

國立臺灣大學生物資源暨農學院農藝學研究所生物統計組

碩士論文

Graduate Institute of Agronomy, Division of Biometry

College of BioResources and Agriculture

National Taiwan University

Master Thesis

利用微陣列資料探索大腸桿菌直接與間接之基因調控關係

Using Microarray Data to Explore Primary and Nonprimary

Regulatory Events of *Escherichia coli*

楊西閔

Hsi-Min Yang

指導教授: 劉力瑜 博士

Advisor: Li-yu Daisy Liu, Ph.D.

中華民國九十九年六月

June, 2010

摘要

近年來，微陣列晶片(Microarray)常用來同時檢驗大量的基因表現值並建立基因調控網路。本篇研究則是利用公開下載的大腸桿菌的微陣列資料庫，對於不同的實驗條件下，探討其目標基因與其下游直接與間接調控基因的相關性。我們利用相關係數(Correlation Coefficient)與本質相關係數(Coefficient of Intrinsic Dependence, CID)來檢驗直接與間接調控關係基因的相關性，配合集群分析(Cluster Analysis)與文氏圖(Venn Diagram)，希望可以從這些統計分析中找到微陣列資料上的數據特性，以應用在預測未知的調控關係。

關鍵字：微陣列晶片、相關係數、本質相關係數、集群分析、文氏圖



Abstract

In recent years, microarray have been widely used to simultaneously monitor expression of a large number of genes and to establish gene regulatory networks. In this study, we use the public-available of E. coli microarray database, to rediscovery the regulation events between the target genes and their downstream targets. We use the coefficient of intrinsic dependence and correlation to identify the genes regulated by modE. We also use the cluster analysis to visualize the possible patterns of expression under different regulatory mechanisms. The results may provide insight into the prediction of gene regulatory mechanisms in the future.

Keywords : Microarray , Correlation Coefficient , Coefficient of Intrinsic Dependence , Cluster Analysis , Venn Diagram

目 錄

1	前言	1
2	材料與方法	4
2.1	研究材料	4
2.1.1	微陣列基因表現資料庫	4
2.1.2	基因調控網路資料庫	5
2.2	統計分析方法	7
2.2.1	基因選擇	7
2.2.2	集群分析	9
2.3	分析流程整理	10
3	結果與討論	12
3.1	基因選擇能力比較	12
3.1.1	基因排列次序	12
3.1.2	顯著基因個數	15
3.2	群集分析	20
4	結論	25
	參考文獻	27
	附錄一	30

表目錄

1.1 基因調控網路模型 3



圖目錄

2.1 利用 Pathway Tools 顯示大腸桿菌 modE 基因調控網路。	6
3.1 合併排序折線圖	14
3.2 顯著基因個數文氏圖 (Mo Dataset)	17
3.3 顯著基因個數文氏圖 (Non Mo Dataset)	18
3.4 顯著基因個數文氏圖 (G445 Dataset)	19
3.5 群集分析分群樹狀圖 (Mo dataset)	21
3.6 群集分析分群樹狀圖 (Non Mo dataset)	22
3.7 群集分析分群樹狀圖 (G445 dataset)	23
3.8 不同分群數之錯分率曲線圖	24

第一章 前言

生物技術不斷的改進下，生物技術重心已漸漸從以往的定序實驗轉移到現在基因資料庫分析，當基因定序實驗已日漸成熟的今日，學者們的下一個步驟即是為基因所包含的訊息做解碼。在大量的基因資訊下，透過統計的分析，挑選出真正影響生物體功能之基因，將可縮小實際實驗的成本與時間。一般在探索基因資訊，最常採用的技術是基因微陣列 (Microarray) 技術。

微陣列晶片早在1995年已被提出(Schena *et al.*, 1995)，係指在微小面積的基質上種植高密度的生物探針，做為大量篩檢及平行分析的工具。微陣列晶片具備快速、方便、經濟、省時等特性，適用於大量基因表達、篩檢、及比對等研究，可以應用於病原體基因檢測、基因表現比較、基因突變分析、基因序列分析、及新藥物開發等領域 (Spellman *et al.*, 1998; Golub *et al.*, 2002)。

cDNA 微陣列晶片是利用 cDNA 作為探針 (Probe)，以高密度點陣固定在經表面化學塗布處理過的玻璃載體表面上，而受測的檢體則是 mRNA。將玻璃片與檢體進行雜合試驗 (hybridization)，由於DNA為雙股螺旋結構具有互補的專一特性，就如同拉鍊般的性質，檢體中的標的核酸，會雜合固定在cDNA微點陣玻璃上含有互補的核酸序列的探針的點；再經過清洗將沒有雜合的樣本核酸去掉，就可以記錄下有雜合反應的點的位置。因此，只需要一次的檢驗，cDNA微點陣能夠將成千上萬的基因表現的樣式 (gene expression pattern) 記錄下來。目前的方法容許偵測到極微細的細胞內變化，甚至可測得一個細胞內少數幾個訊息RNA的改變，使得研究者能透過晶片上的數據得整體性訊息。

當微陣列晶片應用在多種生物體及各項不同的實驗條件下，晶片資料量亦逐漸累積，即出現了微陣列晶片資料庫 (Microarray Database)；微陣列晶片資料庫蒐集了大量的晶片資料，建立於網際網路伺服器上，讓使用者可以方便快速的查詢與利用，不但可以統合大量晶片資料，也可節省重複實驗的時間與資源，以前的晶片樣本數稀少的問題，也透過資料庫的利用，得以解決(Faith *et al.*, 2007)。目前較常使用的資料庫如：Stanford Microarray Database(SMD), Gene Expression Omnibus(GEO), Many Microbe Microarrays Database(M3D), RegulonDB, 與 EcoCyc 等 (Salgado *et al.*, 2005)。

在生物體中，DNA序列可轉錄產生RNA，而RNA又可以轉譯產生維持正常生物體運作機能所需要的蛋白質，但外在環境的變化與刺激，生物體為維持正常運作機能，基因的表現則會改變，這種基因模式的改變，稱之為基因調控(Gene Regulation)，在生物體內，基因調控往往並非是簡單的一對一的調控，而是複雜的網狀相互調控。利用基因微陣列資料配合統計分析，可找出基因與基因之間的調控關係，例如 Sheen-Orr *et al.* (2002) 中提出如一對一、一對多、多對多等基因調控模型。而由簡單調控模型建構而成之基因間鍊狀或是網狀的關係，我們稱之為基因調控網路 (gene regulatory networks, GRN)(Davidson *et al.*, 2006; Shen-Orr *et al.*, 2002)。現階段有許多學者利用不同的方法與不同的資料，也是想透過建立基因調控網路，來了解基因之間的交互作用，表 1.1 列出目前常用的基因調控網路建構方式 (Zare *et al.*, 2009)。

本篇研究是利用資料庫內大腸桿菌微陣列晶片資料，配合已確認之大腸桿菌基因調控網路關係，利用相關係數與本質相關係數 (Hsing *et al.*, 2005) 對目標基因的各種基因調控關係進行統計分析，將兩種相關分析結果互相比較，期望能找到不同調控關係的特性，並在未來作為判定不同調控關係的基礎。

表 1.1: 基因調控網路模型

基因網路建立模式	特性
Differential Equation Models	利用時間序列型(time-course)的mRNA資料，可建立小型且量化的基因網路。(Singer <i>et al.</i> , 1973)
Boolean Networks	利用時間序列型(time-course)的mRNA資料，可建立小型的基因網路。(Akutsu <i>et al.</i> , 2000)
Bayesian Networks and Graphical Models	測量邊際與獨立條件的基因，利用mRNA資料可以建立大型且高複雜的基因網路。(Friedman <i>et al.</i> , 2000)
Relevance Networks	測量線性與非線性的相關性基因，可建立不具方向性的基因網路。(Butte <i>et al.</i> , 2002)
Matrix Decomposition	利用基因表現量資料配合完整的潛在相關性網路資訊，可以重新定義和量化基因網路。(Liao <i>et al.</i> , 2003)
Supervised Methods	利用轉錄因子配合部分相關性網路資訊，可建立基因網路。(Mordelet <i>et al.</i> , 2008)

