白質—DNA 的互相影響的資訊以及調控因子結合區等資訊加入貝氏網路；Eran Segal 等人(Segal et al. 2001)將啓動區序列等資訊加入其機率模型，重建基因網路；Lee 等人(Lee et al. 2005)將生物上的註解(biological annotation)加入貝氏網路；Hsu 也將調控因子結合區納入其系統的考量。

因此，我們在實驗時(見第四章)，分別做了幾項實驗：

<1> 單獨使用本研究方式；

<2> 以本研究方式爲前處理，並以 TNP 進行後階段的預測；

## 2.4 一些待解決的問題

### 2.4.1 將時間差納入考量

由於生物實驗中，受調控基因的表現變化常常需要一段時間才會表現出來，因此，是否將這段時間差納入考量，成爲一個網路重建方法是否健全的關鍵因素之一。

在線性皮爾森相關係數(2.1.1.2)中，基因表現間的時間差可由下列方式決定：(Schmitt et al. 2004)

$$c_{ij} = \max_{\tau} \left| r_{ij}(\tau) \right|$$

$$where \quad r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii}(\tau) S_{jj}(\tau)}}$$

$$where \quad S_{ij}(\tau) = \left( x_i(t) - \bar{x}_i \right) \left( x_j(t+\tau) - \bar{x}_j \right)$$

其中τ爲欲知的時間差；xi(t)爲在時間點 t 基因 i 的基因表現；$\bar{x}_i$ 爲基因 i 基因表現的平均值。對於兩基因 i 與 j，當求得 cij 時，τ即爲基因 i 與基因 j 之間的時間差。

Zou(data et al. 2005)提出一直覺但合理的判斷時間差的方法，如圖 2.4-1。若於時間點 T 數值高於 1.2 或低於 0.7，則視 T 爲此基因之開始變化點。基因 G1 其開始變化點必須早於基因 G2，則調控模組(G1,G2)才能成爲候選，進入之後的預測階段。也就是說， Zou 所建構的網路，其調控基因的開始變化點一定較受調控基因之開始變化點早。

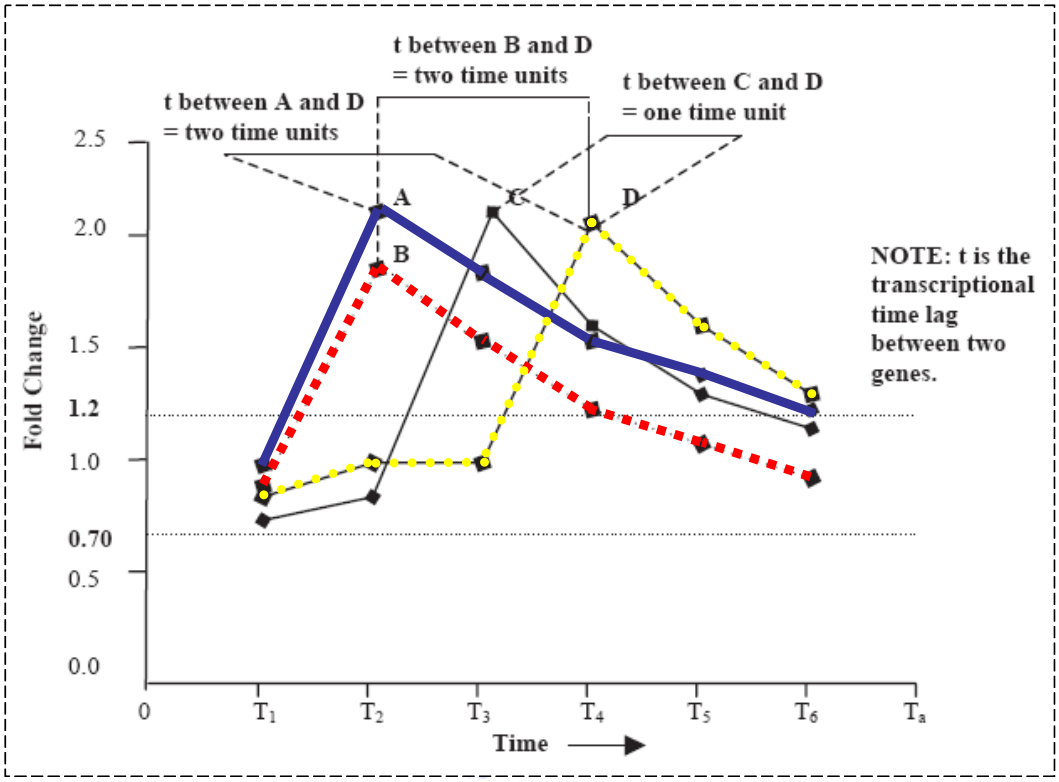



圖 2.4-1 A、B、C、D的開始變化點(initial change)依序爲$T_2$、$T_2$、$T_3$、$T_4$。

這樣的假設的對於大部份的情況似乎成立，並且有效地降低動態貝氏網路預測所需的時間，但是 Min Zou et al 宣稱，預測結果有少部分的遺漏也是由於這樣的假設。在我們建構的系統中，我們將這項個限制交給使用者去決定，亦即使用者可以根據不同的需求而決定，在預測的過程中是否需要啓動這項限制條件。

# 第3章　研究方法與實驗設計

在本章中，我們將詳細地說明本研究所使用的方法與整個實驗流程。

圖爲本研究的流程。流程中各個步驟將於以下各節中介紹。

3.1. 提出本研究所基於的假設。

3.2. 關於實驗資料的前處理。目的在於補足(內插)原始基因表現資料的遺失。

3.3. 將基因表現資料轉換爲足以表現基因變化程度的”slope-data”。試圖找出該基因在該組實驗中，變化較劇烈的時間點。如圖 3.0-1(b)。

3.4. 將 affected-data 的數值(連續型態)轉換爲若干個字元(離散型態)，以方便接下來的實驗步驟。如圖 3.0-1 (c)。

3.5. 有了 3.4 所得的字串，對於任兩基因之間，我們可以將代表其基因變化的字串做排比，判斷此二基因是否有調控關係。如圖 3.0-1 (d)。

3.6. 重建調控網路。經過多次上述幾個步驟，我們可以得到許多由調控因子跟被調控基因所形成的子網路，將這些子網路組合起來，即可得到一較完整的調控網路。如圖 3.0-1 (e)。

3.7. 研究方法討論。討論本實驗方法的特點。如圖 3.0-1 (f)。

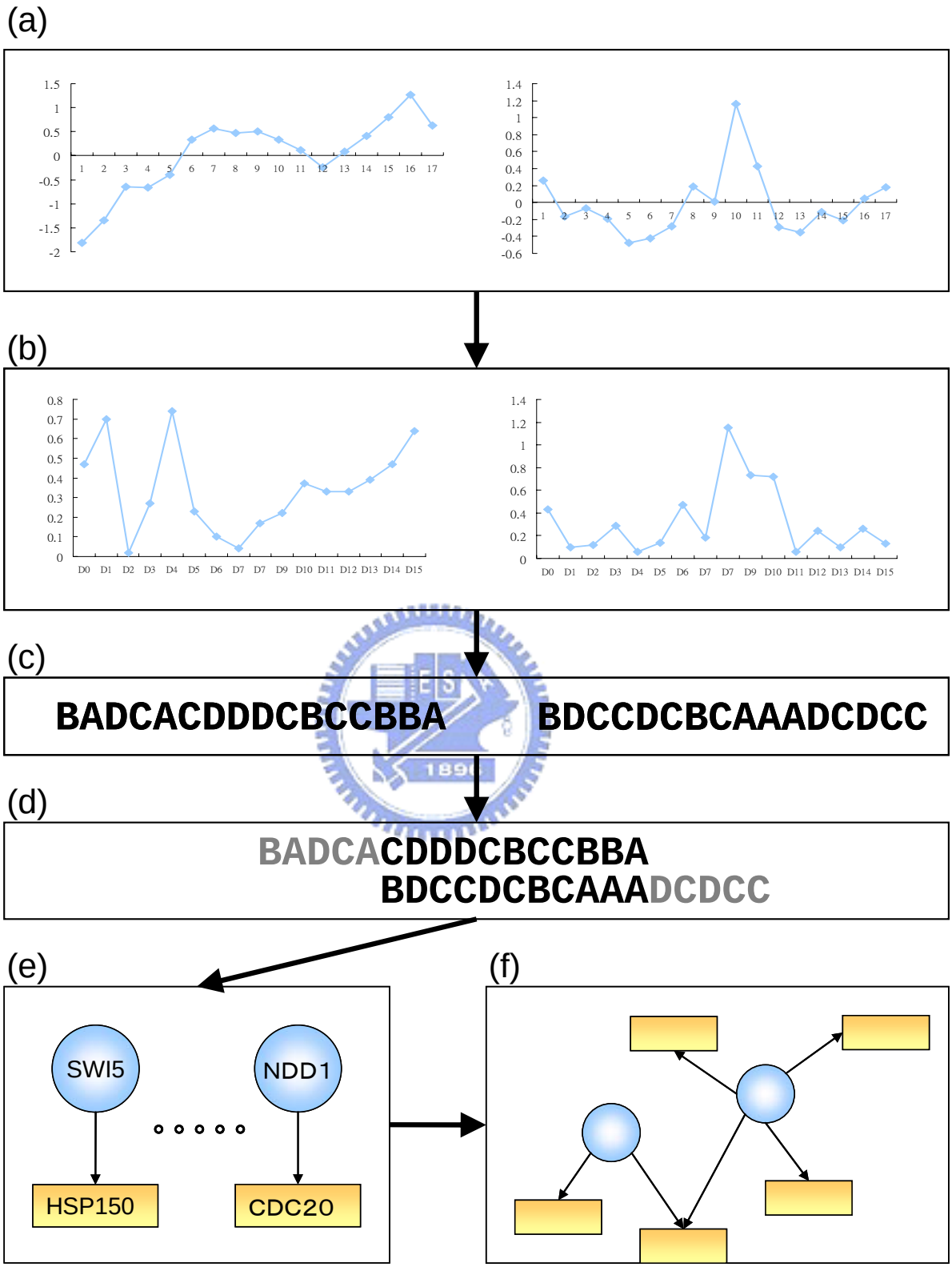



圖 3.0-1 本實驗之實驗流程。(a)爲基因之實驗表現資料；(b)轉換過所得之 slope-data；(c)將 slope-data 轉爲字串；(d)兩字串排比之結果；(e)經過字串排比之後，分數高於平均之調控模組；(f)由調控模組建構出基因調控網路。

## 3.1 研究假設

本研究所使用的方法是基於下列假設：

<1> 調控因子與受調控基因之間的關係，確實存在於基因表現資料(來自微矩陣實驗)。

<2> 基因產物的行爲，大部份決定於基因表現程度。因此，構成轉錄因子的基因，其基因表現程度與轉錄因子的行爲相關。

<3> 轉錄因子所調控的基因，其基因表現程度會受轉錄因子影響；間接的，也會與構成轉錄因子的基因之表現程度相關。

<4> 調控基因與受調控基因在基因表現資料之走勢不必全然相同或相反，但必然存在對應的變化片段。

上述假設<1>，所有重建基因網路，或與調控關係相關的研究，只要有用到微矩陣實驗資料的實驗流程，都必然基於此一假設之上。

<2><3>則是許多重建轉錄調控網路的研究所會用到的假設。

另外，在大部分類似的研究中，通常假設調控基因與受調控基因，其基因表現資料在同一實驗中，必然完全相同或完全相反。也就是說，若在調控基因的基因表現資料中存在一段向上的反應，則必然可以在受調控基因的表現資料中，找到相對應的向上趨勢；反之亦然。但是經由觀察發現，這樣的假設並不全然成立，因此重新定義新的假設<4>，期望得到較前人良好的結果。

## 3.2 基因表現資料的前處理

由於微矩陣實驗常受到環境、儀器、人爲等因素影響，可能造成實驗數據無法判讀，導致基因表現資料在某些時間點的遺失，如圖 3.2-1 中，基因YJL043W之表現資料，於時間點$T_4$、$T_5$、$T_{12}$遺失。但我們不希望因爲這些少數的遺失而完全放棄對該基因實驗的採用，因此我們使用cubic spline，試圖還原這些遺失的資料，如圖 3.2-2，我們將遺失的各點還原。

在我們的前處理中，只要資料遺失的時間點少於 50%，我們就使用 cubic spline 進行內插。

| Gene Name | $T_0$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ |
|---|---|---|---|---|---|---|---|---|
| YJL043W | -0.39 | 0.42 | 0.19 | 0.42 | *NA* | *NA* | -1.81 | -0.58 |

| $T_8$ | $T_9$ | $T_{10}$ | $T_{11}$ | $T_{12}$ | $T_{13}$ | $T_{14}$ | $T_{15}$ | $T_{16}$ |
|---|---|---|---|---|---|---|---|---|
| -0.58 | 1.19 | 0.31 | 0.52 | *NA* | 0.42 | 0.78 | -0.81 | -0.07 |

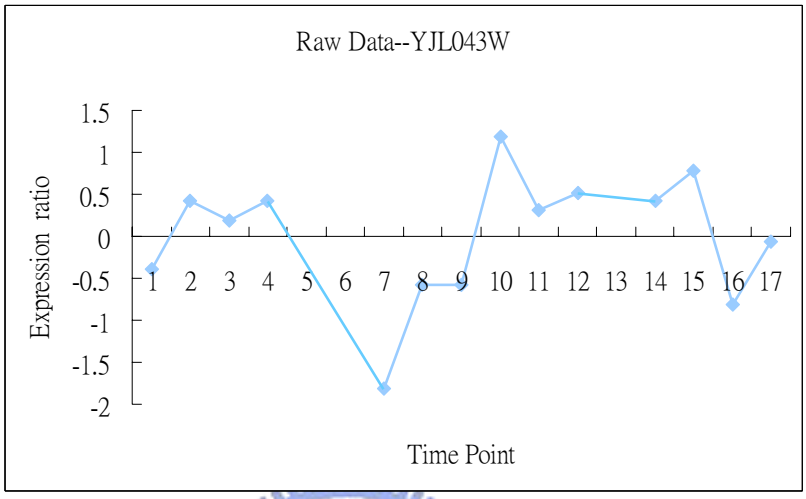



圖 3.2-1 YJL043W 原始的基因表現資料及走勢

| Gene Name | $T_0$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ |
|---|---|---|---|---|---|---|---|---|
| YJL043W | -0.39 | 0.42 | 0.19 | 0.42 | -0.45 | -1.7 | -1.81 | -0.58 |

| $T_8$ | $T_9$ | $T_{10}$ | $T_{11}$ | $T_{12}$ | $T_{13}$ | $T_{14}$ | $T_{15}$ | $T_{16}$ |
|---|---|---|---|---|---|---|---|---|
| -0.58 | 1.19 | 0.31 | 0.52 | 0.43 | 0.42 | 0.78 | -0.81 | -0.07 |

