以間接關聯規則探勘基因表現微陣列資料

Mining Gene Expression Data with Indirect Association Rules

曾新穆(Vincent S. Tseng)
國立成功大學
資訊工程學系
tsengsm@mail.ncku.edu.tw

劉又誠(Yu-Cheng Liu)
國立成功大學
資訊工程學系
uchenliu@idb.csie.ncku.edu.tw

辛致煒(J. W. Shin)
國立成功大學
醫學院微免所寄生蟲學科
hippo@mail.ncku.edu.tw

摘要

資料探勘(Data Mining)為近幾年來應用在微陣列分析(Microarray Analysis) 上十分熱門的研究技術，其目的在從大量的基因表現(Gene Expression)資料中，萃取出有用的知識，以提供研究生物研究學者，在進行研究時當作參考。本研究中，我們應用資料探勘中的間接關聯規則，套用到基因的微陣列分析，並且以<X, Y | M>來表示 X 與 Y 透過 M 形成間接關聯規則，代表 X 與 M 可能為某一生物反應下的參與基因，而 Y 與 M 可能為另一生物反應下的參與基因，其中代表 M 為兩種生物反應的必要因素，有助於在不同生物反應下找尋共同的關聯性研究。最後並以 Gene Ontology 來驗證其關聯性的正確性。經由實驗分析證實，我們提出的方法架構確實可以找到不同於傳統關聯規則，能夠提供生物學家於基因關聯性研究中，更多不同的參考。

關鍵詞：Data Mining, Microarray, Gene Expression Analysis, Indirect Association Rule

一、緒論

隨著資訊電子化的來臨，資料的蒐集愈來愈容易，相對之下，處理資料的能力也更顯重要；對於如何在大量資料中擷取所需要的資訊，就變成是一個相當重要的課題，因此 Knowledge Discovery in Database (KDD) 的議題也因此而產生。

在生物資訊領域中，由於生物學家善加利用新的科技來加速研究的步調，近年來發展迅速，舉凡多重基因比對(multiple gene alignments)[5]、基圖的辨識(motif identify) [4]、微陣列分析(Microarray analysis)[8]、蛋白質結構預測(protein structure prediction)[14]跟生物反應路徑(pathway)[13]中，電腦計算功能都扮演了相當重要的角色；而在眾多的議題中，微陣列相關研究便是其中一個相當重要的主題。過去生物學家使用傳統的方式，一次只能夠檢視幾十個基因，但藉由微陣列的幫助，則可同時篩選大量(上千)的基因表現值以供生物學家參考；也由於微陣列技術的發展，大量的基因表現之分析變得更為困難與重要，因為要如何面對這樣龐大的資料，同時在其中找出生物學家感興趣的資訊，用人工的分析方式已經無法滿足研究人員的需求。

對於龐大的基因資料，專家們一開始使用統計學的方式來做分析，不過由於某些限制，使得這些工作往往相當的花費時間，因此有人開始利用資料探勘(Data Mining)的技術來輔助探索其中重要的資訊；其中常被使用的技術有分群(Clustering)[10] [12]、關聯規則(Association Rule)[7][8]跟分類(Classification)[3]，利用這些技術，生物學家可以在大量資料中挖掘出他們比較感到興趣的資訊，以加速研究的腳步；然而對於這些方式有一個共同的問題，因為微陣列通常是一個數字型態的基因表現值，對於這樣的資料型態如何進行適當且

正確的處理，也是一件相當值得研究的事情。

過去生物學家很常使用分群(clustering)[10][12]的技術應用於微陣列資訊的分析，但是因為分群技術並不適用於辨識基因之間的關係，所以有些生物學家開始使用資料探勘的另外一項技術—關聯規則(Association Rule)[1][2]。在使用關聯規則於分析基因表現資料(gene expression data)的狀況下，每個關聯規則項目可將基因敘述為強烈表現(expressed)或者抑制(repressed)，用於敘述在細胞環境(cellular environment)下相關聯的表現基因。例如，{cancer}⇒{gene A↑, gene B↓, gene C↑}，代表在資料中挖掘(mined)出此一規則，在檢驗出罹癌細胞中，A 基因為高度表現(highly expressed)，B 基因為高度抑制(highly repressed)，而 C 基因為高度表現，此三個基因現象往往共同表現。