第二章 材料與方法

在基因調控網路中，直接調控關係的基因，是比間接調控關係的基因較容易從基因表現量中被檢測出來 (Liu *et al.*, 2009)。本研究主要目的是嘗試利用網路上可公開下載之微陣列資料重建已知的大腸桿菌調控網路，分析材料之來源將於第一節詳細介紹。分析流程的第一部分是探討利用統計方法搜尋直接或間接調控的下游基因，第二部分是透過集群分析解析已知的直接與間接調控關係可能具有之基因表現型態；其中第一部分又常被稱為特徵選擇或基因選擇。本章第二節將分別敘述基因選擇與集群分析所採用的統計方法。第三節則將本論文的資料分析流程做一整理。

第一節 研究材料

本節分別敘述「微陣列基因表現資料庫」與「基因調控網路資料庫」的來源。

一、 微陣列基因表現資料庫

為了探討利用微陣列資料重建大腸桿菌基因調控網路的可行性，我們從公開資料庫中下載三筆資料進行分析，分別敘述如下。

第一筆資料下載至 M^{3D} database (<http://m3d.bu.edu>)，包含 445 片 Affymetrix Antisense2 晶片、共 7312 probes (4345 基因) 表現的結果 (Faith *et al.*, 2007)，是結合不同實驗室在不同處理條件下 (如：酸鹼值、生長期、熱休克、氧氣濃度、基因干擾等) 所得之微陣列基因表現，其中混合了時間序列 (time-course) 與非時間序列 (steady-state) 的數據，該資料庫提供利用 MAS5、RMA、GCRMA、Dchip PM 等四種方式移除陣列間差異後的基因表現結果。本論文僅採

用RMA校正後的微陣列資料作為分析的材料，簡稱為「G445 資料集」(G445 dataset)。

第二筆與第三筆資料為 G445 資料集的子集合。為了解重建 modE 調控網路最適合採用的微陣列資料型態，我們從 445 晶片資料中選出 70 組共145個樣本與 modE 調控相關的實驗 (第二筆資料) 與 152 組共362個樣本與 modE 調控無關的實驗 (第三筆資料) 分別作為分析材料。modE 基因是一個對鉬酸鹽反應的轉錄因子 — 鉬酸鹽可與modE基因結合，形成一種modE-molybdate的化合物，藉以增進鉬酸鹽相關功能的操縱子表現效果(Anderson *et al.*, 2000)。在本篇研究中，我們之所以選定modE作為我們有興趣的目標基因，是因為其調控網路的複雜度較低，根據基因調控網路資料庫 (詳見下節)，該基因只作為調控的角色，並無其它基因調控之，因此也是一個較單純的調控網路。

第二筆資料所採用與基因 modE 調控相關的實驗資料，係指實驗內容為加入濃度 2.91×10^{-6} mM 的鉬酸胺(ammonium molybdate) 處理後的 E. Coli 基因表現值，再經過RMA校正後所得結果，本文中簡稱「Mo資料集」(Mo dataset)。而第三筆資料採用的是與基因 modE 調控無關的實驗，其資料是利用 M3D 網站中全部的 E. Coli 資料，除去第二筆資料及時間序列型態的資料後，所留下來的綜合性實驗基因表現值，再經過RMA校正後所得結果，本文中簡稱「非Mo相關基因集」(Non Mo dataset)。Mo dataset 與 Non Mo dataset 各包含 4298 個基因表現結果。

二、 基因調控網路資料庫

本篇報告所使用來探索大腸桿菌之基因調控網路的工具，是名為 Pathway Tools 的軟體，由 Peter D. Karp 和其工作團隊在 SRI 公司的生物訊息研究小組中所展發的，該程式是建立在 Pathway/Genome Database (PGDB) 上，主要包含了二個公開的資料庫 EcoCyc (<http://ecocyc.org/>) 和 HumanCyc (<http://humancyc.org/>)

資料庫，而 Pathways Tools 整合了二個資料庫後，更提供了離線的創建、編輯、查詢及視覺化的圖形呈現，並提供學術上的免費使用 (<http://biocyc.org/download.shtml>)。在本篇文章中，我們是利用它的大腸桿菌基因調控網路資料庫，該資料庫所提到的調控機制均有相關文獻利用實驗佐證其正確性 (literature-based curated)。圖 2.1 為 Pathway Tools 所呈現的大腸桿菌 modE 基因調控網路，共有 145 個基因受 modE 調控。

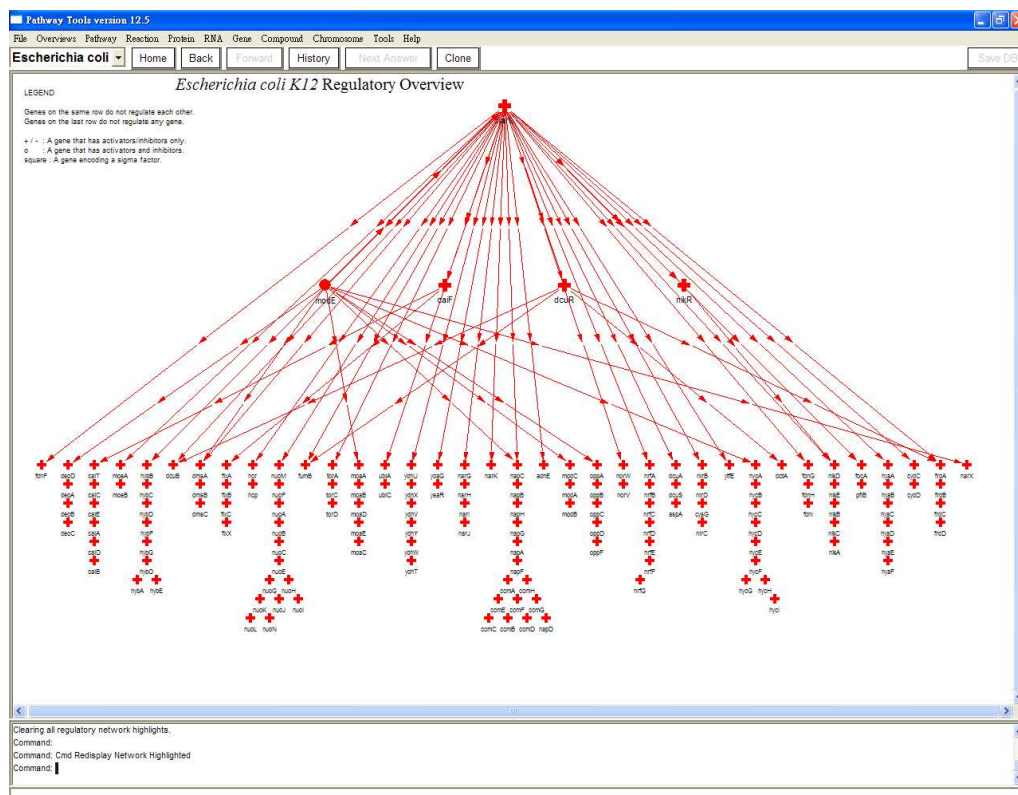


圖 2.1: 利用 Pathway Tools 顯示大腸桿菌 modE 基因調控網路。

利用 Pathway Tools 可將受 modE 調控的 146 個下游基因依不同的調控關係加以分級如下：

Level 1(Lv. 1): 由 modE 直接調控的基因，包含 deoD 等共 28 個基因。

Level 2(Lv. 2): 由受 modE 直接調控的基因所直接調控的基因，共 78 個。例如基因 hcr 即屬此類：modE → narL → hcr，其中箭頭代表直接調控關係。

Level 3(Lv. 3): 例如基因 caiT (modE → narL → caiF → caiT) 等

共11個基因。

Level 2 & 1(Lv. 1;2): 既屬於 Lv. 1 也屬於 Lv. 2 的基因，例如基因 napC (modE → napC 或 modE → narL → napC) 等共18個基因。

Level 3 & 2(Lv. 2;3): 既屬於 Lv. 2 也屬於 Lv. 3 的基因，例如基因 nikE (modE → narL → nikE 或 modE → narL → nikR → nikE) 等共11個基因。

146 個基因的分級結果詳列於附錄一。

第二節 統計分析方法

一、 基因選擇

1. 相關係數 Correlation Coefficient

假設 X 和 Y 為隨機變數，其平均值為各為 μ_X 和 μ_Y ，其共變異數 (covariance) 為：

$$Cov(Y_1, Y_2) = E[(X - \mu_X)(Y - \mu_Y)],$$

相關係數 (correlation coefficient, ρ) 是一種由共變異數除去各變數之單位的純量，其定義為：

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

其中 σ_X 和 σ_Y 為 X 與 Y 的標準差 (standard deviations)。由樣本計算相關係數估計值為：

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_j (X_j - \bar{X})^2 \sum_k (Y_k - \bar{Y})^2}},$$