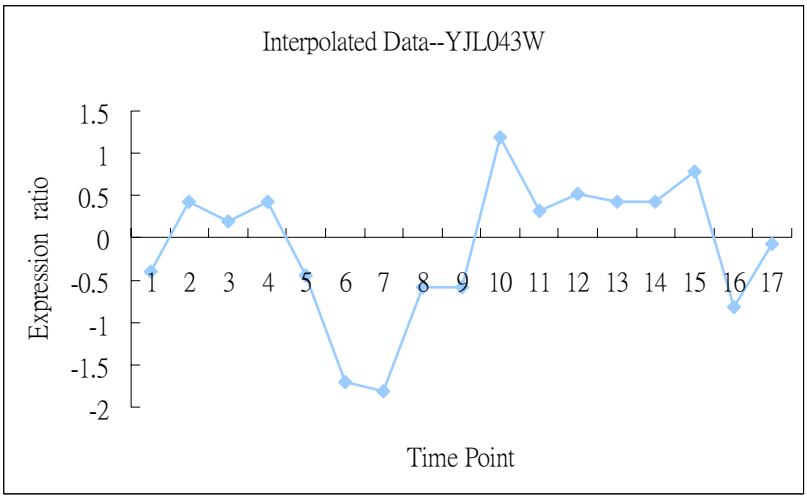



圖 3.2-2 經由 cubic spline 內插後，YJL043W 的基因表現資料及走勢

## 3.3 計算 slope-data

在此一步驟中，我們將基因表現資料轉換爲較能突顯此基因變化的 slope-data。方法爲計算基因表現資料中，相鄰兩點的數值變化。因爲大部分的研究都認爲，調控關係發生在這些變化較爲劇烈的時間點上。

基於實驗假設<4>，我們的計算方式與前人不同。我們只考慮兩點間變化的絕對值(即變化強度)，而不考慮其正、負號(變化的方向)。方法如下：

針對某一基因於某一實驗，若其原始資料爲

$$D = (d_i),\ \forall_i = T_1 \sim T_N,\ T_1, \ldots, T_N \text{爲此實驗之時間點}$$

則計算所得之 slope-data 爲

$$D' = (d'_j),\ \text{其中}\ d'_j = |d_{j+1} - d_j|,\ \forall_j = T_1 \sim T_{N-1}$$

| Gene Name | $T_0$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ | $T_6$ | $T_7$ |
|---|---|---|---|---|---|---|---|---|
| YJL043W | -0.39 | 0.42 | 0.19 | 0.42 | -0.45 | -1.7 | -1.81 | -0.58 |

| $T_8$ | $T_9$ | $T_{10}$ | $T_{11}$ | $T_{12}$ | $T_{13}$ | $T_{14}$ | $T_{15}$ | $T_{16}$ |
|---|---|---|---|---|---|---|---|---|
| -0.58 | 1.19 | 0.31 | 0.52 | 0.43 | 0.42 | 0.78 | -0.81 | -0.07 |

| Gene Name | $T_0T_1$ | $T_1T_2$ | $T_2T_3$ | $T_3T_4$ | $T_4T_5$ | $T_5T_6$ | $T_6T_7$ | $T_7T_8$ |
|---|---|---|---|---|---|---|---|---|
| YJL043W | 0.81 | 0.23 | 0.23 | 0.87 | 1.25 | 0.11 | 1.23 | 0.0 |

| $T_8T_9$ | $T_9T_{10}$ | $T_{10}T_{11}$ | $T_{11}T_{12}$ | $T_{12}T_{13}$ | $T_{13}T_{14}$ | $T_{14}T_{15}$ | $T_{15}T_{16}$ |
|---|---|---|---|---|---|---|---|
| 1.77 | 0.88 | 0.21 | 0.09 | 0.01 | 0.36 | 1.59 | 0.74 |

## 3.4 轉換成字串

前一步驟之後，我們得到了 slope-data，用於表示某基因在某實驗的變化過程；接下來，透過下列規則將 slope-data 轉換成由{ A、B、C、D }所組成的字串，其中每一字元代表相鄰兩時間點基因表現的變化程度。方法如下：

***<1> SortedData ← 將 slope-data 由大到小排列；***

***<2> HighAVG ←SortedData 前 1/2 資料的平均；***

***<3> AVG ←SortedData 所有資料的平均；***

***<4> LowAVG ←SortedData 後 1/2 資料的平均；***

轉換所得的字串：

$$ExpressionString[i]=\begin{cases}'A', 若 slope-data[i] \le HightAVG ;\\ 'B', 若 AVG \le slope-data[i] < HighAVG ;\\ 'C', 若 LowAVG \le slope-data[i] < AVG ;\\ 'D', 其他 ;\end{cases}$$

如此，對於任一筆基因表現資料，我們都可以得到對應的字串，用以表達此基因在實驗過程中的變化量。圖 3.4-1 爲基因 YJL043W 其 slope-data，與轉換所得的字串。

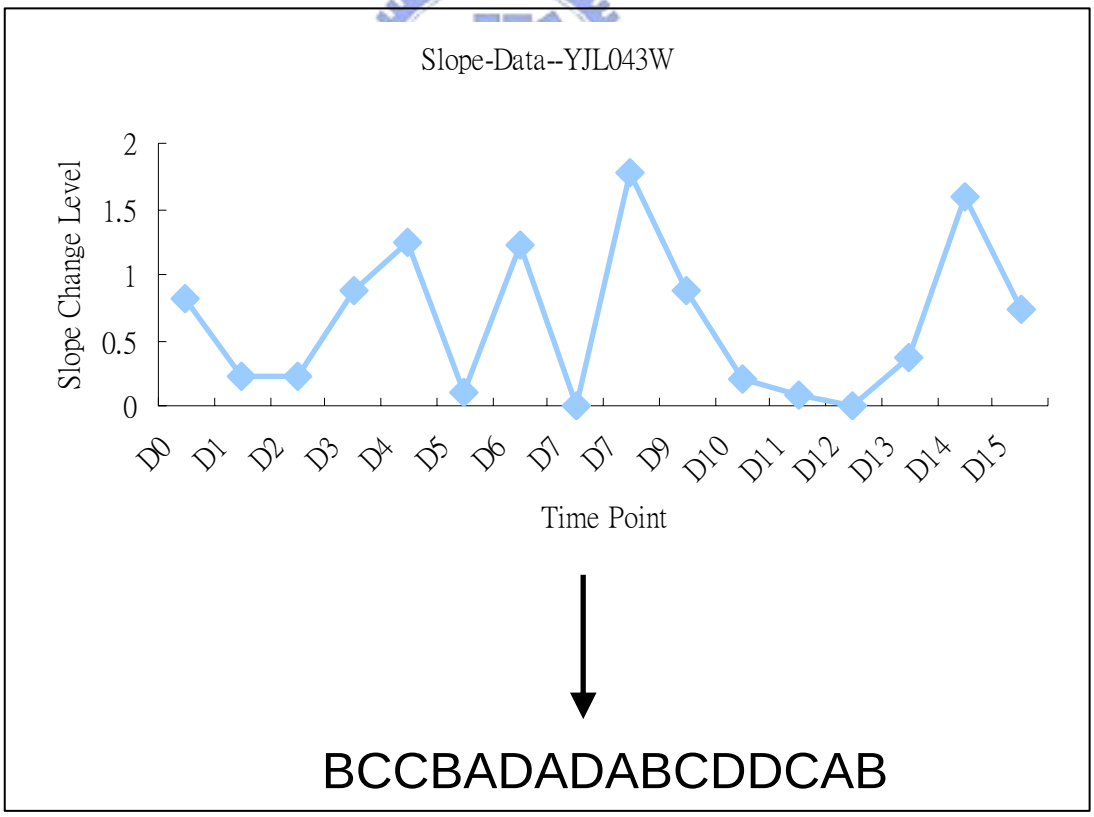



圖 3.4-1 YJL043W 的 Slope-data 與轉換後的字串

## 3.5 字串的排比

每一個基因表現實驗資料，都可由上述幾個步驟，轉換成相對應的字串。接下來，我們定義一評分矩陣(scoring matrix)，用於排比任兩字串並給與評分。

| — | X | | | | |
|---|---|---|---|---|---|
| **A** | -5 | 6 | | | |
| **B** | -5 | 5 | 6 | | |
| **C** | -5 | 5 | 5 | 6 | |
| **D** | -5 | 5 | 5 | 5 | 6 |
| | **—** | **A** | **B** | **C** | **D** |

Scoring Matrix (a)

| — | X | | | | |
|---|---|---|---|---|---|
| **A** | -5 | 6 | | | |
| **B** | -5 | 6 | 6 | | |
| **C** | -5 | -4 | 5 | -3 | |
| **D** | -5 | -5 | -4 | -3 | -3 |
| | **—** | **A** | **B** | **C** | **D** |

Scoring Matrix (b)

一般而言，評分矩陣的配分是經驗法則(或由眾多已知的實驗資料歸納而得)。例如，在 DNA 序列尋找共同結構元(motif)的研究裡，研究人員利用歸納所得的評分矩陣，去尋找共同結構元，對於不同的結構元家族，其評分矩陣亦不完全相同。

觀察上面兩個矩陣，除了配分不同外，其中矩陣(a)對於相同字元的配對(match)，都給予相同的評分；而矩陣(b)對於相同字元的配對給予不同的評分。對於本實驗，矩陣(b)的配分方式較符合要求，因爲在本實驗將資料離散化的過程中，字元”A”代表基因表現較爲劇烈，”Ｂ”次之，字元”D”則變化較不顯著。而利用微矩陣重建基因網路的方法，關心的是變化較爲顯著的片段，因此對於變化較劇烈的片段(如 A==A、A==B、B==B)我們可以給予較高的分數，以突顯其重要性。

圖 3.5-1 爲基因 YDR146C 與 YJL159W 字串透過上述矩陣(b)，並以全域排比 (global alignment)排比之結果。

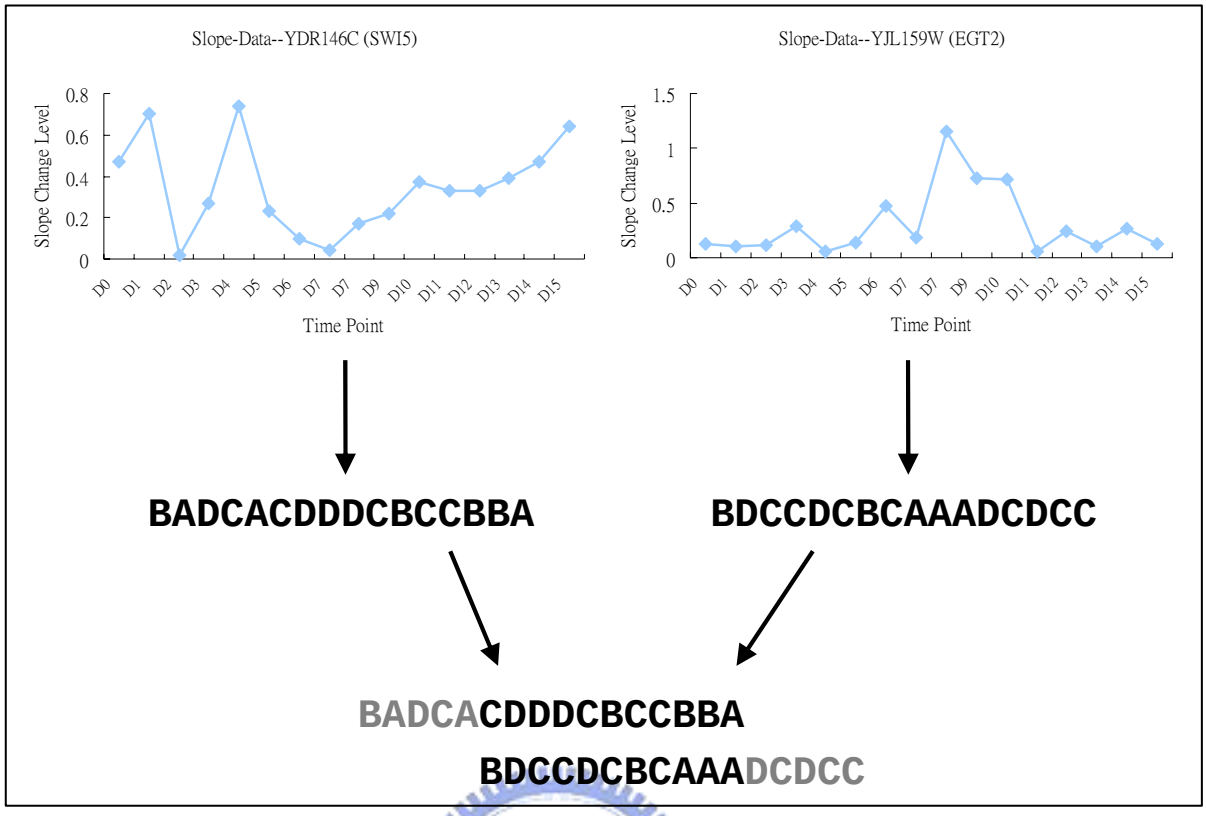



圖 3.5-1 YDR146C 與 YJL159W 字串的排比結果

## 3.6 重建基因調控網路

由前述步驟，我們可決定任兩個基因，在使用者設定的標準之下，是否有可能具調控關係。將這些子網路連結起來，即可得到一較完整之調控網路。

## 3.7 方法特色

根據觀察，在基因表現實驗的關係中，兩基因間的關係可分爲正相關、負相關、具時間差與具時間差之相關(Yu1 et al. 2003)。其中正相關只的是兩基因的基因表現呈現高度的相似，如圖 3.7-1(a)；而負相關只的是兩基因的基因表現呈現高度的相反，如圖 3.7-1 (b)；所謂的具時間差(time shift)描述的是，在大部分的情況下，調控基因欲影響受調控基因，期間常常存在一時間差，亦即受調控基因其基因表現需要一對時間才會反應，如圖 3.7-1 (c)；當然無論是活化或抑制都可能存在這樣的時間差，如圖 3.7-1 (d)。