然而，在於現實狀況中亦存在另一種相當重要的關係。例如，gene set A 與 gene set B 的表現關聯性很高，gene set A 與 gene set C 的表現關聯性很高，但 gene set B 與 gene set C 的表現關聯性卻很低，代表 gene set A 與 gene set B 可能為某一細胞環境下的生物反應中的參與基因，而 gene set A 與 gene set C 可能為另一細胞環境下的生物反應中的參與基因，代表 gene set A 為兩種不同生物反應中的必要因素。若此兩種不同生物反應，代表是兩種不同癌症的生物反應，那麼 gene set A 可能就是不同癌症中的相同的因素基因。對於不同疾病之間的關聯性研究，將會有極大的重要性。本研究主要是應用間接關聯規則(Indirect Association Rule)[19][20]來挖掘出此一現象的基因，並以 Gene Ontology [22][23]來驗證其關聯性的正確性。

本研究的章節安排如下。下一節將對本研究相關的文獻進行探討。第三節將說明本研究所提出的方法架構。第四節為實驗結果。最後並在第五節為本研究做結論。

二、文獻探討

2.1 生物資訊學上的相關研究

過去 R. Chen 等學者曾經將組織(tissue) 中的基因相關資訊用於尋找轉錄因子(transcription factors)跟基因表現值之間的關係[7]，他們將每一個組織的跟一群轉錄因子作結合，再把轉錄因子對應到特定的基因表現值上面，並且將每一個組織的反應看成一筆交易(transaction)資料，經由轉換後的資料使用關聯規則來找出轉錄因子跟目標基因表現值之間的關聯性。Creighton[8]等人利用關聯規則於其研究，整個流程可以分成兩個步驟，首先將基因表現值資料分類，再對分類完畢的實驗使用關聯規則；分類的部分將基因表現值分為已超過 0.2 表示成表現狀態(up-regulated)跟小於-0.2 表示為抑制狀態(down- regulated)，然後再使用關聯規則找出它們之間的關聯性；找出來的規則如：

$$G_a\uparrow \rightarrow G_b\downarrow$$

該規則顯示當基因 G_a 為表現狀態的時候，G_b 基因將會被抑制。

而 Kotala 學者等人[11]，利用 Peano Count Tree (P-Tree)，套用編碼的方式，找出類似“$\{G_1,...,G_n\}\rightarrow G_m$” 的規則，表示一群基因$\{G_1,...,G_n\}$跟 G_m 之間的表現有一定程度的關聯性。

其他的研究，例如使用分群方式[9][17]區分基因之間屬於哪一個群組，或者分類[3]的方式，去區分某個基因可能是屬於某個家族，皆為之前相關的研究。

2.2 關聯規則

關聯規則是於 1993 年，首先由 IBM

研究員 Agrawal 所提出的演算法[1][2]，原本是用於分析龐大的資料庫裡面的交易資料，試圖發覺其中有隱含的模式(pattern)，並且進一步找到相關的規則。這個方法，一開始被應用於商業銷售的資料分析，來調整商業的銷售的策略。例如，使用者在買個人電腦，同時會有 70%的機會同時購買印表機，這樣個人電腦的廠商跟印表機的廠商就可以應用一些搭售的促銷策略來增加銷售量跟市場佔有率。

在由 Agrawal[1][2]跟他的團隊提出的 Apiroi 演算法。主要觀念是我們擁有一個資料庫 D 跟項目集合 (Itemset) $I=\{I_1,...,I_n\}$，項目集合代表每個可能出現在資料庫的項目，一個資料庫 D 由一群交易 $T=\{T_1,...,T_m\}$組成，每個 T_i 都是項目集合的子集合。經由 Apriori 演算法找出的關聯規則，其表示法為 $A\rightarrow B$，其中 $A\subseteq I$，$B\subseteq I$ 且 $A\cap B=\phi$，其中 A 稱為規則的 LHS (Left Hand Side)，B 稱為規則的 RHS (Right Hand Side)；而在關聯規則中有兩個評估關聯規則的指標，用以表示 A 跟 B 之間的關係強度，分別是支持度(support)跟信心度(confidence)；支持度的定義是 Support(AB) = P(AB) = |AB|/m，表示 AB 集合在整個資料庫中出現的比例，信心度的定義則是 Confidence(AB) = P(B|A) = |AB|/|A|，表示在出現 A 集合的前提下出現 AB 集合的機率，表示整個規則的強度。