其中 \bar{X} 與 \bar{Y} 為樣本均值。

相關係數介於 -1 至 1 之間，越接近 1 (或 -1) 表示 X 與 Y 越有明顯的正比 (或反比) 線性關係。相關係數是最常被用來偵測基因調控關係的統計方法之一 (Liu *et al.*, 2009)，當相關係數越接近 1 (或 -1)

表示兩基因間的正向 (或逆向) 調控關係越強。本篇論文爲了同時考慮正向與逆向調控關係，我們將相關係數 r 取平方，可以得到決定係數 (coefficient of determination, R^2)，其範圍介於 0 至 1 之間。我們利用決定係數來檢定目標基因與其它基因之間的相關性， R^2 越接近 1 表示兩者相關性越大，反之越接近 0 則兩者相關性越小。若隨機變數 X 與 Y 服從二維常態分佈且互相獨立，其相關係數的抽樣分佈近似於自由度爲 $n - 2$ 的學生氏 t 分佈，因此欲檢定兩隨機變數是否獨立時之 p value 可藉由自由度爲 $n - 2$ 的學生氏 t 分佈計算得之。

2. 本質相關係數 Coefficient of Intrinsic Dependence(CID)

在給定隨機變數 X 下連續隨機變數 Y 之本質相關係數 (coefficient of Intrinsic Dependence, CID) 定義爲：

$$\text{CID}(Y|X) = \frac{\int_{-\infty}^{\infty} \text{Var}(E(\delta_Y(u)|X))dF_Y(u)}{\int_{-\infty}^{\infty} \text{Var}(\delta_Y(u))dF_Y(u)}, \quad (2.1)$$

本式裡的 F_Y 是 Y 的累積機率密度函數， $E(\delta_Y(u)|X)$ 是隨機變數 $\delta_Y(u)$ 給定 X 的條件期望值，且 $y \in (-\infty, \infty)$ ， $\delta_y(u) = I(y \leq u)$ ， $u \in (-\infty, \infty)$ 。由樣本計算本質相關係數估值爲：

$$\hat{\text{CID}} = \frac{1}{C(n)} \sum_{j=1}^k \frac{n_j}{n^2} \sum_{i=1}^n \left\{ \hat{F}_{n_j}(y_i) - \hat{F}_n(y_i) \right\}^2,$$

其中 $C(n) = 1/6 - 1/(6n^2)$ ，爲計算 Y 給定 X 之條件分佈估值 $\hat{F}_{n_j}(y_i)$ ，需將成對的 X 與 Y 觀測值分爲 k 個子集合，第 j 個子集合之樣本大小爲 n_j ，則

$$\hat{F}_{n_j}(y_i) = \frac{1}{n_j} \sum_{l=1}^{n_j} I(y_l \leq y_i \text{ 且 } y_l \in \text{第 } j \text{ 個子集合});$$

而 Y 的邊際分佈估值爲

$$\hat{F}_n(y_i) = \frac{1}{n} \sum_{l=1}^n I(y_l \leq y_i)。$$

我們利用本質相關係數來檢定目標基因與其下游基因之間的相關性，其範圍介於 0 至 1 之間，越接近 1 表示目標基因與下游基因相

關性越大，反之越接近 0 則目標基因與下游基因相關性越小。由於本質相關係數具有不對稱的特性 (亦即 $CID(Y|X) \neq CID(X|Y)$)，本論文將分別計算以目標基因 (modE) 之表現量作為 X 或 Y 之本質相關係數值。欲檢定兩隨機變數 (基因) 是否獨立時，我們隨機排列 X 與 Y 之觀測值並計算本質相關係數以模擬在兩變數獨立時之本質相關係數結果，重複模擬 10000 次後計算其中大於原始之本質相關係數之個數後，再除 10001，可以得到該 X 與 Y 之本質相關係數之 p value。本質相關係數無需對隨機變數 X 與 Y 之分佈做任何假設，且 X 與 Y 之相關性不限於線性關係，特別適用於微陣列資料之分析，能廣泛偵測到基因間不同類型的調控機制 (Liu *et al.*, 2009)。

3. 基因選擇之分析流程

搜尋並分析 modE 下游基因方法如下：

- 1) 計算每一個基因與 modE 的決定係數和本質相關係數及兩者之 p value。
- 2) 依不同調控級別，比較決定係數和本質相關係數之顯著基因個數。
- 3) 依不同調控級別，計算該級別內的基因決定係數和本質相關係數對於晶片的全部基因中的次序，並製作趨勢圖以判定並比較各種統計方法選擇下游基因的能力。
- 4) 依不同調控級別加入最低合併次序趨勢圖。最低合併次序定義為：在同一筆資料下，對全部的基因做三種相關係數計算後，加以排序列名次；每個基因即包含了三種統計方法的三種名次，我們則取最低的名次與該名次所屬的統計方法，將各層級的基因群之最低次序合併後，即可以描繪出該群基因的相關係數方法的變化趨勢。合併次序趨勢圖主要是在討論各統計方法是否找到相同的下游基因。

二、 集群分析

群集分析 (Cluster analysis) 主要的目的是將資料做分群，而分群的依

據則是利用資料間的差異性，將差異較小的資料分成一群，而群與群之間的異差較大。群集分析常用在探索大量資料可能存在的趨勢，本文則利用此分析法來對已選定的基因群試作分群，以觀察資料特性與分群結果。由於不預設分群數，所以利用摺疊集群法 (Hierarchical cluster) 中的完全聯結法(Complete Linkage) 對 modE 相關基因群做摺疊集群分析，依其基因表現量為觀測值，將 146 個基因 (含 modE 與其 145 個下游基因) 當作集群中的個體，對其分群，使每個群集中之個體有最相近的特徵。其集群法演算過程 (clustering algorithm) 如下：

- 1) 將 146 個基因各自為一群，共有 146 個小群，即是每小群只包含一個基因，設第 i 個基因的表現量為 $\{X_{i1}, X_{i2}, \dots, X_{ip}\}$ ，其兩兩基因間的距離構成的距離矩陣為

$$D_n = \{d_{ij}\}, d_{ij} = \text{歐氏距離} = [\sum_{K=1}^p (X_{iK} - X_{jK})^2]^{1/2}。$$

- 2) 在距離矩陣中尋找最小的 d_{ij} ，將有最小距離的基因 i 與基因 j 合併成群。
- 3) 利用完全連結法求群與其它群之間的距離為兩群間兩個最遠個體的距離。
- 4) 重複步驟 2 與 3 依序合併最小距離的兩群，最後成一大群。

集群分析完成後，將其分群成 2 至 10 群，觀察其每種分群之錯分率 (misclassification rate) 之差異來判斷何種分群數較為合適。

第三節 分析流程整理

彙整本論文資料分析流程如下：

步驟1：取得大腸桿菌微陣列晶片

本篇文章中，所分析的資料有三種：1.非特定的綜合實驗微陣列晶片 (G445 dataset)；2.與目標基因相關實驗之微陣列晶片 (Mo dataset)；3.與目標基因非相關實驗之微陣列晶片 (Non Mo dataset)。

步驟2：資料預處理

在非特定的綜合實驗微陣列晶片這組資料中，我們只取出其中的RMA正規化後的基因表現量資料，當作我們分析的對象；而目標相關與非相關的微陣列晶片則是已經正規化後的基因表現量資料，可以直接進行分析。配合後面的分析，我們將把各組資料分成全基因資料（包含所有基因）與目標基因相關的部分資料（僅包含146 modE下游基因）。對於目標基因與其直接與間接調控關係的基因群，我們按照不同層級的調控關係，將其分成數群。

步驟3：全基因資料分析

針對全基因資料，計算每個基因的決定係數與本質相關係數的統計值後，再把目標基因與其相關調控基因依不同層級篩選出來，對這些基因在全基因裡面依不同的統計值做排名。因此，我們可以得到在不同的統計方法中不同層級的目標基因群在全基因的名次，從中觀察不同的統計法的名次變化。

步驟4：部分基因分析

對於部分基因資料，我們計算目標基因群的各種統計方法的p-value，對於這些p-value我們將其依照不同層級其不同的顯著水準做文氏圖（Venn Diagram），去觀察同資料下何種統計方法可以篩選出較多的基因及不同的資料在篩選結果有何不同。