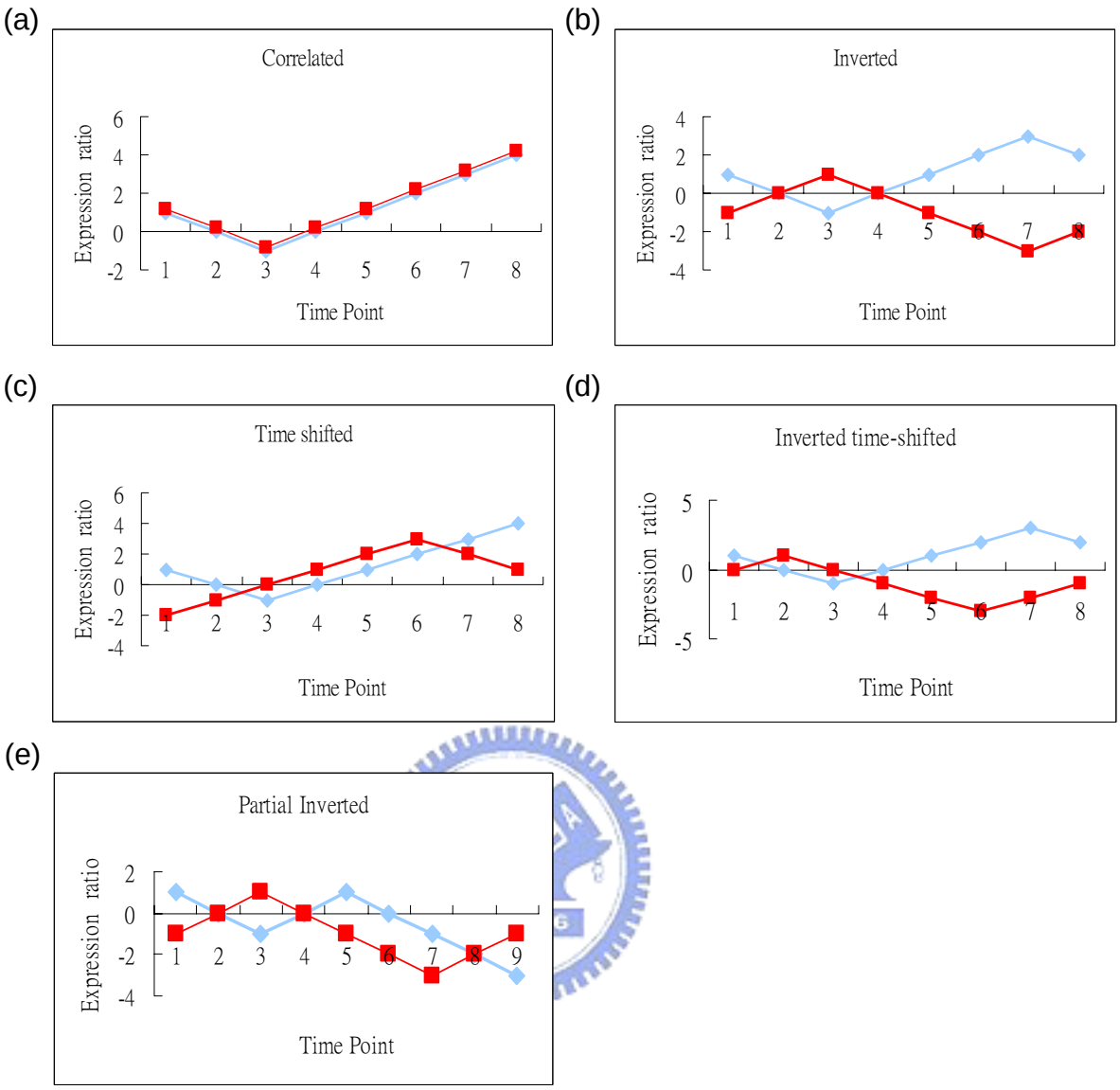



圖 3.7-1 調控基因與受被調控基因其基因表現資料的關係

根據前幾節的介紹，我們的研究可以正確地偵測上述四種現象。另外，除了前述幾個關係之外，我們發現調控基因在同一實驗中，其基因表現不一定與受調控基因之基因表現完全相同(或相反)，而是部分成正相關，部份成負相關，如圖(e)。基於這樣的觀察，我們定義一與前人不同的方法，希望能將對這樣的現象的考量，納入重建網路的系統中(見 3.3)，以獲得更多良好的預測結果。圖 3.7-2 爲一已知(Zou et al. 2005)具調控關係之基因對：(SWI4 , HTA1)。觀察其基因表現資料，發現在第 1 到第 14 個時間點之間，其走勢大致相同；而第 142 到第 17 個時間點，其走勢成反方向。即圖 3.7-1(e)所描述的現象。

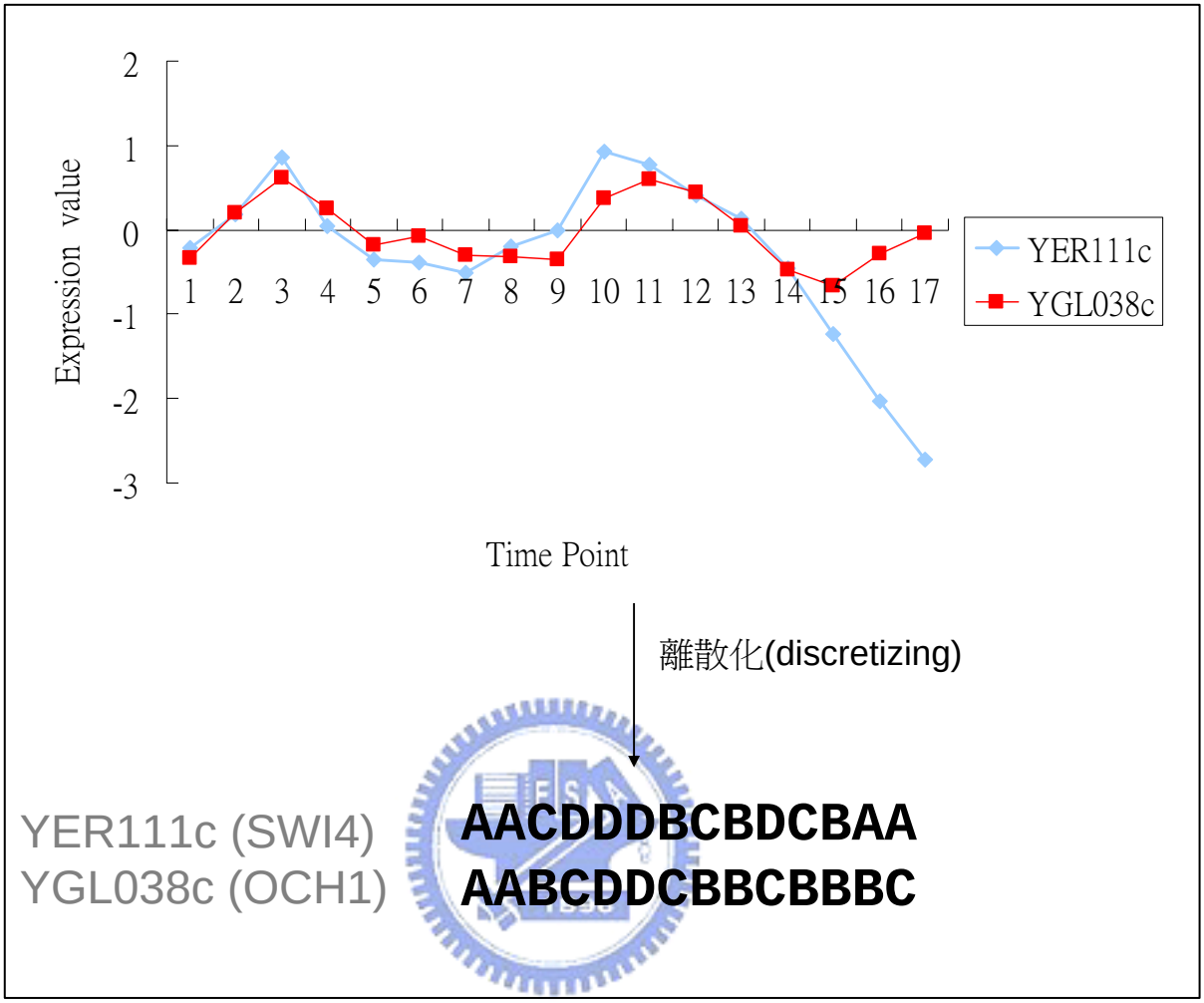



圖 3.7-2 已知的調控關係，並存在如圖 3.7-1(e)所描述的現象(Zou et al. 2005)。

# 第4章　實驗結果與討論

在第三章中，我們描述了將實驗資料轉換爲字串，並進行排比的方法與標準。本章將展現一連串實驗與比較，證明我們的方法的確可以找出基因間相互的關係。

## 4.1 實驗資料

隨著微矩陣實驗普遍地被應用，許多慷慨的研究者，大方地將這些寶貴的資料公開，讓生物資訊學者進行分析研究。針對重建調控網路，常用的資料包括 Cho(Cho et al. 1998)以及 Spellman 等人(Spellman et al. 1998)在 1998 年提供的實驗數據。這些資料主要包含 alpha、cdc15、cdc28 和 elutriation 等四組實驗。對於實驗所使用的資料，我們皆採用的是 cdc28 這一組，其理由如下：

<1> Cdc28 爲一時間序列的實驗，符合本實驗方法的要求。圖 3.7 所描述的五個基因表現的關係，都是基於時間序列的資料上；

<2> 對於所有的研究方法，實驗所採集的時間點愈多，愈有利於良好結果的預測，本研究亦不例外。該組資料包含 18 個時間點，且兩時間點的間隔較短(10 分鐘)，所測得的基因表現的數據較爲準確；

<3> 相較於 alpha 那一組資料，在 cdc28 中，其遺失的數據較少。若根據 3.2 的資料前處理的方法，alpha 資料中，有 103 個基因其遺失的時間點多於 50%；而 cdc28 資料中，只有 53 個基因其遺失的時間點多於 50%。

<4> 此實驗資料較爲完整，且被許多相關的研究所採用(Zou et al. 2005)，如此可方便將本實驗方法與前人研究之結果做比較。

## 4.2 實驗方法的可行性評估

實驗目的：

藉由此實驗，證明本研究所提出之方法，對於具調控關係之基因對，的確可以提供較高之鑑別度。

實驗資料：

在本實驗中，我們將所提出的方法，套用在下列兩組資料之基因表現資料 cdc28，並比較其結果。

<1> Zou 等人(Zou et al. 2005)提供的 54 組已知的調控基因對(如表 4.2-1)；

<2> 針對酵母菌(Saccharomyces cerevisiae)的 6178 基因任取兩個；

| Regulator | Target | Regulator | Target | Regulator | Target |
|---|---|---|---|---|---|
| ACE2 | SPO12 | NDD1 | ACE2 | SWI4 | PCL2 |
| FKH1 | BUD8 | NDD1 | CDC20 | SWI4 | CLB6 |
| FKH1 | HFF_1 | SWI4 | PCL1 | SWI4 | CLB2 |
| FKH1 | ACE2 | SWI4 | CLN2 | SWI4 | PLB3 |
| FKH1 | UTR2 | SWI4 | OCH1 | SWI4 | BUD9 |
| FKH1 | SWI6 | SWI4 | HO | SWI4 | HTA2 |
| FKH1 | SWI5 | SWI4 | SWE1 | SWI5 | EGT2 |
| FKH2 | GIC1 | SWI4 | GIN4 | SWI5 | MFA2 |
| FKH2 | CLB4 | SWI4 | RNR1 | SWI5 | YLR463 |
| FKH2 | ACE2 | SWI4 | MNN1 | SWI5 | HSP150 |
| MBP1 | CLB6 | SWI4 | NDD1 | SWI6 | HTB2 |
| MCM1 | STE6 | SWI4 | FKS1 | SWI6 | SIM1 |
| MCM1 | PIR3 | SWI4 | SPT21 | SWI6 | YPR075C |
| MCM1 | CLN2 | SWI4 | RSR1 | SWI6 | CDC6 |
| MCM1 | CLN3 | SWI4 | CWP1 | SWI6 | AGA1 |
| MCM1 | GIN4 | SWI4 | YBR071W | SWI6 | SPO12 |
| MCM1 | SIM1 | SWI4 | BUD4 | SWI6 | CIS3 |
| MCM1 | MFA1 | SWI4 | MBP1 | SWI6 | RSR1 |

表 4.2-1 Zou 等人所提論文中，已知的 54 組調控基因對。

實驗結果：

表 4.2-2 爲此實驗的結果比較。針對資料<1>，我們得到的平均分數爲 8.04；針對資料<2>，我們共取了 2000 組基因對進行字串排比，其平均爲分數爲 1.46。由實驗結果可以發現，跟隨機選取的基因對比較，已知具調控關係的基因對，在我們的方法中，的確可以得到較高的分數。

| Data Set | Average Scores |
|---|---|
| 54 known gene-gene pairs | 8.04 |
| 2000 random gene-gene pairs | 1.46 |

表 4.2-2 針對已知的調控基因對與隨機取得的基因對，所得的評分結果。

## 4.3單獨使用 TNP 與本研究方法之結果比較

本實驗的最終目的爲：比較本研究與 TNP 兩者重建基因調控網路能力。實驗可分爲三部分；其中，4.3.1 節的實驗是爲了決定 TNP 與本實驗的最佳門檻值(threshold)；而 4.3.2 節的實驗則使用 4.3.1 節中決定之門檻值，對兩種方法進行比較；爲了排除設定門檻值可能造成的偏頗，4.3.3 捨棄門檻值的設定，採用排名的方式來比較兩方法的預測能力，以做到更爲客觀之比較。