利用上面兩個評估準則，Apriori 演算法定義了兩個門檻(threshold)，分別是最小支持度(minimum support)跟最小信心度(minimum confidence)；所有找出來的關聯規則必須同時大於最小支持度跟最小信心度。

Aprior 演算法，利用了一個數學原則，改善原本以暴力法來組合所有可能集合的效率不彰的問題。此一重要的規則是：一個大項目集合(Large Itemset)的子集合亦必然是大項目集合。因此 Agrawal 等人利用該規則反覆進行產生準大項目集合(Candidate Large Itemset)之產生跟裁減(pruning)行為，裁減準大項目集合成為大項目集合，以減少不必要的組合；整個演算法就是重複產生大項目集合跟準大項目集合，再做裁減行為的步驟。演算法如下：

Algorithm Apriori

1. $L_1=\{\text{Large-1-itemsets}\}$
2. for($k=2;L_{k-1}\neq\phi;k++$) do begin
3. $\quad C_k=\text{apriori-gen}(L_{k-1})$
4. $\quad$ for all transactions $t\in D$ do begin
5. $\quad\quad C_t=\text{subset}(C_k,t)$
6. $\quad\quad$ forall candidates $c\in C_t$ do
7. $\quad\quad$ c.count++
8. $\quad$ end
9. $\quad L_k=\{c\in C_k \mid c.\text{count}\geq\text{minsup}\}$
10. end
11. Answer=$\cup_k L_k$

Apriori 演算法大致上可以分成下的步驟：

1. 從頭到尾掃描資料庫 D 一次，選出支持度大於最小支持度的項目，我們稱之為大項目1 的項目集合，其中1表示此項目集合的長度。掃描結果為所有長度為1大項目集合的集合，我們稱之為 L_1。
2. 根據大項目集合 L_{k-1} 使用priori-gen函式產生準大項目集合C_k。
3. 利用subset 函式讀取資料庫 D 以計算 C_k 裡各準項目集合的支持度。
4. 挑選 C_k 裡面支持度大於最小支持度的項目集合，成為長度 k 的大項目集合 L_k。
5. k 加1，然後再重複2 到4 的步驟，直到無法找出準大項目集合為止。

最後，Apriori 會產生出各種長度的大項目集合，支持度皆大於使用者所定義的最小支持度。接下來，便是用最小信心度來探勘關聯規則，對每一個大項目集合

L，找出所有可能的子集合 A，則規則可以表示成：

$$A \rightarrow (L - A)$$

其信心度為 P（L|A），即為在 A 出現的情況下，L 出現的機率，亦即將 L 的支持度除以 A 的支持度。然後挑選出所有信心度大於最小信心度的規則，就是最後的關聯規則探勘結果。

Apriori 裡有兩個重要的函式 apriori-gen 和 subset，這兩個函式是各 Apriorilike 演算法都會利用到的程序，分別用於產生可能的準大項目集合跟計算每個項目集合的支持度，包括本論文都有延伸利用到，因此我們將介紹這兩個函式：

apriori-gen

1. insert into Ck
2. select p.item1, p.item2, …, p.itemk-1, q.itemk-1
3. from Lk-1 p, Lk-1 q
4. where p.item1=q.item1, …, p.itemk-2=q.itemk-2, p.itemk-1<q.itemk-1

此函式的主要功能是用來產生可能的大項目集合，它利用 L_{k-1} 來產生長度為 k 的項目集合 C_k，也就是候選項目集合。

Subset

1. Subset_collection=ϕ
2. For each c in C_k
3. If c is subset of t
4. Subset_collection= Subset_collection∪c
5. Return Subset_collection