步驟5：部分基因集群分析

利用部分基因的基因表現量做集群分析，觀察各種不同的調控分層與集群分析的分群有何相關性與特徵。

第三章 結果與討論

第一節 基因選擇能力比較

一、 基因排列次序

圖 3.1 為三種資料 (G445、Mo、Non Mo datasets) 在不同的調控層級下，不同統計方法所得的排序結果，其中 x 軸為選取基因個數， y 軸為選取基因中屬於 EcoCyc 資料庫特定層級的 modE 下游基因數。圖中黑線代表相關系數排序折線圖；紅線代表給定 modE 基因表現量下，CID 值排序折線圖 (以 modE 基因表現量為 (2.1) 式中的 x 、其他基因為 (2.1) 式中的 y ，以下簡稱 cidx)；綠線代表給定被調控基因的基因表現量下，CID 值排序折線圖 (以 modE 基因表現量為 (2.1) 式中的 y 、其他基因為 (2.1) 式中的 x ，以下簡稱 cidy)。圖 3.1 中由點與線構成的折線圖為最低合併次序趨勢圖。

由 modE 直接調控的第一層級 (lv1) 基因來看，Mo 與 Non Mo 資料下，可以發現單獨的統計方法對全基因名次排序中，cidy 與 cidx 表現的趨勢相似，而且比 cor 的方法可以找到較低名次的基因，而在 G445 的資料中則是與 Mo、Non Mo 呈現不同的趨勢，cor 方法在前半部分的基因比 cidx 與 cidy 方法較好，而後半部則是 cidx 略佳一些。

再看到 modE 間接調控的 lv2 層級，與 lv1 呈現相似的趨勢；Mo 與 Non Mo 資料下，cidy 與 cidx 比 cor 方法較容易找到名次較低的 lv2 基因群，而 G445 則是 cor 與 cidx、cidy 各佔一半，與 lv1 不同的是，前 1/3 的基因群，以 cidx 表現較好，而中段基因則是 cor 較好，後段基因則是差不多。

再看到lv3的調控層級，三種資料都可以看到共同的趨勢：cor比cidx與cidy方法找到同基因較低的名次，而且與合併排序折線大多重疊在一起。

最後看到多重層級調控關係(lv 1;2與lv 2;3)下的基因群排序趨勢，在Mo與Non Mo資料下，通常cidx較其它二種方法找到同基因較低的名次，而G445的資料則是以cor方法較其它二種方法略佳。

從合併的折線圖來看的話，通常都是優於單獨一種統計值的排序方法，只有lv3的調控層級下，大部分都是cor方法最快找到最低名次的基因，因此與合併折線圖的趨勢幾乎重疊。



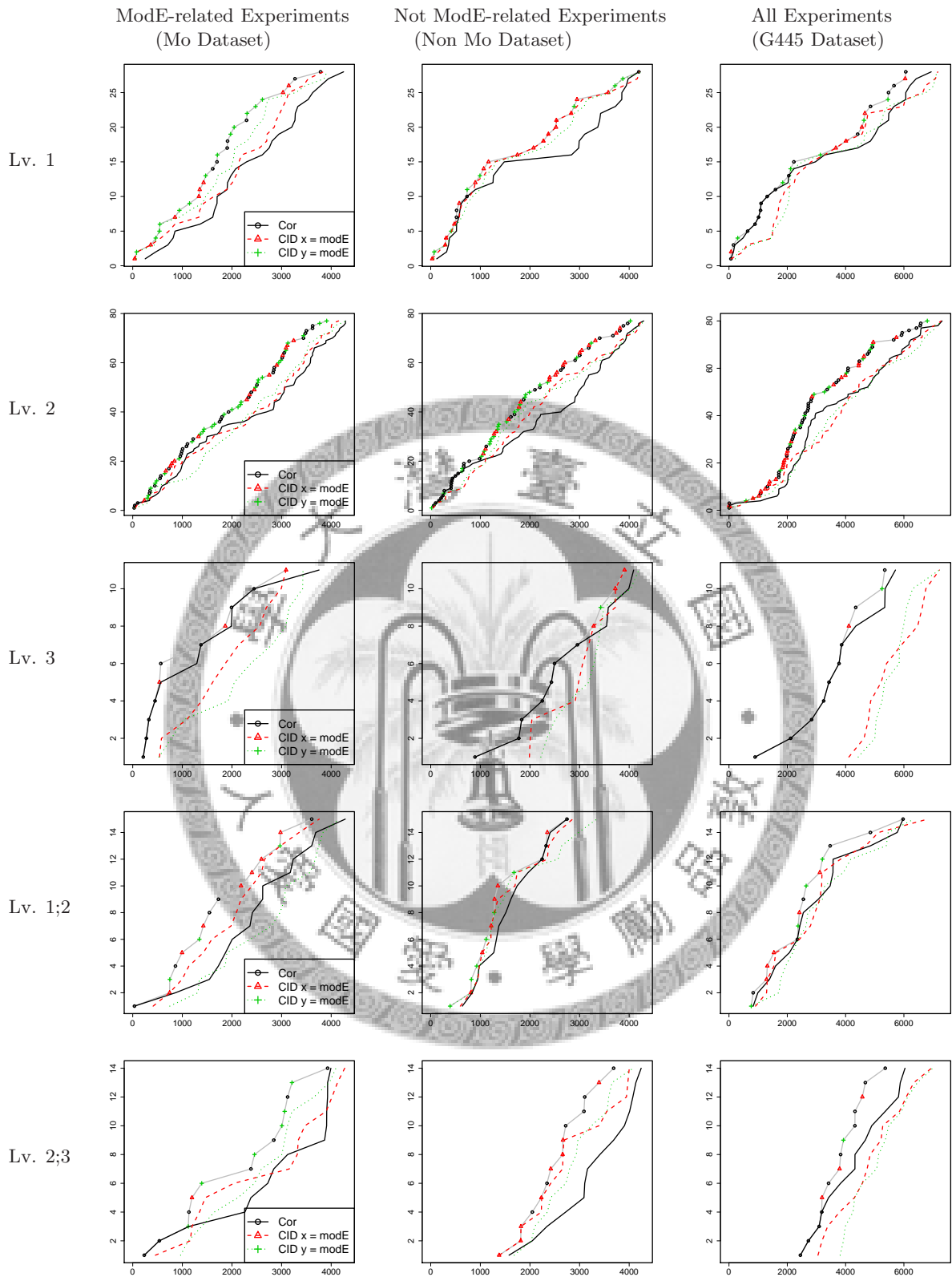


圖 3.1: 合併排序折線圖。三種資料在不同的調控層級下，不同的統計方法排序折線圖。黑線代表決定係數排序折線圖；紅線代表給定ModE基因表現量下，CID值排序折線圖；綠線代表給定被調控基因的基因表現量下，CID值排序折線圖。

二、 顯著基因個數

圖 3.2、3.3、3.4 比較三種資料 (G445、Mo、Non Mo datasets) 在不同的調控層級下，不同統計方法找到的顯著基因個數，代表各統計方法的檢定力。從 modE 直接調控的第一層級 (lv1) 來看，可以從三筆不同的資料發現一些特性，三種統計方法都有找到的基因數量較單獨一種方法找到的數量較多，其次則是在 cidx 與 cidy 這二種方法的交集下找到的基因數量較 cor 或是其與 cid 交集的數量都較多，從每一筆資料來看的話，Mo 這筆資料落在三種方法之外的基因較其它二筆資料多，且三種統計方法皆找到的基因也較其它二筆資料少。

再來看到 modE 的第二層級調控基因群 (lv2)，亦可以發現與 lv1 時的資料特色，三種統計方法的交集基因較其它單獨或是兩種交集方法的基因較多其次是 cidx 與 cidy 的交集基因數量雖然沒有三種方法交集下的數量多，但明顯多過其它方法，成為第二多的群組，而這個特性在 Mo 資料中就不若其它二組資料的明顯，在 Mo 資料中反而 cor 與 cidy 的交集和其自身的數量較多，另一個與 lv1 資料相同的特性就是 Mo 資料落在三種統計方法外的數量也較其它二筆資料多。