### 4.3.1 決定 TNP 與本研究方法之門檻值

實驗目的：

在 TNP 中以 PF 作爲重建基因調控網路之門檻值；本研究方法則以字串排比之分數(Scores)作爲門檻值。本實驗目的是找出最適當的 PF 與 Scores，以利 4.3.2 之實驗進行。

實驗資料：

我們以 TNP 所採用的 26 個轉錄因子所組成的轉錄調控模組(如表 4.3-1)，爲重建基因網路的目標。

| |
|---|
| ABF1、ACE2、ADR1、BAS1、BAS2、GAL4、GCN4、GCR1、HAP1、HSTF、LEU3、HATalpha1、HATalpha2、MBF、MCM1、MIG1、PDR3、PHO4、PUT3、RAP1、REB1、Repressor of CAR1、SBF、STE12、SWI5、TBP |

表 4.3-1 本實驗所使用之 26 個調控因子

關於驗證的資料，我們使用記錄於資料庫 SCPD(Zhu et al. 1999)之調控關係來決定某一方法之預測能力。

SCPD 全名爲 Saccharomyces Cerevisiae Promoter Database，集合了生物學家在酵母菌上的實驗結果，並整理出酵母菌基因與轉錄因子等相關資訊，包含酵母菌之轉錄因子結合區、轉錄起始位置之資訊，另外還提供部份經過生物實驗驗證之基因調控模組。在我們的實驗中，將以這些已知的證據，來比較兩個重建調控網路方法之預測能力。

實驗方法：

分別實作 TNP 與本實驗之研究方法，並針對同一組基因表現資料，以比較兩者重建基因調控網路之能力。

如 1.1 所言，一個轉錄因子爲一種特定的蛋白質，可能由多個蛋白質所組成，而這些蛋白質又由不同的基因所合成。在 26 個轉錄因子中，其中 23 個是由單一基因所合成之蛋白質;有 3 個則是由多基因所合成之蛋白質複合體(protein complex)，分別爲 Repressor of CAR1、MBF、SBF。其組合基因如表 4.3-2。

| **TF name** | **Component Gene(ORF)** | **Reference** |
|---|---|---|
| Repressor of CAR1 | RPD3(YNL330c)、SIN3(YOL004w)、UME6(YDR207c) | Wingender E. et al. 1996 |
| MBF | SWI6(YLR182w)、MBP1(YDL056w) | Sellman et al. 1998 |
| SBF | SWI4(YER111c)、SWI6(YLR182w) | Sellman et al. 1998 |

表 4.3-2 由多基因所組成之轉錄因子。

針對這些由多基因合成之調控因子，TNP 利用方程式 2.1(參見 2.2.1)，增加其變數以估計該調控因子與可能被調控基因之間的相關性。當調控因子只由單一基因所組成，則其估計之回歸方程式爲：

$$Y(t) = W_t X_1(t) + \varepsilon \ ;$$

若調控因子由兩個基因所合成(如 MBF)，則其估計之回歸方程式則爲：

$$Y(t) = W_{1,t} X_1(t) + W_{2,t} X_2(t) + \varepsilon$$

其中$Y(t)$爲可能受調控基因於時間點$t$之基因表現程度；$X_i(t)$爲調控因子第$i$個組成基因，於時間點$t$之基因表現程度。

在本研究所提出的方法中，針對這些由多基因組成之調控因子，我們分別將其組成基因與可能被調控基因之基因表現資料，做多次的字串排比，最後計算其平均，作爲此預測調控模組之給分。例如，考慮調控因子 Repressor of CAR1，與可能受調控基因 YAL001C，我們必須分別對基因對(YNL330c , YAL001C)、(YOL004w ,YAL001C)、(YDR207c , YAL001C)進行字串排比。如表 4.3-3，其得分分別爲-6、1、13，平均爲 2.67。因此預測過程中，調控模組(Repressor of CAR , YAL001C)的得分爲 2.67。

| 組成基因 | 字串排比結果 | 得分 |
|---|---|---|
| YNL330c (RPD3)<br>YAL001C | -BC--C--DCBAAACCDDC<br>ABCBBCDDDCCAAA----- | -6 |
| YOL004w (SIN3)<br>YAL001C | --DDDDCCCCBAABBA<br>ABCBBCDDDCCAA--A | 1 |
| YDR207c (UME6)<br>YAL001C | AAABBBCCDDDD-CC<br>AB-CBBCDDDCCAAA | 13 |

表 4.3-3 調控模組(Repressor of CAR , YAL001C)之預測結果

爲了決定兩方法之門檻值，PF 與 Scores，有較爲公平的評量方式，我們引用第三種重建基因網路之方式，Pattern Match(Hsu et al. 2004)，並分別與此兩方法比較，以決定其門檻值。Pattern Match 是利用判斷某基因之上游區，是否包含某調控因子之轉錄因子結合區(見 1.1)，以決定此基因是否受該轉錄因子所調控。基因上游區序列、轉錄因子結合區等資訊皆可以由 SCPD 資料庫所取得。

圖 4.3-1 爲 TNP 於不同 PF 值的設定下，與 Pattern Match 比較的結果。其中，X 軸代表不同的 PF 值；Y 軸爲預測精確度較高的調控模組數。以 PF=0.4 爲例，在此設定之下，上述 26 個調控模組中，Pattern Match 表現較好的有 11 組；而 TNP 表現較好的有 15 組。

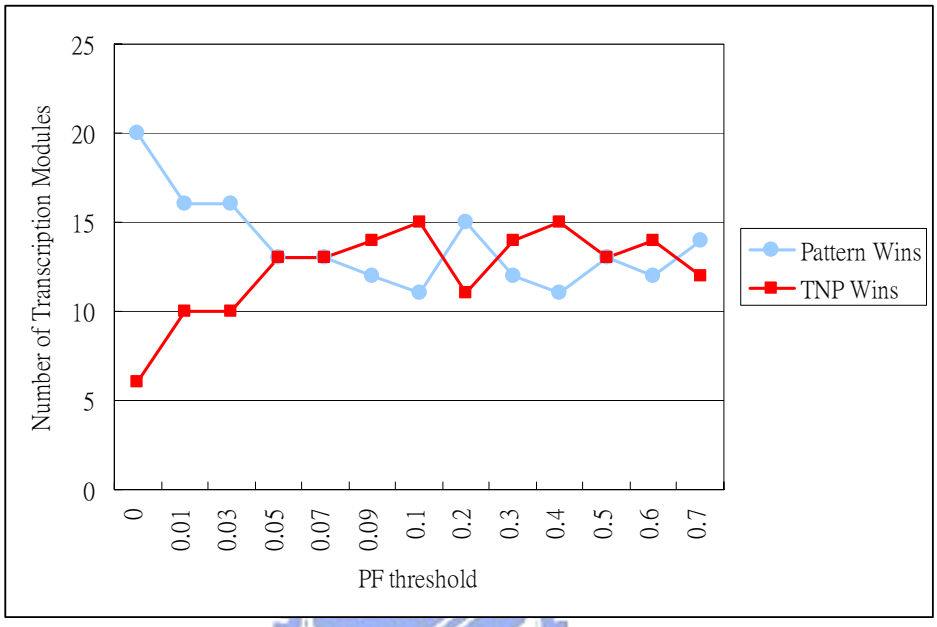



圖 4.3-1 TNP 於不同 PF 值之下，與 Pattern Match 之預測結果比較。
X 軸代表不同的 PF 值；Y 軸爲預測精確度較高的調控模組數。

圖 4.3-2 爲就所提出的方法於不同 Scores 值的設定下，與 Pattern Match 比較的結果。其中，X 軸代表不同的 Scores 值；Y 軸爲預測精確度較高的調控模組數。以 Scores = 6.5 爲例，在此設定之下，上述 26 個調控模組中，Pattern Match 表現較好的有 8 組；而本研究方法表現較好的有 18 組。

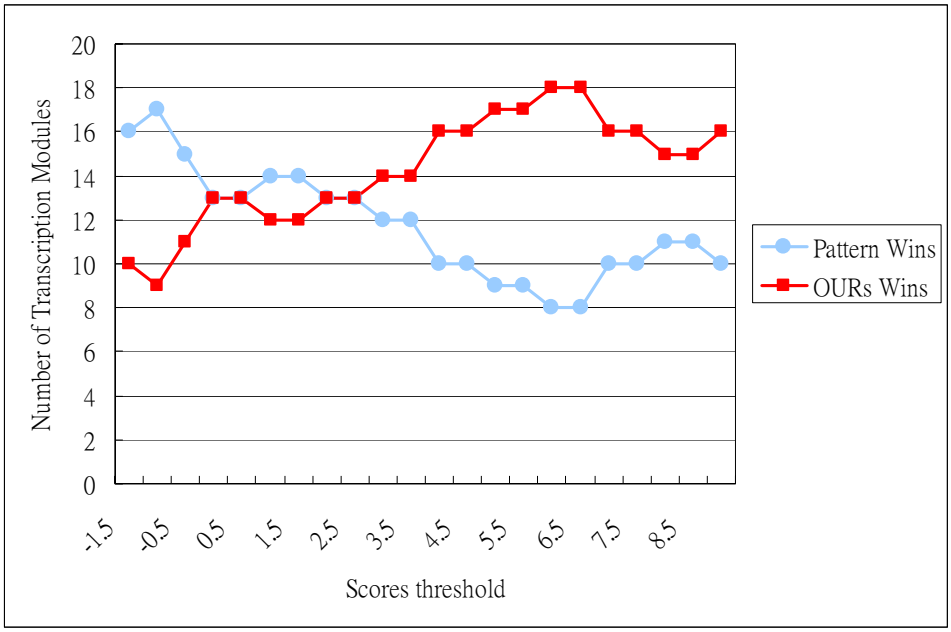



圖 4.3-2 本研究所提方法於不同 Scores 值之下，與 Pattern Match 之預測結果比較。X 軸代表不同的 Scores 值；Y 軸爲預測精確度較高的調控模組數。

實驗結果：

由實驗數據發現，根據所使用的基因表現資料，TNP 在 PF = 0.4 時，有較好的表現；而我們所提的方法則是在 Scores = 4.0 時，有較好的表現。因此，在 4.3.2 的比較實驗中，我們將以這兩個設定值來進行預測能力的比較 (結果可參考 4.3-4) 。

### 4.3.2 TNP 與本研究方法之比較 – 使用設定門檻值

實驗目的：

藉由本實驗，比較 TNP 與本研究所提方法之重建調控網路之能力。並根據所得的結果，分析兩者之優劣。

實驗資料：

同 4.3.1。以 TNP 所採用的 26 個轉錄因子所組成的轉錄調控模組(如表 4.3-1)，爲重建基因網路的目標。

實驗方法：

有了 4.3.1 節所得的 PF 值與 Scores，本實驗只需將此二數值分別設定於 TNP 與本系統，並針對所得結果比較即可。

實驗結果：

我們將前述 26 個轉錄因子其組成基因，與其他酵母菌之基因，分別使用 TNP 與本研究方法進行調控網路之重建，並以資料庫 SCPD(Zhu et al. 1999)記錄之調控關係來驗證兩者之實驗結果。結果如表 4.3-4。

表 4.3-4 大約可分爲三大欄：”Pattern”、”TNP”與”OURs”，分別記錄三方法的實驗結果。其中”P”爲該系統認爲具調控關係之基因個數；”TP”爲預測結果之 True Postive 個數，即預測結果中，於 SCPD 中有記錄的調控基因數；” Precision”爲預測之精確度，定義爲：

精確度(precision) = 正確預測的基因數(TP) / 預測會被調控的基因數(P)

而” selectivity”爲精確度提升倍數，依據 Hsu 等人定義爲：

方法二針對方法一之精確度提升倍數(selectivity) = 方法二之精確度（precision of method 2）/方法一的精確度（precision of method 1）。

TNP 與 OURs 的精確度提升倍數都是跟 Pattern 作比較；另外，OURs 中 selectivity*是記錄本研究所提方法，對於 TNP 的精確度提升倍數。當數值大於 1，則表示針對該調控模組，本研究所提方法之表現較爲優秀；若數值小於 1 則代表，TNP 之預測能力較出色；另外，若 TNP 預測之 true positive 個數爲 0，而本研究方法預測之結果不爲 0 時，計算精確度提升倍數會造成分母爲 0，我們以”>>”表示。而“Coverage”爲預測結果之正確涵蓋率，定義爲：

正確涵蓋率(coverage) = 正確預測的基因數 / SCPD 中對應的轉錄調控模組所含基因數。

一般來說，精確度與正確涵蓋率無法兼得，亦即愈高的精確度，往往伴隨著較低的正確涵蓋率；反之亦然。因此，我們引用 F score(Lewis, D., and Gale,1994)來比較兩方法的預測能力。F score 同時將精確度與正確涵蓋率，定義爲：

F score = 2*精確度*正確涵蓋率 / (精確度+正確涵蓋率)

由表 4.3-4 我們發現，在 26 組調控模組中，本研究所提方法，其精確度較 TNP 爲提升的有：ABF1、ACE2、ADR1、BAS1、BAS2、GAL4、HAP1、LEU3、HATalpha2、MBF、MCM1、MIG1、PHO4、RAP1、REB1、Repressor of CAR1、SBF、STE12，共計 18 組；而 GCN4、GCR1、HSTF、HATalpha1、PDR3、PUT3、SWI5、TBP 等 8 組準精確度則是下降。值得注意的是，單獨使用 TNP 預測時，有 2 個調控模組(BAS2、LEU3)無法預測到任何已記錄於 SCPD 之調控關係，而使用本研究所提之方法之後，即可改善這些模組之預測能力