Subset 以 apriori-gen 產生的準大項目集合，必須再以 Subset 函式計算其內各項目集合的支持度，才能產生出大項目集合。Subset 會取用兩個參數—準大項目集合 C_k 以及一筆交易紀錄 t。然後 Subset 會找出在 t 裡面，含有 C_k 裡的項目集合的集合 C_t。於是我們就可以將 C_t 裡的項目集合的支持度計數加 1，表示 C_t 的項目集合出現次數增加一次。因此，每一次 C_k 產生之後，Apriori 必須掃描一次資料庫，將每筆交易紀錄 t 送入 Subset 計算 C_k 的支持度。

2.3 負關聯規則

負關聯規則(Negative association rules) [6] 為 Brin 等人在延伸關聯規則的架構下首先提出的觀點，而後續亦有相關的研究發展出挖掘負關聯規則的方法[18, 22, 26]。而所謂的負關聯規則是用來找尋出共同發生頻率極低的項目集合。若是負關聯規則存在項目集合 X 與 Y 之間，則可表示成 $X \rightarrow \overline{Y}$ 或 $Y \rightarrow \overline{X}$，代表 X 與 Y 極少共同發生在同一筆交易中。

在[16]的研究中提出，若用直覺性的方法(naive approach)來從包含大量資料的資料庫中挖掘負關聯規則，則往往會挖掘出極大量使用者不感興趣的規則。故在此篇先前的研究中，該作者應用先前的關聯規則與專業的背景知識(domain knowledge)來限制挖掘的方向，以得到較少但使用者真正感興趣的負關聯規則。

而在[21]中提出了一個架構能夠同時挖掘出關聯規則與使用者較感興趣的負關聯規則，而且對於所處理的資料不需要專業的背景知識，且能更具體的表達出實際上在不同項目集合之間的關係。雖然在此架構下能夠同時挖掘出關聯規則與使用者感興趣的負關聯規則，但在空間使用的需求上仍然非常巨大。另一個問題是，所找出來的負關聯規則在實際應用上仍然過多、不便於使用。

2.4 間接關聯規則

間接關聯規則與負關聯規則有相當密切的關係，兩者皆用來挖掘出項目集合間

沒有足夠高支持度的關係。而間接關聯規則提供了一個更有效率的方法來挖掘使用者感興趣的負關聯規則，不需要使用負項目(negative items)或專業的背景知識，只需從“被期待為高頻率出現”的項目集合當中，探索出“非高頻率項目對”(infrequent itempairs)即可。

一對項目對 {x, y} 若是透過一個中介 (mediator) M 形成間接關連規則，則必需滿足以下條件：

1. Support({x, y})<t_s
   (Itempair Support Condition)

2. 存在一個非空集合 M：
   (a) Support({x}∪M)≥t_f,
   Support({y}∪M)≥t_f
   (Mediator Support Condition)
   (b) Dependence({x}, M)≥t_d,
   Dependence({y}, M)≥t_d
   (Mediator Dependence Condition)

門檻(threshold) t_s 為項目對支持門檻(itempair support threshold)，t_f 為中介支持門檻(mediator support threshold)，而 t_d 為中介信賴門檻(mediator dependence threshold)。在實際應用當中，$t_f \geq t_s$。

在此篇研究當中以 <X, Y | M> 來表示 X 與 Y 透過 M 形成間接關聯規則。而用 IS 測量(IS measure) [18]來衡量信賴度(dependence)。在條件 2(b)中，集合 X 與 Y 分別與集合 M 用 IS 測量計算其信賴度。

2.5 Gene Ontology

隨著後基因體(post-genome)時代的來臨，以及愈來愈多的基因體資料產生出來，生物學家們急需要一個工具，可以很有系統性的去查詢、整合基因的資料、產物及基因所具有的功能，在這種需求下，Gene Ontology [22][23]就因此被創造出來，至今已經整合非常多包括植物和動物的基因資料庫，而它又可以分為三個大類，包含分子功能(MF: Molecular Function)、生物反應過程(BP: Biological Process)以及細胞位置(CC: Cellular Component)。