接下來的三個層級可能是因為本身所包含的基因總數較少，所以不像第一、第二調控層級有明顯的特性，從第三調控層級中可以發現與之前共同的特性就是在三種統計方法的交集中，包含最多數量的基因，而 Mo 資料下，cor 方法可以包含的基因較 cidy 與 cidx 較多，但其它二筆資料則剛好相反，cidx 與 cidy 所包含的基因較 cor 的多。三種方法外的基因也已經減少到沒有明顯的差距了。從多種層級調控 (lv1;2,lv2;3) 下的基因群來看，lv1;2 的層級下，特別不同的的資料特性出現在 Non Mo 這筆資料中，三種顯著水準下，全部的基因都被三種方法所包含，与其它二筆資料有明顯的不同。而 G445 的資料則是與前幾面的層級特性相似，三種方法交集下的基因數量最多，cidy 與 cidx 的交集次之，cor 方法下的基因數量最少或是沒有。Mo 的資料則是隨著顯著水準的差著，有不同的變化，比較一致的結果是在 cidy 的方法下，不同的顯著水準都包含了主要的基因數量。而 lv2;3 因為

總數基因已經減少很多，所以前面層級的資料特性都不明顯。但從落在三種方法外的基因數來看，Mo 資料與前面層級的一樣的特性，都較其它二種資料多。

整體來看的話，可以發現一些資料上面與不同層級的特性，三種統計方法的交集總是包含最多數量的基因，而其次的數量是在 cidx 與 cidy 的交集中。而 Mo 資料的特性與 Non Mo 與 G445 較不同的是，其落在三種統計方法外的基因數都較多。且 cidy 與 cidx 的交集基因數量也較少。Cor 方法找到的基因會較其它資料多一些。在多重層級的調控方面，三種資料與三種統計方法可能因為基因總量太少，而沒有明顯的資料特性。