表 4.3-4 Pattern、TNP、與本研究所提方法之實驗結果與比較。其中 P 爲系統預測之受調控基因數；TP 爲記錄於 SCPD 已知的受調控基因數；Precision 爲預測精確度；Selectivity 爲針對 Pattern 方法之精確度提升倍數；F score 一爲常用之評量標準，較單獨考慮 Precision 或 coverage 爲客觀；Selectivity*爲本系統針對 TNP 之精確度提升倍數；若 TNP 預測之 true positive 個數爲 0，而本研究方法預測之結果不爲 0，計算精確度提升倍數會造成分母爲 0，因此以”>>”表示。

| | Pattern | | | TNP (PF = 0.4) | | | | | OURs (Scores = 6.5) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Family | TP | P | Precision ($*10^{-1}$) | TP | P | Selectivity | coverage | Fscores ($*10^{-1}$) | TP | P | Selectivity | Selectivity* | Coverage | Fscores ($*10^{-1}$) |
| ABF1 | 17 | 2907 | 0.058 | 10 | 1462 | 1.170 | 0.526 | 0.135 | 12 | 1261 | 1.627 | 1.391 | 0.632 | 0.188 |
| ACE2 | 1 | 1820 | 0.005 | 1 | 1104 | 1.649 | 1.000 | 0.018 | 1 | 858 | 2.121 | 1.287 | 1.000 | 0.023 |
| ADR1 | 2 | 4907 | 0.004 | 1 | 2625 | 0.935 | 0.500 | 0.008 | 1 | 2278 | 1.077 | 1.152 | 0.500 | 0.009 |
| BAS1 | 3 | 1562 | 0.019 | 1 | 753 | 0.691 | 0.200 | 0.026 | 1 | 472 | 1.103 | 1.595 | 0.200 | 0.042 |
| BAS2 | 1 | 5722 | 0.002 | 0 | 2716 | 0.000 | 0.000 | 0.000 | 1 | 1607 | 3.561 | >> | 1.000 | 0.012 |
| GAL4 | 6 | 344 | 0.174 | 3 | 157 | 1.096 | 0.500 | 0.368 | 2 | 69 | 1.662 | 1.517 | 0.333 | 0.533 |
| GCN4 | 9 | 6050 | 0.015 | 4 | 3165 | 0.850 | 0.444 | 0.025 | 3 | 2421 | 0.833 | 0.980 | 0.333 | 0.025 |
| GCR1 | 6 | 5874 | 0.010 | 1 | 3164 | 0.309 | 0.167 | 0.006 | 0 | 2252 | 0.000 | 0.000 | 0.000 | 0.000 |
| HAP1 | 4 | 60 | 0.667 | 3 | 44 | 1.023 | 0.600 | 1.250 | 2 | 26 | 1.154 | 1.128 | 0.400 | 1.333 |
| HSTF | 6 | 5103 | 0.012 | 5 | 2551 | 1.667 | 0.833 | 0.039 | 0 | 809 | 0.000 | 0.000 | 0.000 | 0.000 |
| LEU3 | 2 | 37 | 0.541 | 0 | 16 | 0.000 | 0.000 | 0.000 | 1 | 16 | 1.156 | >> | 0.500 | 1.111 |
| HATalpha1 | 3 | 1988 | 0.015 | 2 | 997 | 1.329 | 0.667 | 0.040 | 0 | 471 | 0.000 | 0.000 | 0.000 | 0.000 |
| HATalpha2 | 7 | 2126 | 0.033 | 5 | 1319 | 1.151 | 0.714 | 0.075 | 3 | 638 | 1.428 | 1.240 | 0.429 | 0.093 |
| MBF | 6 | 1643 | 0.037 | 6 | 1094 | 1.502 | 1.000 | 0.109 | 4 | 716 | 1.530 | 1.019 | 0.667 | 0.111 |
| MCM1 | 25 | 2441 | 0.102 | 7 | 1325 | 0.516 | 0.280 | 0.104 | 7 | 677 | 1.010 | 1.957 | 0.280 | 0.199 |
| MIG1 | 8 | 633 | 0.126 | 3 | 290 | 0.819 | 0.300 | 0.201 | 3 | 167 | 1.421 | 1.737 | 0.300 | 0.343 |

(續)

| | Pattern | | | TNP (PF = 0.4) | | | | | OURs (Scores = 6.5) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Family | TP | P | Precision ($*10^{-1}$) | TP | P | Selectivity | coverage | Fscores ($*10^{-1}$) | TP | P | Selectivity | Selectivity* | Coverage | Fscores ($*10^{-1}$) |
| PDR3 | 6 | 177 | 0.339 | 6 | 104 | 1.702 | 1.000 | 1.091 | 3 | 58 | 1.526 | 0.897 | 0.500 | 0.938 |
| PHO4 | 4 | 2160 | 0.019 | 1 | 1156 | 0.467 | 0.250 | 0.017 | 1 | 684 | 0.789 | 1.690 | 0.250 | 0.029 |
| PUT3 | 1 | 275 | 0.036 | 1 | 152 | 1.809 | 1.000 | 0.131 | 0 | 106 | 0.000 | 0.000 | 0.000 | 0.000 |
| RAP1 | 14 | 1983 | 0.071 | 10 | 1206 | 1.174 | 0.625 | 0.164 | 11 | 718 | 2.170 | 1.848 | 0.688 | 0.301 |
| REB1 | 12 | 1409 | 0.085 | 5 | 762 | 0.770 | 0.357 | 0.129 | 5 | 635 | 0.925 | 1.200 | 0.357 | 0.155 |
| Repressor of CAR1 | 13 | 455 | 0.286 | 8 | 300 | 0.933 | 0.615 | 0.511 | 1 | 5 | 7.000 | 7.500 | 0.077 | 1.111 |
| SBF | 3 | 3611 | 0.008 | 3 | 2430 | 1.486 | 1.000 | 0.025 | 2 | 965 | 2.495 | 1.679 | 0.667 | 0.041 |
| STE12 | 4 | 4682 | 0.009 | 2 | 2242 | 1.044 | 0.500 | 0.018 | 1 | 653 | 1.792 | 1.717 | 0.250 | 0.030 |
| SWI5 | 1 | 4567 | 0.002 | 1 | 2993 | 1.526 | 0.500 | 0.007 | 0 | 1255 | 0.000 | 0.000 | 0.000 | 0.000 |
| TBP | 17 | 4831 | 0.035 | 14 | 2839 | 1.401 | 0.778 | 0.098 | 9 | 1869 | 1.368 | 0.976 | 0.500 | 0.095 |

### 4.3.3 TNP 與本研究方法之比較 – 使用排名值

實驗目的：

對於不同的基因表現資料，門檻值的設定可能有所不同。預測結果的精確度，也會因爲門檻值的設定而改變，因此，爲了排除在設定門檻值時，可能造成的偏頗，本實驗將捨棄門檻值的設定，而採用排名的方式來比較兩方法的預測能力，以求更爲客觀之比較。

實驗資料：

同 4.3.1。以 TNP 所採用的 26 個轉錄因子所組成的轉錄調控模組(如表 4.3-1)，爲重建基因網路的目標。

實驗方法：

對於每一個調控基因，我們將 TNP 所預測之可能受調控基因，以 PF 值由小到大排列(PF 值愈小，其可信程度愈高，見 2.2.1)；同理，將本研究所預測之可能受調控基因以字串排比的得分 Scores 值由大到小排列。取出兩者的前 N 名，計算其中 True Positive 的個數以比較兩者預測的準確程度。

實驗結果：

根據實驗方法，我們分別取了 9 組不同的 N (10、50、100、200、300、400、500、600、700)，來比較兩方法預測結果的精確度。結果如表 4.3-5。其中”Top N”一欄記錄兩個方法中，所預測的前 N 組調控模組。”Ours wins”一欄記錄在 26 個調控模組中，本研究所預測的結果較 TNP 所預測準確之調控模組個數；反之，記錄於”TNP wins”；” Families wins by our methods”一欄則標明，在本方法中預測較爲準確之調控模組。

| Top N | TNP wins | Ours wins | Families wins by our methods |
|---|---|---|---|
| 10 | 4 | 3 | HAP1、PDR3、TBP |
| 50 | 6 | 3 | GAL4、MCM1、SBF |
| 100 | 5 | 4 | GAL4、MCM1、MIG1、RAP1 |
| 200 | 3 | 7 | GAL4、MCM1、MIG1、PHO4、RAP1、REB1、SBF |
| 300 | 4 | 9 | MATalpha2、MBF、MCM1、MIG1、PHO4、RAP1、REB1、Repressor of CAR1、SBF |
| 400 | 5 | 7 | BAS1、MATalpha2、MBF、MCM1、PHO4、RAP1、SBF |
| 500 | 7 | 6 | BAS1、MATalpha2、MBF、MCM1、PHO4、RAP1 |
| 600 | 8 | 6 | BAS1、MATalpha2、MBF、MCM1、PHO4、RAP1 |
| 700 | 7 | 7 | BAS1、BAS2、MATalpha2、MCM1、PHO4、RAP1、REB1 |

表 4.3-5 使用排名值比較本研究所提之方法與 TNP 之預測能力。

## 4.4 融合本實驗方法與 TNP 之結果比較

實驗目的：

觀察表 4.3-4，我們可以發現，在設定最佳門檻值的情況，ABF1、ACE2、ADR1、BAS1、BAS2、GAL4、HAP1、LEU3、HATalpha2、MCM1、MIG1、PDR3、PHO4、RAP1、REB1、Repressor of CAR1、SBF、STE12 等 18 個調控因子，本研究所提之方法，擁有較好的預測能力；而 GCN4、GCR1、HSTF、HATalpha1、MBF、PUT3、SWI5、TBP 等 8 個調控因子，TNP 所預測的結果較爲準確。因此，此一實驗的目的是藉由融合此兩種重建調控網路之方法，以提高預測的準確性。

實驗資料：

同 4.3。以表 4.3-1 中 26 個調控因子所形成之調控模組爲對象，並以 SCPD 資料庫記載之調控模組爲驗證資料。

實驗方法：

4.3.1 節的實驗決定了 TNP 與本研究方法之最佳門檻值；4.3.2 節實驗中，依據 4.3.1 的結果設定最佳門檻值，進行重建網路的工作。本實驗以聯集的方式將此兩方法的結果結合並比較其精確度的改變。以調控因子 ABF1 爲例，TNP 門檻值 PF=0.4 時，預測的受調控基因有 1463 個，而其中 10 個於 SCPD 有記載(True Positive)；本研究方法之門檻值 Scores=4 時，預測的受調控基因有 1426 個，而其中 13 個於 SCPD 有記載；兩者取聯集之後，預測的受調控基因有 2180，而其中有 15 個於 SCPD 有記載。詳細結果如表 4.4-1。

實驗結果：

透過融合兩種方法，我們將所得結果列於表 4.4-1。”Family”一欄爲該調控因子所形成之調控模組；”TNP”一欄爲單獨使用 TNP 所預測之結果；”Combined”爲結合本研究方法所得的結果；”Precision”、” Selectivity”、” Coverage”與”F scores”定義同 4.3.2 節，”F score*”爲結合兩方法之後，F score 高於單獨使用 TNP 時之 F score 之改進倍數，定義爲：

F scors* = 結合兩方法之 F score / 單獨使用 TNP 之 F score

另外，若計算評量值時，分母爲 0，則以”>>”表示之。

結果發現，結合兩種方法後，ABF1、BAS1、BAS2、GAL4、LEU3、HATalpha2、MCM1、MIG1、PHO4、RAP1、REB1、STE12 等 12 個調控因子，其預測精確度較單獨使用 TNP 時爲高；而 ACE2、ADR1、GCN4、GCR1、HAP1、HSTF、HATalpha1、MBF、PDR3、PUT3、Repressor of CAR1、SBF、SWI5、TBP 等 14 個調控因子，使用結合兩者的方法，對於精確度的提升沒有幫助。將此結果與 4.3.2 之實驗結果比對發現，針對 ABF1、BAS1、BAS2、GAL4、LEU3、HATalpha2、MCM1、MIG1、PHO4、RAP1、REB1、STE12 等調控因子，本研究方法的表現較好，因此結合兩方法之後，這些調控模組的改善程度較爲顯著。值得注意的是，BAS2 與 LEU3 在單獨使用 TNP 時，無法預測到任何已知的調控模組，但結合本方法之後，情況明顯改善。

| | TNP ( PF=0.4 ) | | | | | Combined (PF=0.4,Scores = 6.5 ) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Family | TP | P | Precision ($*10^{-1}$) | Coverage | F score ($*10^{-1}$) | TP | P | Selectivity | Coverage | Fscore* |
| ABF1 | 10 | 1462 | 0.0684 | 0.5263 | 0.1352 | 15 | 2097 | 1.0458 | 0.7895 | 1.0494 |
| ACE2 | 1 | 1104 | 0.0091 | 1.0000 | 0.0181 | 1 | 1410 | 0.7830 | 1.0000 | 0.7831 |
| ADR1 | 1 | 2625 | 0.0038 | 0.5000 | 0.0076 | 1 | 3602 | 0.7288 | 0.5000 | 0.7289 |
| BAS1 | 1 | 753 | 0.0133 | 0.2000 | 0.0265 | 2 | 988 | 1.5243 | 0.4000 | 1.5257 |
| BAS2 | 0 | 2716 | 0.0000 | 0.0000 | 0.0000 | 1 | 3397 | >> | 1.0000 | >> |
| GAL4 | 3 | 157 | 0.1911 | 0.5000 | 0.3681 | 4 | 191 | 1.0960 | 0.6667 | 1.1032 |
| GCN4 | 4 | 3165 | 0.0126 | 0.4444 | 0.0252 | 4 | 3883 | 0.8151 | 0.4444 | 0.8155 |
| GCR1 | 1 | 3164 | 0.0032 | 0.1667 | 0.0063 | 1 | 3979 | 0.7952 | 0.1667 | 0.7955 |
| HAP1 | 3 | 44 | 0.6818 | 0.6000 | 1.2500 | 3 | 48 | 0.9167 | 0.6000 | 0.9231 |
| HSTF | 5 | 2551 | 0.0196 | 0.8333 | 0.0391 | 5 | 2940 | 0.8677 | 0.8333 | 0.8680 |
| LEU3 | 0 | 16 | 0.0000 | 0.0000 | 0.0000 | 1 | 26 | >> | 0.5000 | >> |
| HATalpha1 | 2 | 997 | 0.0201 | 0.6667 | 0.0400 | 2 | 1200 | 0.8308 | 0.6667 | 0.8313 |
| HATalpha2 | 5 | 1319 | 0.0379 | 0.7143 | 0.0754 | 6 | 1531 | 1.0338 | 0.8571 | 1.0346 |
| MBF | 6 | 1094 | 0.0548 | 1.0000 | 0.1091 | 6 | 1277 | 0.8567 | 1.0000 | 0.8574 |
| MCM1 | 7 | 1325 | 0.0528 | 0.2800 | 0.1037 | 12 | 1606 | 1.4143 | 0.4800 | 1.4189 |
| MIG1 | 3 | 290 | 0.1034 | 0.3000 | 0.2013 | 5 | 384 | 1.2587 | 0.5000 | 1.2670 |
| PDR3 | 6 | 104 | 0.5769 | 1.0000 | 1.0909 | 6 | 126 | 0.8254 | 1.0000 | 0.8333 |
| PHO4 | 1 | 1156 | 0.0087 | 0.2500 | 0.0172 | 2 | 1482 | 1.5601 | 0.5000 | 1.5612 |
| PUT3 | 1 | 152 | 0.0658 | 1.0000 | 0.1307 | 1 | 208 | 0.7308 | 1.0000 | 0.7321 |
| RAP1 | 10 | 1206 | 0.0829 | 0.6250 | 0.1639 | 13 | 1442 | 1.0872 | 0.8125 | 1.0893 |
| REB1 | 5 | 762 | 0.0656 | 0.3571 | 0.1292 | 8 | 1010 | 1.2071 | 0.5714 | 1.2117 |
| Repressor of CAR1 | 8 | 300 | 0.2667 | 0.6154 | 0.5112 | 8 | 301 | 0.9967 | 0.6154 | 0.9968 |
| SBF | 3 | 2430 | 0.0123 | 1.0000 | 0.0247 | 3 | 2702 | 0.8993 | 1.0000 | 0.8994 |
| STE12 | 2 | 2242 | 0.0089 | 0.5000 | 0.0178 | 3 | 2542 | 1.3230 | 0.7500 | 1.3233 |
| SWI5 | 1 | 2993 | 0.0033 | 0.5000 | 0.0067 | 1 | 3388 | 0.8834 | 0.5000 | 0.8834 |
| TBP | 14 | 2839 | 0.0493 | 0.7778 | 0.0980 | 16 | 3586 | 0.9048 | 0.8889 | 0.9059 |