MF 的用途為以分子的角度，描述基因的產物的活動，例如催化(catalytic activity)或轉錄調控(transcription regulator activity)等，BP 是由一個或是多 MF 來完成，有時很難去分辨 MF 和 BP 的不同，不過 BP 所描述的動作必定是包含一個以上的步驟，不像 MF 只描述單一的生化活動，而 BP 不等於 pathway，它沒有描述 pathway 裡複雜的關係，CC 的用以表示細胞的某個部份或位置，描述基因在細胞的那個位置發揮它的功能。這些分支都是由 GO term 所構成，而在 GO 上每一個節點皆代表一 GO term，而這些 GO term 是有結構性的，在愈高層的 term 所代表的意義愈廣泛，而愈低層 term 的意義愈狹隘，而生物學家可以用利用 GO 所提供的功能來查詢不同階層的 term。如圖 1 所示，我們可以去尋找在 BP 下所有的基因功能，或是單獨只查詢功能為 biological regulation 的基因也可以。

GO term 是以 Direct Acyclic Graph (DAG) 的架構所組合而成，而在 GO 中是用“is-a”和“part-of”的兩種關係來連接 term 和 term 之間的關係。如圖 2 所示，例如：cellular process “is-a” child of biological process，或 regulation cellular process is “part of” cellular process，而其與樹不一樣的地方為一個 term 節點可以有許多的父節點，而在 GO 中大約有 98% 的“is-a”關係和 2%的“part-of”的關係。

三、研究方法

3.1 應用間接關聯規則於基因表答資料

以購物籃分析(market basket analysis)為背景，一筆的基因表答資料可視同為一筆交易資料，而每一個表現值可視為是一個項目。然而在購物籃分析中，任何一個項目在任一筆交易資料中，就僅是被購買或沒被購買兩者其中之一。然而在基因表答資料中，每一個基因皆被付予一個數值，代表此基因在相對應的條件下的表現值。所以在應用間接關聯規則於基因表答資料中，首先就必需先將每個基因的表現值對映到 up (高度表現), down (高度抑制), 或者 normal (兩者皆非)。但在利用基因表現資料來討論基因之間的交互作用，往往只聚焦在討論高度表現與高度抑制的部份，因為只有這兩者在基因之間交互作用中有真正的參與作用。所以在基因表答資料當中，任何一個基因都可以對映成兩個項目(up 與 down)於交易型態資料中，如圖 3 所示。