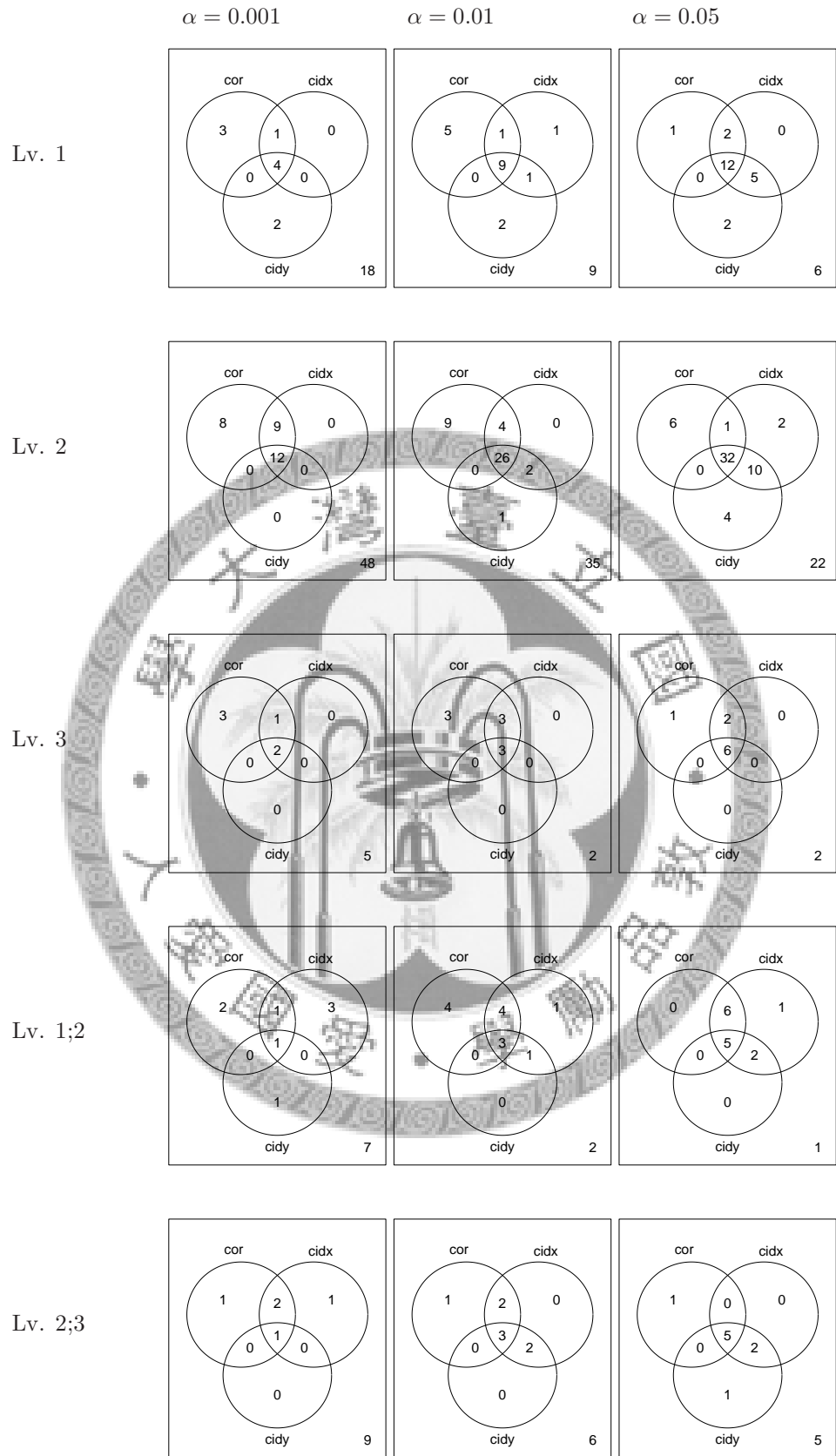


圖 3.2: 顯著基因個數文氏圖 (Mo Dataset), 顯示在不同的調控層級下, 不同的統計方法判定顯著的基因數量。

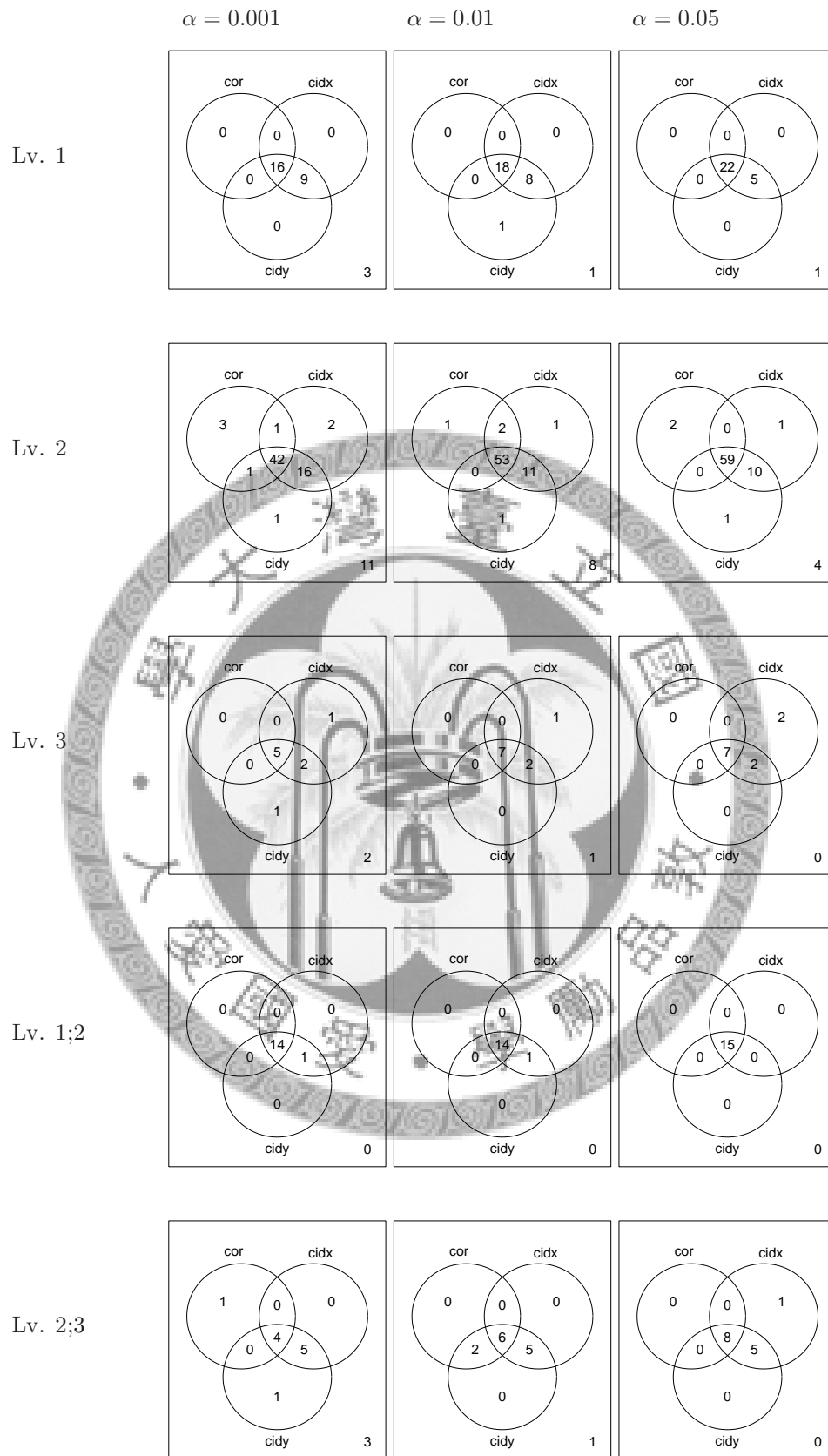


圖 3.3: 顯著基因個數文氏圖 (Non Mo Dataset), 顯示在不同的調控層級下, 不同的統計方法判定顯著的基因數量。

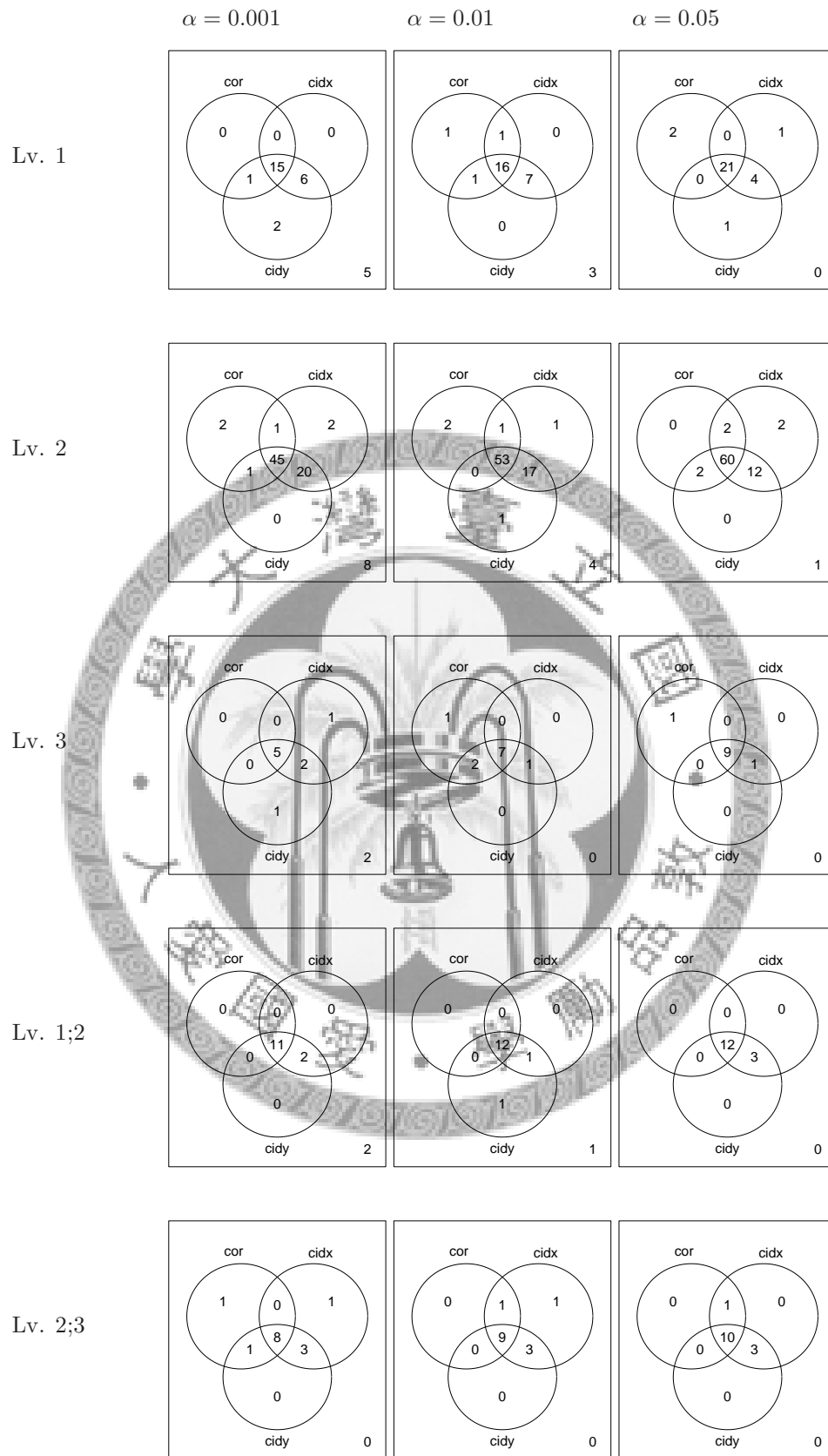


圖 3.4: 方法顯著基因個數文氏圖 (G445 Dataset), 顯示在不同的調控層級下, 不同的統計方法判定顯著的基因數量。

第二節 群集分析

圖 3.5、3.6、3.7 顯示集群分析結果。從 Mo 的資料來觀察分群狀況，可以發現 lv1 (圖上標記1) 與 lv2 (圖上標記2) 的基因較多集群且集群範圍大 (3個或是3個以上的基因集群)；而 lv3 以上(標記為 3、1;2、2;3) 則集群的效果就較不明顯。但還是有部分集群的效果 (2、3個基因集群)。可能因為 lv3 以上調控的關係較為複雜，也有可能基因個數減少的關係。從 Non Mo 的資料來看，一樣有 lv1、lv2 明顯的大分群，而其他較不明顯，但與 Mo 資料不同的是，多重調控 lv1;2 (標記為1;2) 在 Non Mo 的資料中，有較 Mo 的資料中明顯分群，且集群的基因數到達一群中有 5 個相鄰。再看到 G445 的資料，就與前面兩筆不同，雖然 lv1、lv2 有集群的效果，但相鄰的個數大量減少，而且不集中，而 lv3 以上則更分散在資料之中，沒有明顯的集群出現。

我們利用錯分率來對 cluster 的分群結果做討論，由圖 3.8 的結果來看，我們可以發現 G445 的資料錯分率明顯與 Mo、Non Mo 的資料錯分率較高，而 Mo 與 Non Mo 的資料錯分率則有相似的錯分率趨勢。另外我們可以從錯分率的變化趨勢來討論，G445 的資料若要把錯分率降至 0.3 左右，則分群數要達到 70 左右，而 Mo 與 Non Mo 資料則只需要分 30 群左右即可達到，而錯分率要降至 0.2 左右時，G445 資料則需要分群到 90 左右，Mo 與 Non Mo 則是大約50群左右。因此，我們可以推測，如果資料需要集群分析，則同質性較高的 Mo 與 Non Mo 資料的錯分率會比較好一些，但實際上對於我們已知的五種調控層級來說，如果只把資料分成 5 群，三種資料的錯分率都超過 0.4 以上，因此只分 5 群的結果並沒有較好的錯分率。