表 4.4-1 單獨使用 TNP、與結合 TNP 與本研究所提方法之實驗結果與比較。其中 P 爲系統預測之受調控基因數；TP 爲記錄於 SCPD 已知的受調控基因數；Precision 爲預測精確度；coverage 爲正確涵蓋率；Selectivity 爲針對單獨使用 TNP 之精確度提升倍數；F score 一爲常用之評量標準，較單獨考慮 Precision 或 coverage 爲客觀。若計算評量值時，分母爲 0，則以”>>”表示之。

## 4.5 重建生物調控網路

本研究以細胞週期爲例，觀察了 MCM1、ACE2、SWI5、SBF、MBF 等與細胞週期相關之轉錄因子與 CLB1、CLB2、CLN3、SWI4、FAR1、RME1、SIC1、CDC6、CLN1、CLN2、CLB5、CLB6 等基因的關係，並以 Mendenhall 等人之實驗結果(表 4.5-1)來驗證。

| Cell Cycle | TF | Target genes | Functions |
|---|---|---|---|
| M/G1, Early G1 | MCM1 | CLN3 | Cyclin activator of CDC28 in G1. |
| | | SWI4 | DNA binding component of SBF transcription factor. Important for Start-specific expression of CLN1 and CLN2. |
| | | FAR1 | CKI specific for CDC28-CLN complexes. |
| | ACE2 | RME1 | Positive factor in CLN2 expression. Negatively regulates early sporulation-specific genes. |
| | | SIC1 | CKI specific for CDC28-CLB complexes. |
| | SWI5 | CDC6 | Required for DNA replication. |
| | | | Inhibitor of CLB-CDC28 complexes. |
| Start (late G1) | SBF | CLN1 | Cyclin activator of CDC28 at Start . |
| | | CLN2 | Cyclin activator of CDC28 at Start . |
| | MBF | CLB5 | Cyclin activator of CDC28 at Start . |
| | | CLB6 | Cyclin activator of CDC28 at Start . |

表 4.5-1 細胞週期相關的轉錄因子及所調控的基因 (Mendenhall et al. 1998)

圖 4.5-1 爲將上述調控模組爲預測之結果。其中方框爲調控因子，圓圈代表基因，箭頭代表調控關係，而直線則連接調控因子與其組成基因。預測的 20 個調控關係中，其中有 6 個與表 4.5-1 之實驗結果相符。圖 4.5-2 爲使用本研究方法之預測結果。預測的 10 個調控關係中，有 2 個與表 4.5-1 之實驗結果相符。圖 4.5-3 爲將 TNP 與本研究方法結合後，採用交集的方式所預測之結果。預測的 9 個調控關係中，有 2 個與表 4.5-1 之實驗結果相符。

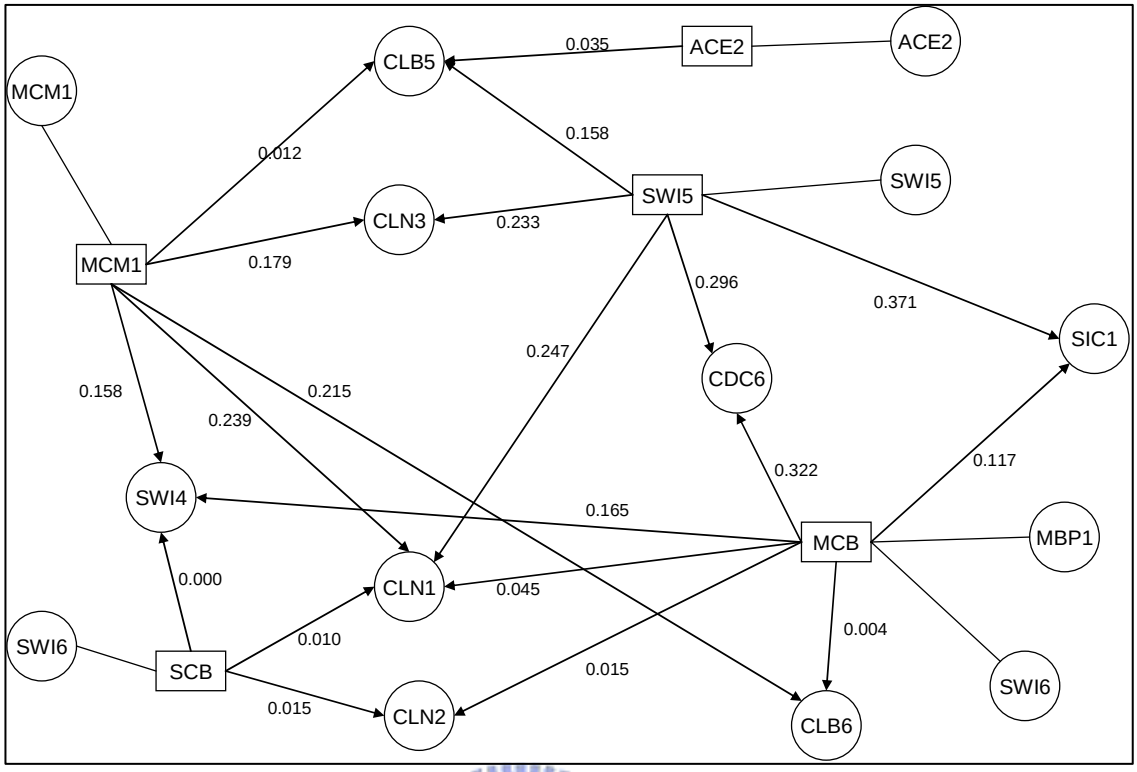



表 4.5-1 單獨使用 TNP 預測細胞週期相關調控模組之結果，PF 門檻值設定爲 0.4。

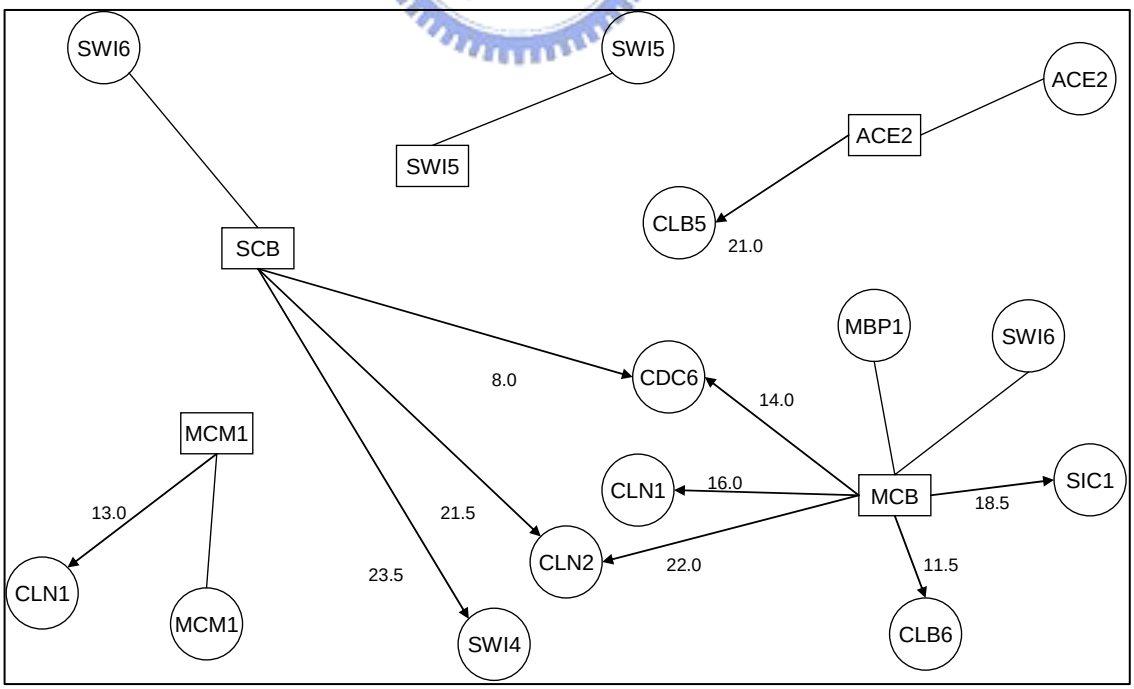



表 4.5-2 單獨使用本研究方法預測細胞週期相關調控模組之結果，Scores 門檻設定爲 6.5。

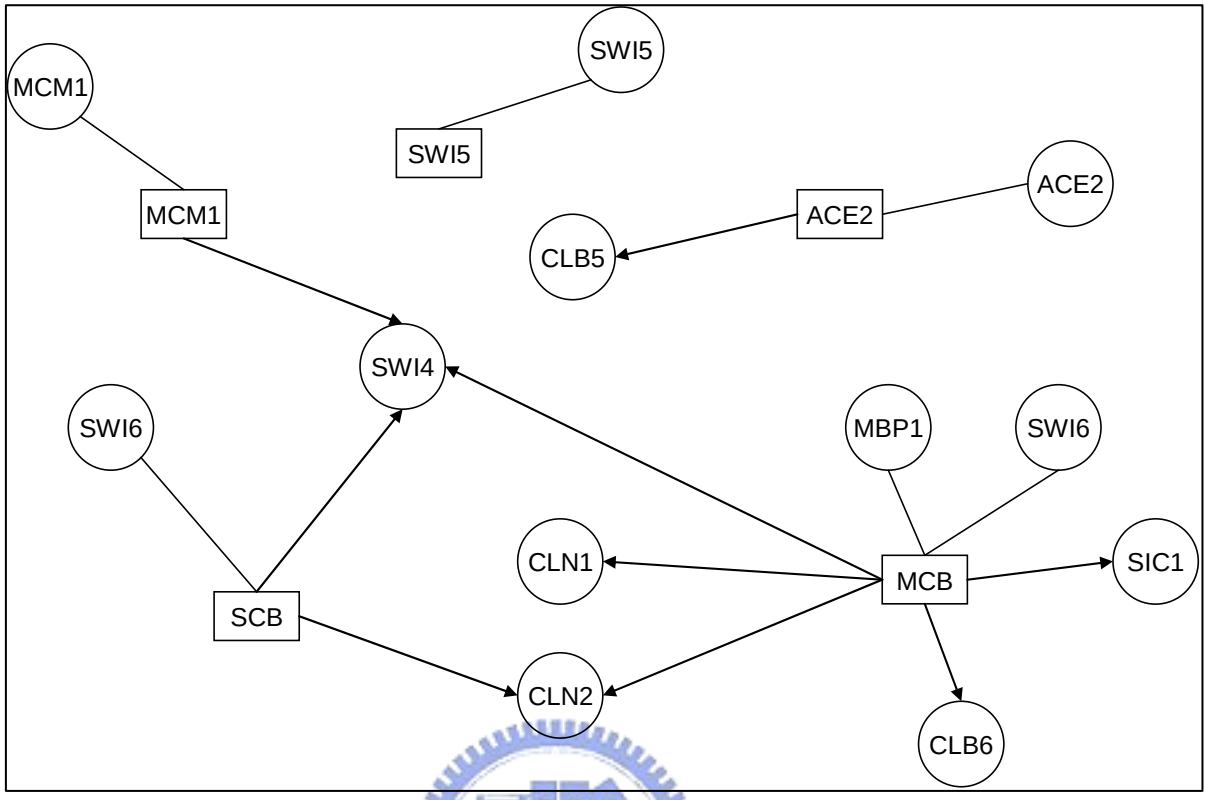



表 4.5-3 交集 TNP 與本研究方法預測細胞週期相關調控模組之結果。PF 門檻值設定爲 0.4，Scores 門檻設定 6.5；

# 第5章 結論與未來研究方向

## 5.1 結論

基於調控因子與受調控基因之間的關係，確實存在於基因表現資料的前提下(參考3.1)，本研究將時間序列的基因表現資料，轉換成數個子母所形成的連續字串，並利用字串比對，決定任兩基因間之相關性。如此的作法，沒有其他方法(如貝氏網路)的高度時間複雜度；而字串排比，亦巧妙的解決基因表現”時間差”(time delay)的問題。