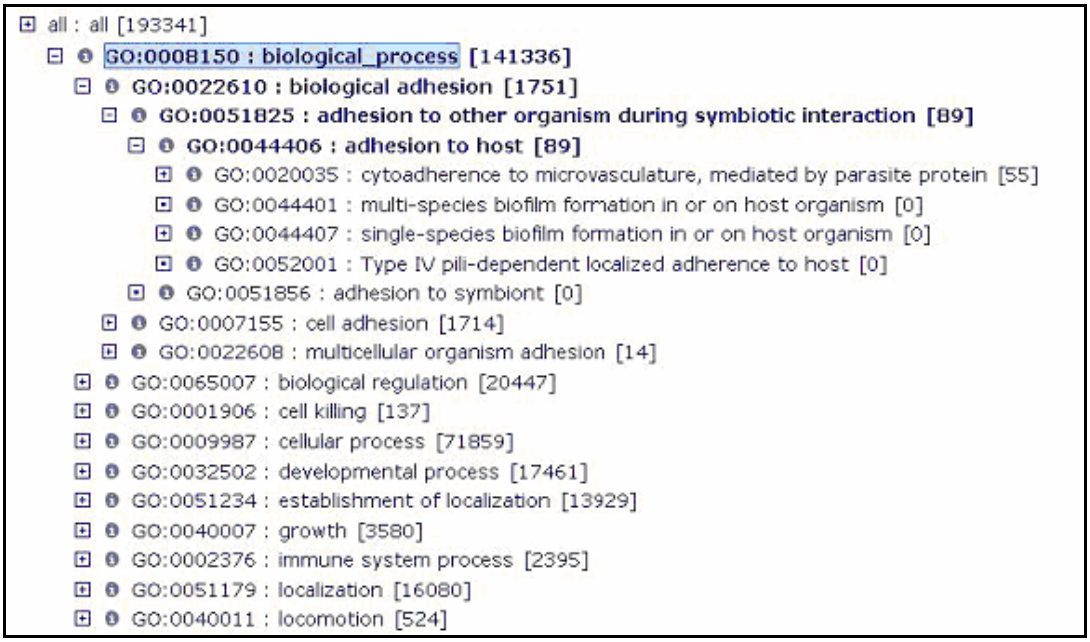

圖 1. Gene Ontology 的架構

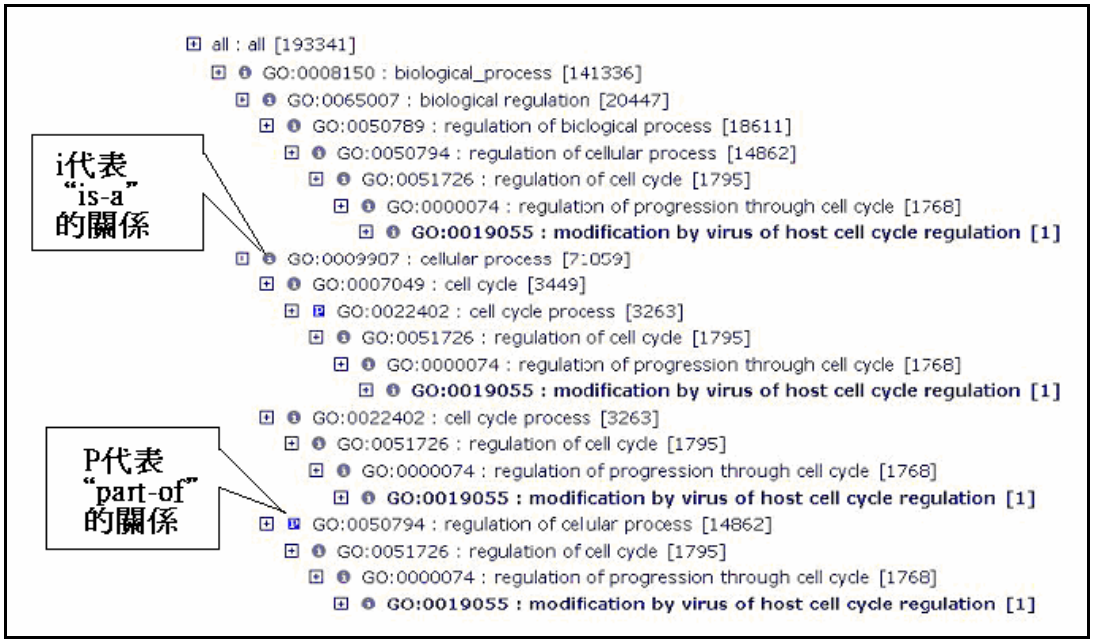

圖 2. Gene Ontology “is-a” 和 “part-of” 的示意圖

將基因分為 up 或 down 的項目後，在某一實驗條件下(可視為購物籃分析中的一筆交易資料)，基因表現可敘述相關聯的基因在細胞環境下的交互作用。例如，如果用基因表現資料來做疾病的觀察，間接關聯規則 ＜gene B↓, gene C↑ | gene A↑＞ 代表，如果 gene A 的表現為 up 而 gene B 的表現為 down，則可能為某一疾病的基因表現，如果 gene A 的表現為 up 而 gene C 的表現為 up，則可能為另一疾病的基因表現，並且 gene A 可能為此兩疾病的致病因素或參與了此兩疾病的作用，可作為醫生診療的參考，或相關基因作用的研究依據。

	Gene 01	Gene 02	Gene 03	Gene 04	Gene 05
Condition 1	-0.212	0.045	-0.023	-0.014	0.336
Condition 2	0.002	-0.077	0.224	0.016	0.07
Condition 3	0.034	-0.256	0.016	-0.081	-0.235
Condition 4	0.085	0.054	-0.033	0.223	-0.048
Condition 5	0.063	-0.033	-0.06	0.016	-0.067

↕

	Gene 01	Gene 02	Gene 03	Gene 04	Gene 05
Condition 1	Down				Up
Condition 2			Up		
Condition 3		Down			Down
Condition 4				Up	
Condition 5					