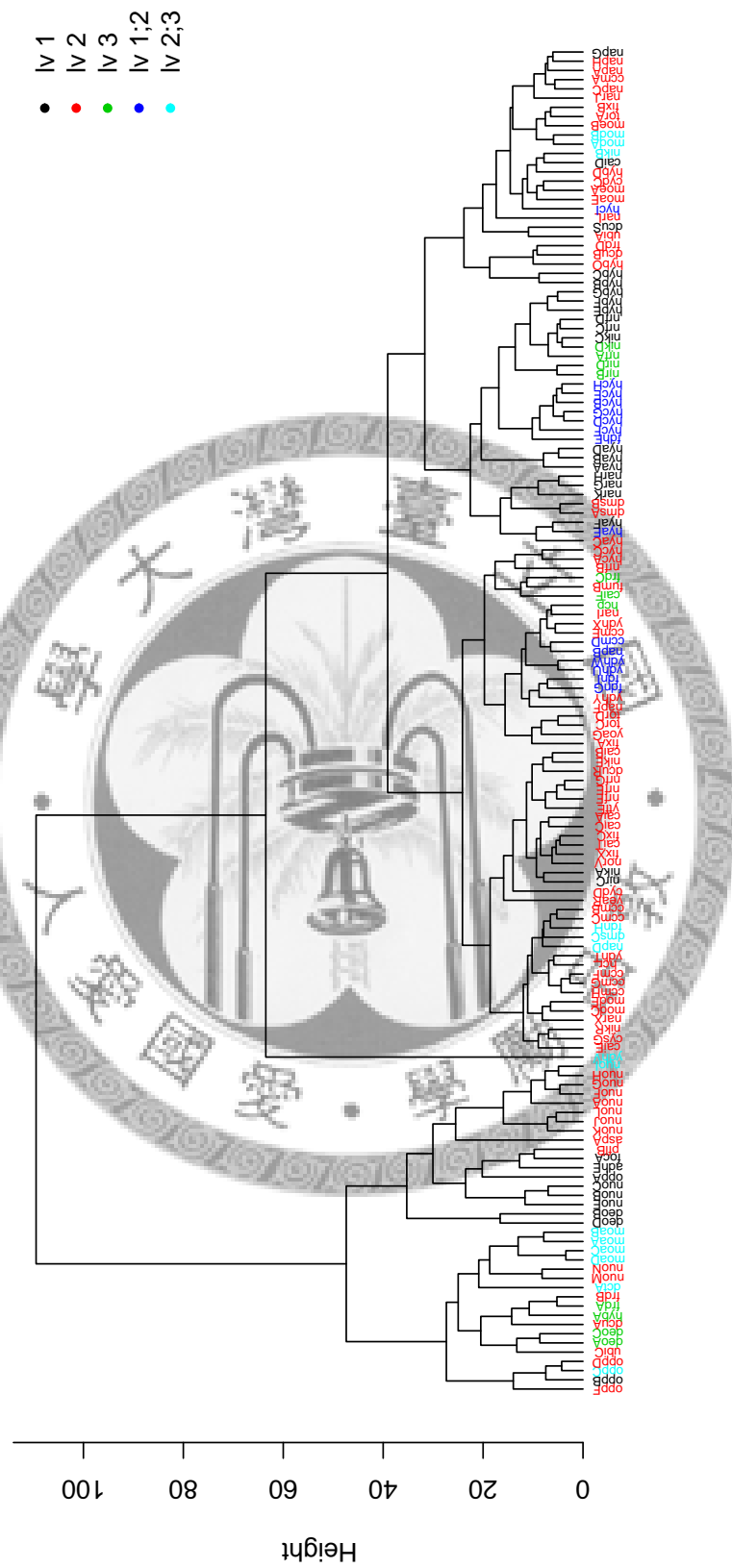


圖 3.6: 群集分析分群樹狀圖 (Non Mo dataset)

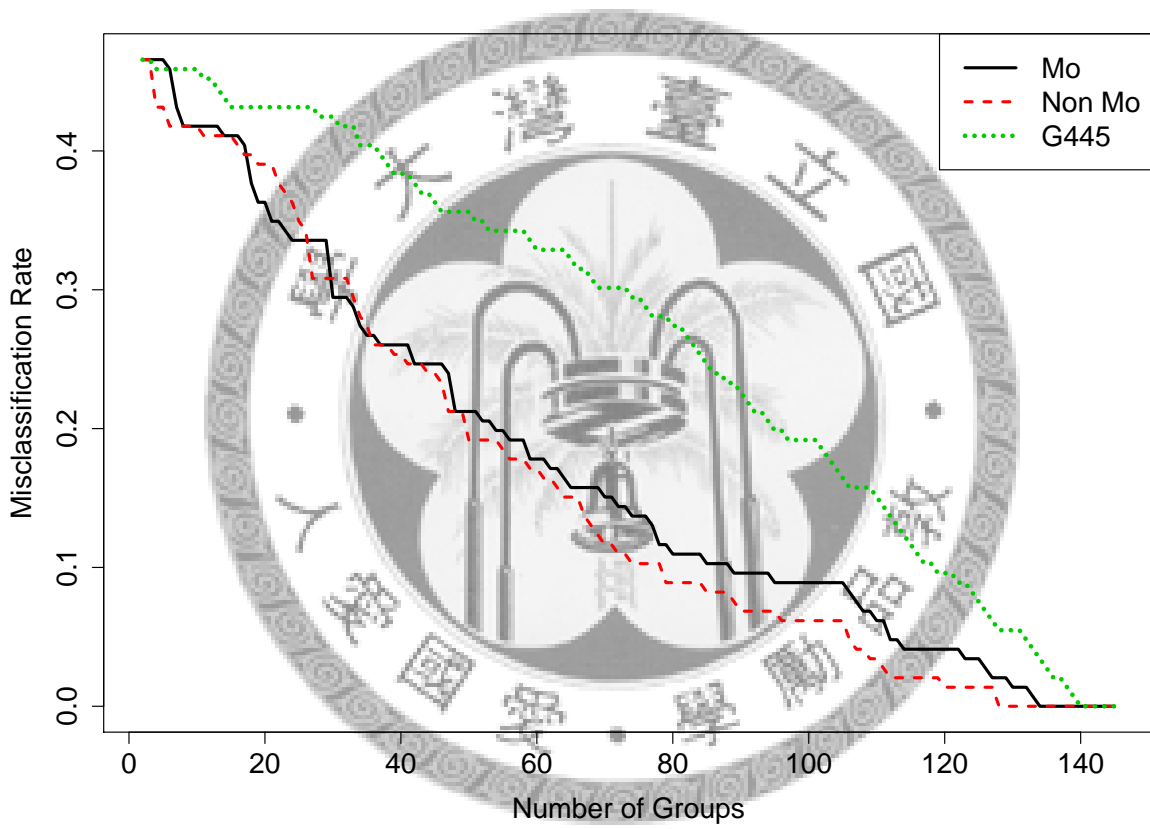


圖 3.8: 不同分群數之錯分率曲線圖

第四章 結論

從合併排序折線圖或是 Venn Diagram 我們都可以發現在不同的資料下，不同層級的基因調控關係，在三種統計方法下，結果都不盡相同。在合併排序折線圖中與我們一開始預期的 cidx 與 cidy 呈現的效果是類似的，在Mo與Non Mo的資料下，通常都是比cor方法較快找到名次低的基因，而G445的資料中，則是cor方法較好。特別在調控層級3(lv3)的情況下，三種資料中cor方法都有明顯的與cidx、cidy的趨勢不同，且較佳。而在Venn Diagram中，我們則可以發現，與我們預期相同的是，cidx與cidy重疊的基因較多，但卻不是最多的，最多是三種方法的交集區域，這可能表示這些基因用三種方法都可以找到，而較對比較少的是cor部分，cor方法找到的基因普遍較少一些。因此，如果要找到該層級最多的基因，可以利用三種方法的交集區去挑選出該層級大部分的基因，再配合cidx與cidy交集區，可以包含更多該層級基因。

另外看到集群分析圖，與我們預期中的有些許出入，相同的是，調控層級1、2集群很明顯，而其它調控層級則不明顯，在G445的資料中連層級1、2都被打散成很多小集群。因此可以推測，在包含各式各樣的實驗中，集群分析較不容易看出集群，調控層級3與多重調控層級的基因集群效果則更不明顯。因此我們建議利用集群分析的資料最好使用同質性較高的實驗資料。

由本篇研究分析方法中，我們尚無法將基因微陣資料百分之百的對應到真實的調控基因中，但部分的基因還是顯示出一些資料上面的特性，至於無法對應的基因調控關係，我們能假設其相關性並非如此的單純與直接，以至於只考慮二點基因之間的相關性，尚不足以推測其真實的調控關係，再加上於由Microarry資料雖然已經很普及，其資料的雜訊依然很多，且資料是一種相對的比較量，對於表現量較低的基

因，沒辦法有很效率的探測出來，因此我們希望可以找到一種雜訊較少的資料，以增加分析上的準確度。目前較新式的基因工具已逐漸發展，預期可以產生出一種絕對量化的基因資料，降低雜訊的干擾，以利於統計分析。希望未來的研究可以朝基因群與基因群之間的關係研究，而不單單只是單獨的看一個一個的基因。



參考文獻

- Akutsu T., Miyano S. and Kuhara S. (2000). Algorithms for Identifying Boolean Networks and Related Biological Networks Based on Matrix Multiplication and Fingerprint Function. *Journal of Computational Biology*, 7(3-4): 331-343.
- Anderson L. A., McNairn E., Lubke T., Pau R. N. and Boxer D.H. (2000) ModE-dependent molybdate regulation of the molybdenum cofactor operon moa in *Escherichia coli*. *Journal of Bacteriology*, 182(24):7035-43.
- Butte A. J. and Kohane I.S.(2002).Mutual information relevance networks:Functional genomic clustering using pairwise entropy measurements.*Pac Symp Biocomput*, 418-429.
- Davidson E. H. and Erwin D. H. (2006).Gene Regulatory Networks and the Evolution of Animal Body Plans.*Science*,Vol. 311. no. 5762, pp. 796 - 800.
- Faith J.J., Hayete B., Thaden J.T., Mogno I., Wierzbowski J., Cottarel G., Kasif S., Collins J. J. and Gardner T. S. (2007). Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLOS Biology*, 5(1):54-66.
- Friedman N., Linial M., Nachman I. and Pe'er D. (2000). Using Bayesian networks to analyze expression data. *J Comput Biol* , 7:601-620.
- Golub T. R., Antipova A. A. and Tamayo P. (2002). A strategy for oligonucleotide microarray probe reduction. *Genome Biology*, 3(12):re-

search0073.