分別針對 Zou 等人論文中 54 組已知的調控基因對與隨機選取之基因組，本研究之評分方式對於已知的調控關係，的確具有較高之辨識程度。另外(見 4.3.1)，針對酵母菌 cdc28 這組基因表現資料，TNP 之最佳門檻值爲 0.4；而本研究方法之最佳門檻值爲 6.5。將 TNP 與本研究方法，分別設定最佳之門檻值，並針對 SCPD 中 26 已知的調控模組，與 TNP 相較，我們順利的改善其中 18 個調控模組預測之準確程度。

對於已知與細胞週期相關的轉錄因子及所調控基因(Mendenhall)，我們分別以 TNP 與本研究方法重建調控網路(見 4.5)。我們所預測的 10 個結果中，有 2 個與已知的結果相符；結合兩方法；結合兩方法後，在預測的 9 個結果中，有 2 個與已知的結果相符。因此，本系統之結果，可以提供生物研究人員之參考。

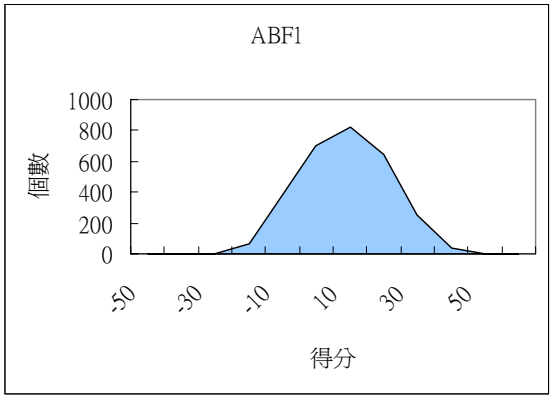
## 5.2 未來研究方向

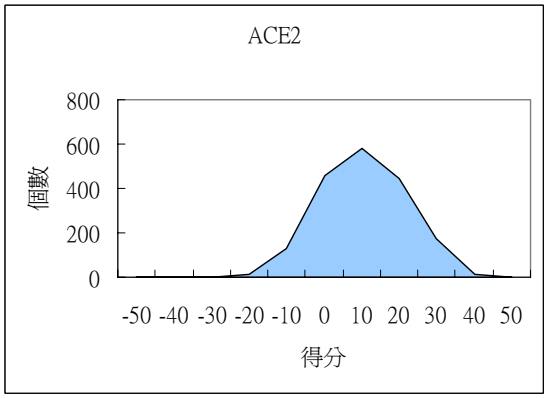
在本研究所提之方法中，每一個步驟都具有調整的彈性。3.4 節中我們將基因表現資料分成 4 種不同的表現程度，但是這並不一定是個準則。分階選擇的不同會影響實驗的精確度，因此，將來可以利用其他的方式(如 SOM)選用不同的分階，討論不同的分階對於實驗所帶來的影響；3.5 節中，我們使用全局排比(global alignment)將轉換後之基因資料作排比。在未來的研究中，可以選用局部排比(local alignment)或者半全局排比(semi-global alignment)來擷取部份的基因表現，讓基因間的關係更爲彈性；評分矩陣的選用，往往是依據經驗法則，如 3.5 節中，我們的評分矩陣亦是如此。但是隨著已知調控模組的增加，我們可以使用其他的學習方法(如基因規劃法，Genetic Programming)，修正評分矩陣，如此必定可以大幅度改進實驗預測之準確性。

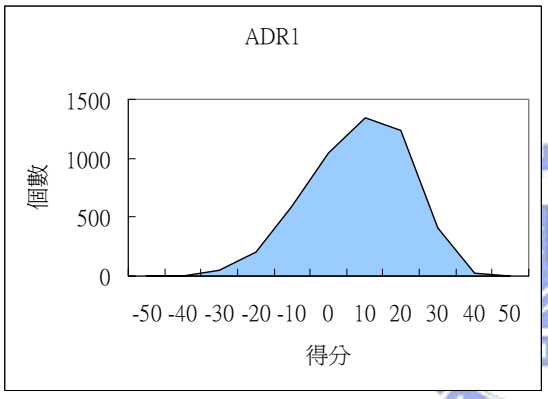
如 2.3 節所述，生物上的許多問題，往往無法單獨倚靠計算機或統計等方法解決，需要加入生物上的資訊才能獲得更好的結果。如 TNP 利用 SCPD 中整理之調控因子結合區序列，掃描基因之上游區，作爲實驗方法得前處理，如此大幅度降低其 false positive，進而提升預測的準確程度。因此，加入適當之生物資訊，亦可是將來繼續努力方向之一。

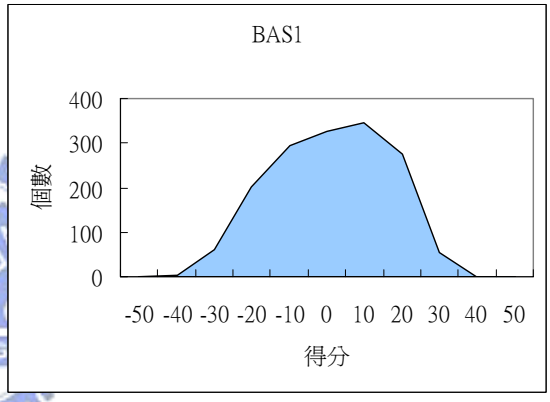
結合多種實驗(或方法)重建調控網路，一直被許多研究人員所採用的方式。Hsu 將 TNP 與 PROSPECT(Fujibuchi et al. 2001)的結合、本研究與 TNP 之結合。將適當的實驗方法結合，亦可改善實驗的結果。因此，如何將前人研究之優點，與本研究結合，亦是未來可行的努力目標。

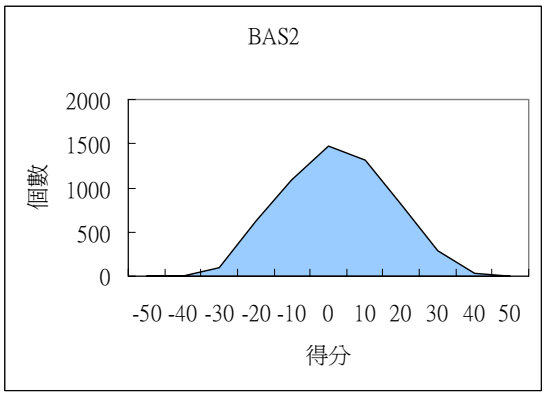
# 附錄一：26 個轉錄調控模組，其字串排比之得分分布。

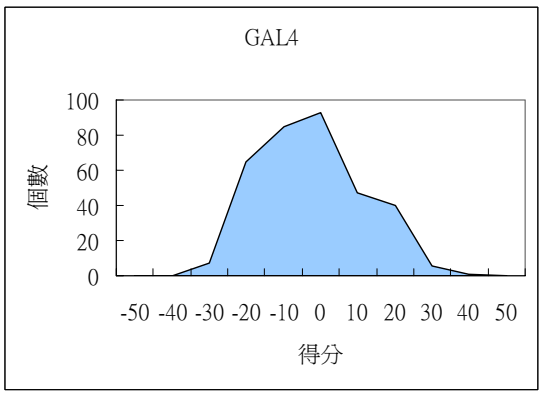


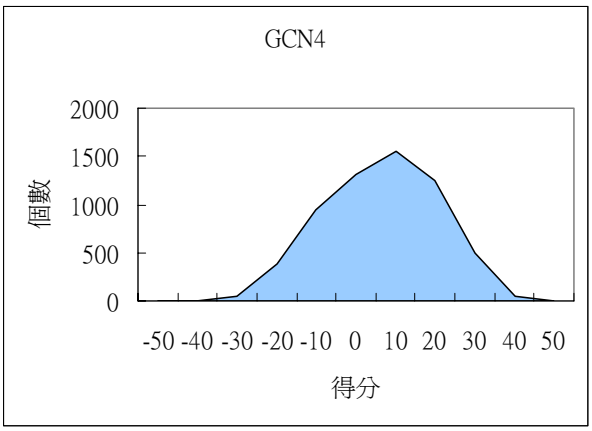



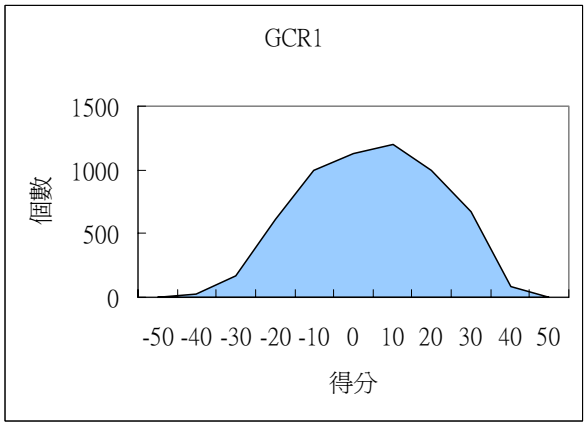



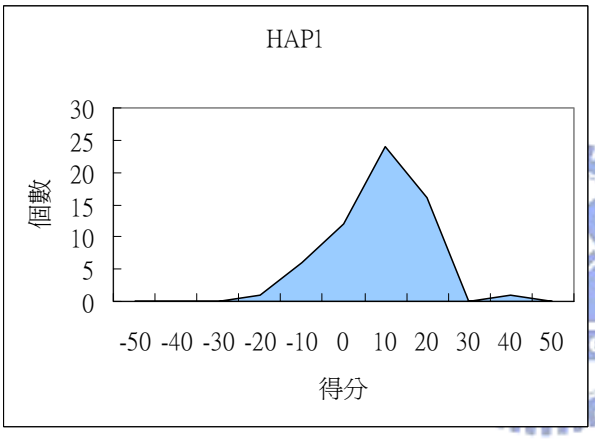



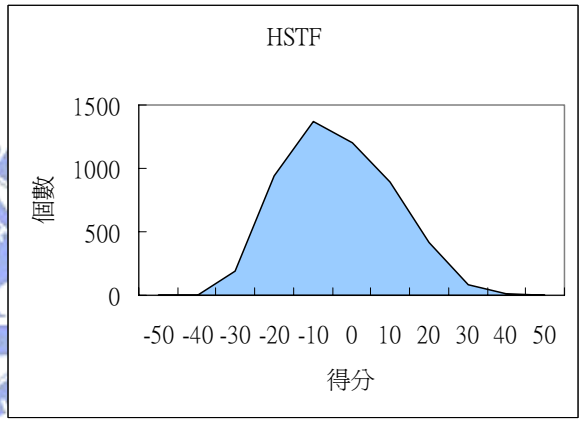



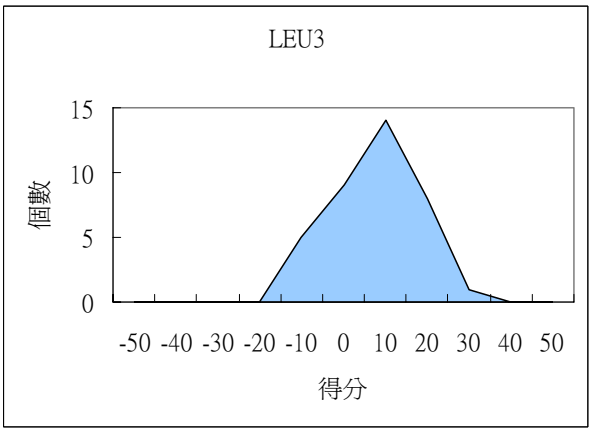



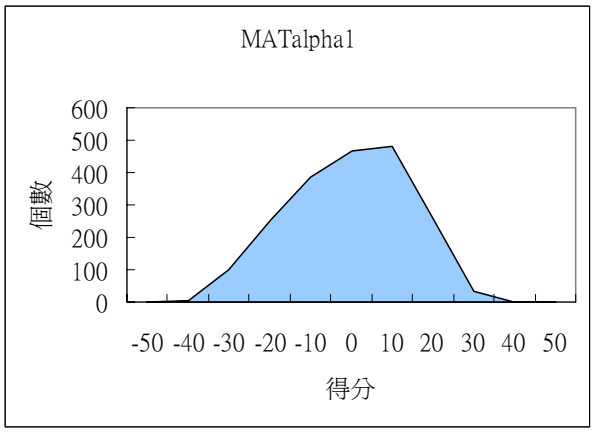

GCN4
個數
2000
1500
1000
500
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

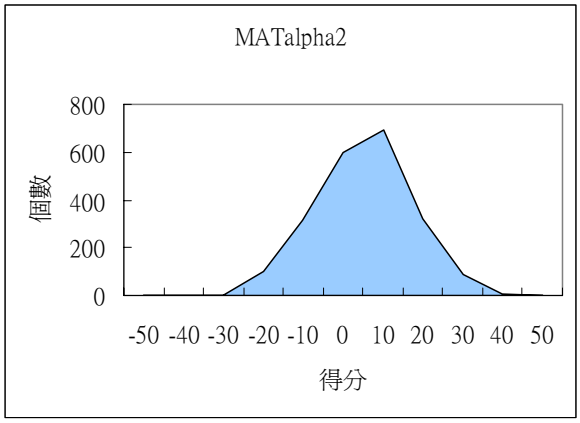

GCR1
個數
1500
1000
500
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

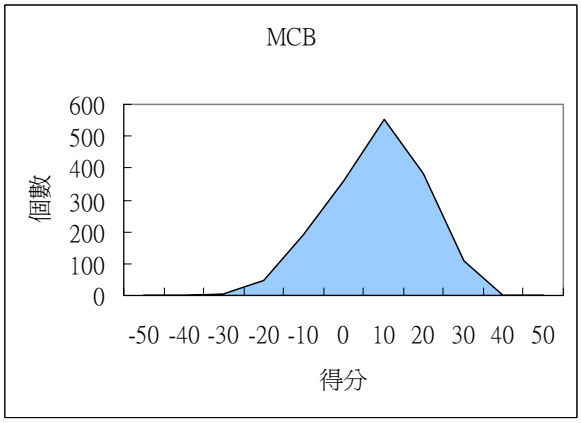

HAP1
個數
30
25
20
15
10
5
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

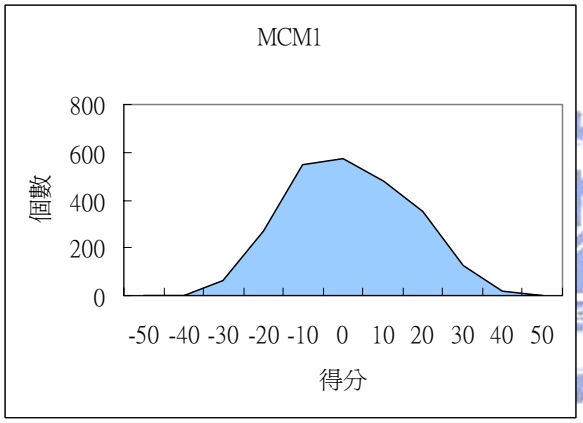

HSTF
個數
1500
1000
500
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

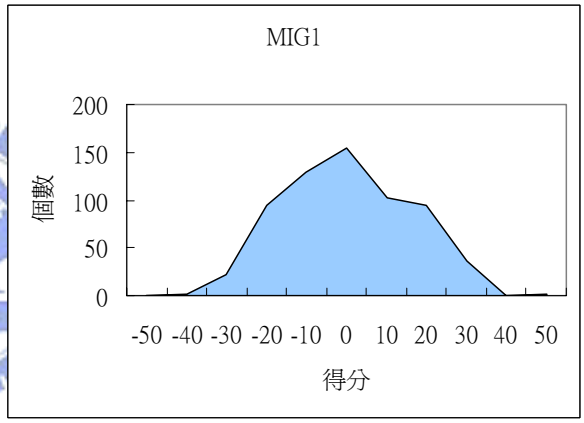

LEU3
個數
15
10
5
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

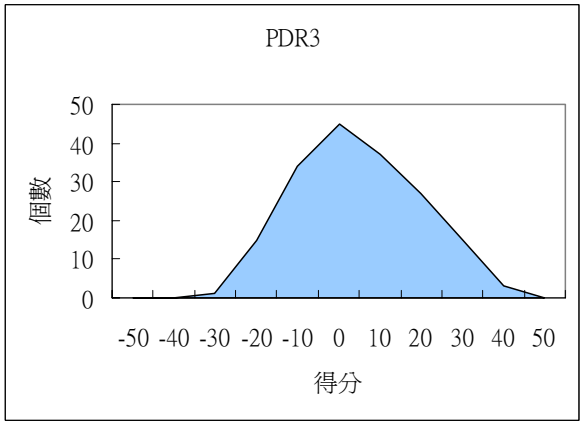

MATalpha1
個數
600
500
400
300
200
100
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

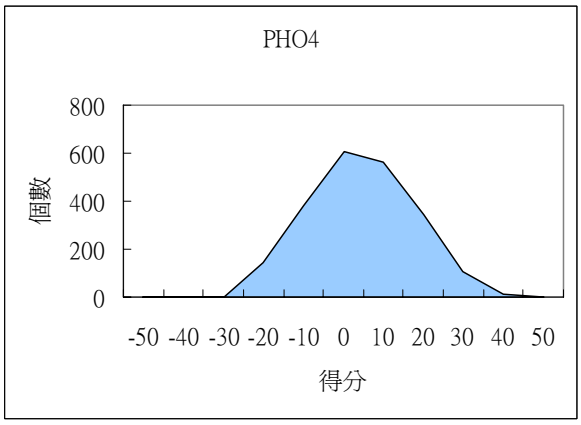

MATalpha2
個數
800
600
400
200
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

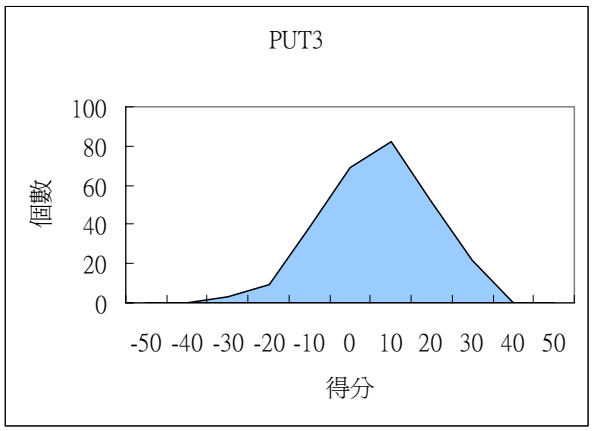

MCB
個數
600
500
400
300
200
100
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

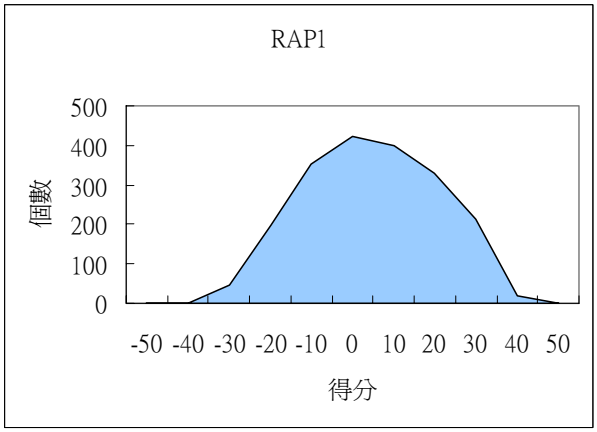

MCM1
個數
800
600
400
200
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

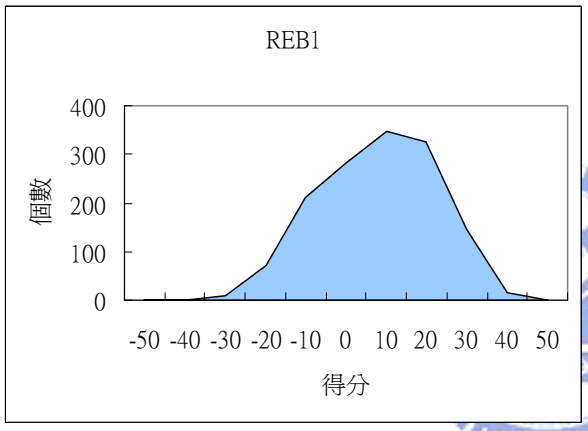

MIG1
個數
200
150
100
50
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

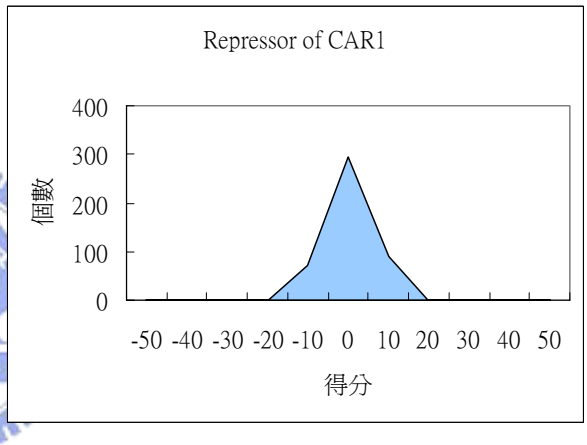

PDR3
個數
50
40
30
20
10
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

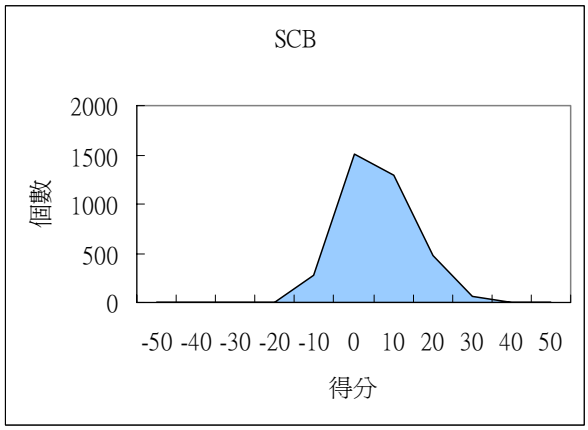

PHO4
個數
800
600
400
200
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

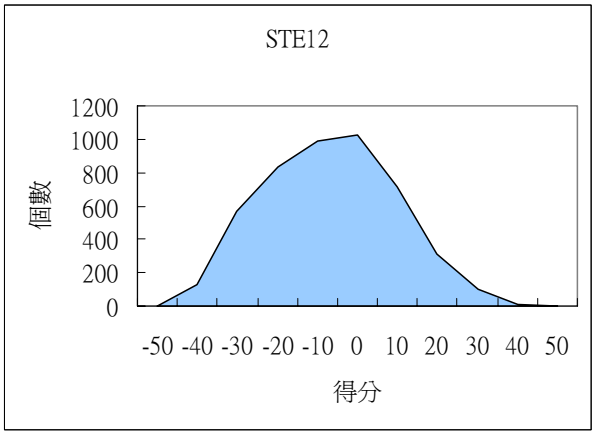

PUT3
個數
100
80
60
40
20
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

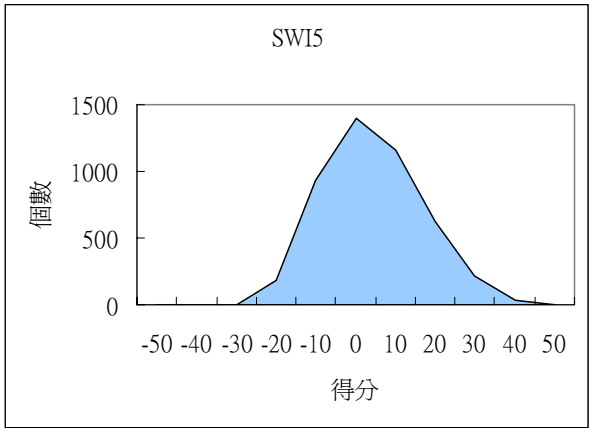

RAP1
個數
500
400
300
200
100
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

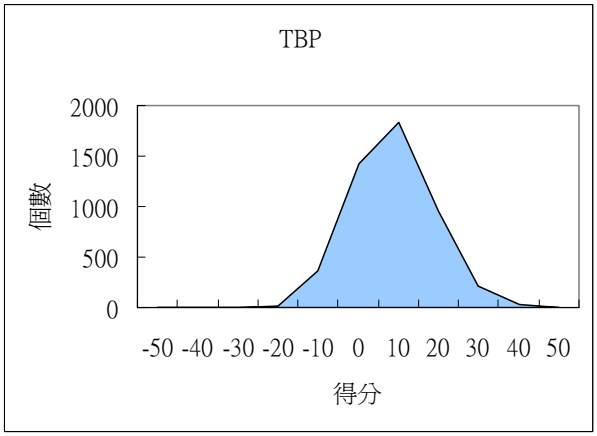

REB1
個數
400
300
200
100
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分


Repressor of CAR1
個數
400
300
200
100
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分


SCB
個數
2000
1500
1000
500
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分


STE12
個數
1200
1000
800
600
400
200
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分


SWI5
1500
1000
500
0
個數
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分


TBP
2000
1500
1000
500
0
個數
-50 -40 -30 -20 -10 0 10 20 30 40 50
得分

# 參考文獻


Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA, 96(12):6745-6750.

Butte AJ, Kohane IS (2000) Mutual Information Relevance Networks:Functional Genomic Clustering Using Pairwise Entropy Measurements.PSB 2000

Cunningham MJ, Liang S, Fuhrman S, Seilhamer JJ and Somogyi R (2000) Gene Expression Microarray Data Analysis for Toxicology Profiling. Annals of the New York Academy of Sciences, 919: 52-67

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW(1998) A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Molecular Cell,2,65–73

D'haeseleer P, Wen X, Fuhrman S, Somogyi R (1999) Linear modeling of mRna expression levels during CNS development and injury. PSB 1999.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 95(25): 14863-14868.

Fuhrman S, Cunningham MJ, Wen X, Zweiger G, Seilhamer JJ, and Somogyi R. (2000) The Application of Shannon Entropy in the Identification of Putative Drug Targets. Biosystems,55,5-14

Fujibuchi W, Anderson JS, and Landsman D (2001) PROSPECT Improves Cis-Acting Regulatory Element Prediction by Integrating Expression Profile Data with Consensus Pattern Searches. Nucleic Acids Res., 29(19): 3988-3996.

Hsu YZ and Hu YJ. (2004) Combining correlations between gene expression profiles with binding sites to reconstruct transcription networks.

Ji L. and Tan KL (2005) Identifying time-lagged gene clusters using gene expression data. Bioinformatics ,21,4,509–516

Kuruvilla FG, Park PJ and Schreiber SL(2002) Vector algebra in the analysis of genome-wide expression data. Genome Biology,3,3

Kwon AT, Hoos HH and Ng R(2003) Inference of transcriptional regulation relationships from gene expression data. Bioinformatics ,19,905–912

Lewis D and Gale W (1994) A sequential algorithm for training text classifiers.

SIGIR1994,3-12

Lee PH and Lee D (2005) Modularized learning of genetic interaction networks from biological annotations and mRNA expression data. Bioinformatics 21,11,2739–2747

Liang S, Fuhrman S, Somogyi R (1998)REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. PSB 1998

Liu TF, Sung WK, Mittal A(2004) Learning Multi-Time Delay Gene Network Using Bayesian Network Framework. Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)

MacQueen JB (1967): "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Murphy K and Mian S(1999) Modelling Gene Expression Data using Dynamic Bayesian Networks.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B.(1998) Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell,9,3273–3297

Schmitt WA, Raab RM, and Stephanopoulos G (2004) Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data. Genome Research,14,1654–1663

Segal E, Barash Y, Simon T, Friedman N, Koller D (2001) From Promoter Sequence to Expression: A Probabilistic Framework.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM (1999). Systematic Determination of Genetic Network Architecture. Nature Genet., 22:281-285.

Tamayo P, Slonim D, Mesirov J, Zhu O, Kitareewan S, Dmitrovsky E, and Golub TR (1999) Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. Proc. Natl. Acad. Sci. USA, 96:2907-2912.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. Nat Genet 22:281-285

van Someren EP, Wesselsa LFA and Reindersa MJT. (2000) Linear Modeling of Genetic Networks from Experimental Data.

Wingender E, Dietze P, Karas H and Knuppel R (1996). TRANSFAC: A Database on

Transcription Factors and Their DNA Binding sites. Nucleic Acids Res., 24:238-241

Yu H, Luscombe NM, Qian J and Gerstein M (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. TRENDS in Genetics,19,8,422-427

Zou M and Conzen SD (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics,21,71–79

Zhu J and Zhang MQ (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics,15,607-611