- Hsing T., Liu L.-Y. D., Brun M., Dougherty E. R. (2005). The coefficient of intrinsic dependence (feature selection using el CID), *Pattern Recognition.*, 623-636.
- Liao J. C., Boscolo R., Yang Y. L., Tran L. M., Sabatti C. and Roychowdhury V. P. (2003). Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*,100(26):15522-15527.
- Liu, L.-Y. D., Chen, C.-Y., Chen, M.-J. M., Tsai, M.-S., Lee, C.-H. S., Phang, T. L., Chang, L.-Y., Kuo, W.-H., Hwa, H.-L., Lien, H.-C., Jung, S.-M., Lin, Y.-S., Chang, K.-J. and Hsieh, F.-J. (2009). Statistical identification of gene association by CID in application of constructing ER regulatory network. *BMC Bioinformatics*,10:85.
- Mordelet F. and Vert J. P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24(16):76-82.
- Salgado H., Santos-Zavaleta A., Gama-Castro S., Peralta-Gil M., Penaloza-Spinola M.I., Martinez-Antonio A., Karp P.D., and Collado-Vides J., (2006). The comprehensive updated regulatory network of Escherichia coli K-12. *BMC Bioinformatics*, 7:5 2006.
- Schena M., Shalon D., Davis R. W. and Brown P. O. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, 270(5235):467-470.
- Shen-orr, S. S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genet.*, 31, 64-68.
- Singer R. H. and Penman S. (1973). Messenger RNA in HeLa cells: kinetics of formation and decay. *J Mol Biol*, 78(2):321-334.
- Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. and Futcher B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccha-


romyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, Vol. 9, Issue 12, 3273-3297

Zare H., Sangurdekar D., Srivastava P., Kaveh M. and Khodursky A. (2009) Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations. *BMC Systems Biology*, 3:39.



附錄一

ModE調控基因群(一)



Lv. 1 Genes		
No.	Genes	Pathway
1	deoD	deoD
2	hycA	hycA
3	hycB	hycB
4	hycC	hycC
5	hycD	hycD
6	hycE	hycE
7	hycF	hycF
8	hycG	hycG
9	hycH	hycH
10	narL	narL
11	narX	narX
12	modC	modC
13	moaA	moaA
14	moaB	moaB
15	moaD	moaD
16	moaE	moaE
17	moaC	moaC
18	modA	modA
19	deoA	deoA
20	deoB	deoB
21	deoC	deoC
22	oppA	oppA
23	oppB	oppB
24	oppC	oppC
25	oppD	oppD
26	oppF	oppF
27	hycI	hycI
28	modB	modB

附錄一.

ModE調控基因群(二)

Lv. 2 Genes 1-26			Lv. 2 Genes 27-52			Lv. 2 Genes 53-78		
No.	Gene	pathway	No.	Gene	pathway	No.	Gene	pathway
1	hcr	narL,hcr	27	nuoN	narL,nuoN	53	dcuS	narL,dcuS
2	dcuA	narL,dcuA	28	narG	narL,narG	54	hybA	narL,hybA
3	focA	narL,focA	29	narH	narL,narH	55	norW	narL,ygbD
4	pflB	narL,pflB	30	narI	narL,narI	56	norV	narL,norV
5	caiF	narL,caiF	31	narJ	narL,narJ	57	aspA	narL,aspA
6	adhE	narL,adhE	32	nirB	narL,nirB	58	ubiC	narL,ubiC
7	fdhF	narL,fdhF	33	nirD	narL,nirD	59	cysG	narL,sysG
8	hyaA	narL,hyaA	34	fdnG	narL,fdnG	60	moeA	narL,moeA
9	hyaB	narL,hyaB	35	fdnH	narL,fdnH	61	moeB	narL,moeB
10	hyaC	narL,hyaC	36	fdnI	narL,fdnI	62	narK	narL,narK
11	hyaD	narL,hyaD	37	nrfA	narL,nrfA	63	nirC	narL,nirC
12	hyaE	narL,hyaE	38	hybB	narL,hybB	64	hcp	narL,hcp
13	hyaF	narL,hyaF	39	hybC	narL,hybC	65	nikR	narL,nikR
14	ubiA	narL,ubiA	40	hybD	narL,hybD	66	hybE	narL,hybE
15	nuoM	narL,nuoM	41	hybF	narL,hybF	67	torD	narL,torD
16	nuoF	narL,nuoN	42	hybG	narL,hybG	68	ytfE	narL,ytfE
17	nuoA	narL,nuoA	43	torA	narL,torA	69	cydC	narL,cydC
18	nuoB	narL,nuoB	44	torC	narL,torC	70	cydD	narL,cydD
19	nuoC	narL,nuoC	45	nrfB	narL,nrfB	71	ydhU	narL,ydhU
20	nuoE	narL,nuoE	46	nrfC	narL,nrfC	72	ydhX	narL,ydhX
21	nuoG	narL,nuoG	47	nrfD	narL,nrfD	73	ydhV	narL,ydhV
22	nuoH	narL,nuoH	48	nrfE	narL,nrfE	74	ydhY	narL,ydhY
23	nuoI	narL,nuoI	49	nrfF	narL,nrfF	75	yoaG	narL,yoaG
24	nuoJ	narL,nuoJ	50	nrfG	narL,nrfG	76	ydhW	narL,ydhW
25	nuoK	narL,nuoK	51	hybO	narL,hybO	77	ydhT	narL,ydhT
26	nuoL	narL,nuoL	52	dcuR	narL,dcuR	78	yeaR	narL,yoaG

附錄一.

ModE調控基因群(三)

Lv. 3 Genes		
No.	Gene	pathway
1	caiT	narL,caiF,caiT
2	detA	narL,dcuR,dctA
3	caiC	narL,caiF,caiC
4	caiE	narL,caiF,caiE
5	fixA	narL,caiF,fixA
6	fixB	narL,caiF,fixB
7	caiA	narL,caiF,caiT
8	caiD	narL,caiF,caiD
9	caiB	narL,caiF,caiB
10	fixC	narL,caiF,fixA
11	fixX	narL,caiF,fixX



附錄一.

ModE調控基因群(四)

Lv. 2 & 1 Genes		
No.	Gene	pathway
1	dmsA	dmsA;narL;dmsA
2	dmsB	dmsB;narL;dmsB
3	dmsC	dmsC;narL;dmsC
4	napC	narL,napC;napC
5	napB	narL,napC;napC
6	napH	narL,napC;napC
7	napG	narL,napC;napC
8	napA	narL,napC;napC
9	napF	narL,napC;napC
10	ccmA	narL,ccmA;ccmA
11	ccmH	narL,ccmH;ccmH
12	ccmG	narL,ccmG;ccmG
13	ccmF	narL,ccmF;ccmF
14	ccmE	narL,ccmE;ccmE
15	ccmC	narL,ccmC;ccmC
16	ccmB	narL,ccmB;ccmB
17	ccmD	narL,ccmD;ccmD
18	napD	narL,napD;napD

附錄一.

ModE調控基因群(五)

Lv. 3 & 2 Genes		
No.	Gene	pathway
1	dcuB	narL,dcuR,dcuB,marL,dcuB
2	fumB	narL,fumB;narL,dcuR,fumB
3	frdA	narL,frdA;narL,dcuR,frdA
4	frdB	narL,frdB;narL,dcuR,frdB
5	frdC	narL,frdC;narL,dcuR,frdC
6	frdD	narL,frdD;narL,dcuR,frdD
7	nikD	narL,nikR,nikD;narL,nikD
8	nikE	narL,nikR,nikE;narL,nikE
9	nikB	narL,nikR,nikB;narL,nikB
10	nikC	narL,nikR,nikC;narL,nikC
11	nikA	narL,nikR,nikA;narL,nikA