圖 3. 基因表現資料對映成交易型態資料的示意圖

3.2 挖掘間接關聯規則於基因表答資料

表1 間接關連規則演算法

1. Extract the large itemsets, $L_1, L_2, \cdots L_n$, using standard mining algorithms.
2. $P = \emptyset$
3. for $k = 2$ to n do
4. $\quad C_{k+1} \leftarrow \text{join}(L_k, L_k)$
5. $\quad$ for each $(a, b, M) \in C_{k+1}$ do
6. $\quad\quad$ if $(sup(\{a, b\}) < t_s$ and $d(\{a\}, M) \geq t_d$ and $d(\{b\}, M) \geq t_d$
7. $\quad\quad\quad P = P \cup (a, b, M)$
8. $\quad$ end
9. end

在先前的研究[19]與[20]中，提出了從項目對中挖掘間接關連規則演算法如表1所示。此演算法可分為兩階段。於第一階段，利用 Apriori 演算法產生所有的大項目集合。第二階段，再利用大項目集合L_k產生k+1階段候選準間接關聯規則。若符合間接關聯規則的 Itempair Support Condition、Mediator Support Condition以及 Mediator Dependence Condition則可成為k+1階段間接關聯規則。

3.3 GO Term 之權重值計算

在第 2.5 節時，有去說明 GO 的整個架構，而 GO 是一圖形結構，在 GO 上的每一個節點皆是代表一 GO term，而兩個 term 之間是用一條邊(edge)來相連。所以有關計算在 GO 中 term 跟 term 之間計算相似度的方法，最直覺的方法就是去計算這兩個 term 在 GO 上的節點距離，或是再加上這兩個 term 在 GO 中的深度來做為相似度的計算。但是一般這樣子會皆把距離的單位都視為是相同的，也就是在 GO 上任意兩個 term 只要它們在 GO 上的節點距離是一樣的，則它們都會有相同的相似度。但是這樣子的做法，完全沒有去考慮到所經過節點的重要性，即使是把深度加進去考慮也沒有辦法解決這個問題。

所以為了要讓相似度的計算結果更加的準確，我們在進行相似度計算之前，會先對於每個不同的 GO Term 給予不同的權值(weight)，這是因為不同的 GO Term 其重要的程度也會不一樣。例如當我們在註解 “transmembrane receptor”時，要比註解 “receptor”有更準確的生物意義，因為 “transmembrane receptor”是在 “receptor” 下的更詳細的註解，在考慮了 GO Term 的權值後可以讓計算出來的相似度更加的準確。而在本論文中，是採用 Resink et al.,[15] 所提出的在 “is-a” 架構下，使用 Information Content 方法來計算相似度，將會比使用傳統上兩個節點間的距離來計

算相似度更加來的精確。

所以我們使用 Information Content 方法來代表每個 GO term 的權重值，而我們在這裡是去計算在 GO 裡面所有的 term 的權重值，使用 Information Content 的方法來計算權重值的時候，愈常來被註解的 GO term 其重要性和權重值愈低，而較不常用來被註解的 GO term 則其重要性和權重值愈高。當一個 GO term 被用來註解一個基因之後，這個 GO term 的註解個數會加一，並且在這個 GO term 上層的所有父節點的 GO terms 的頻率也會累加一。如圖 4、5 所示。

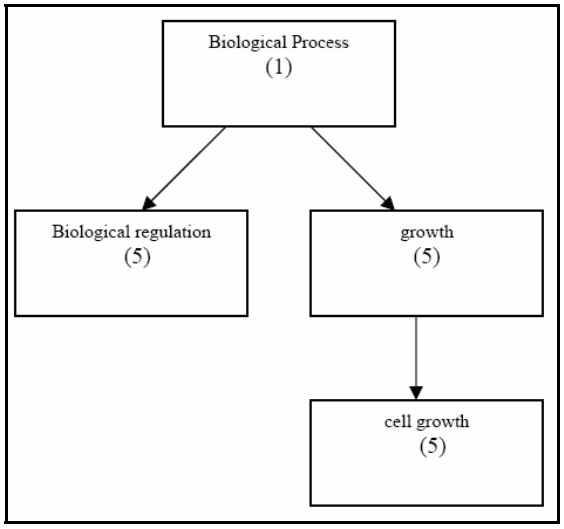

圖 4. 各 GO Term 用來被註解的次數

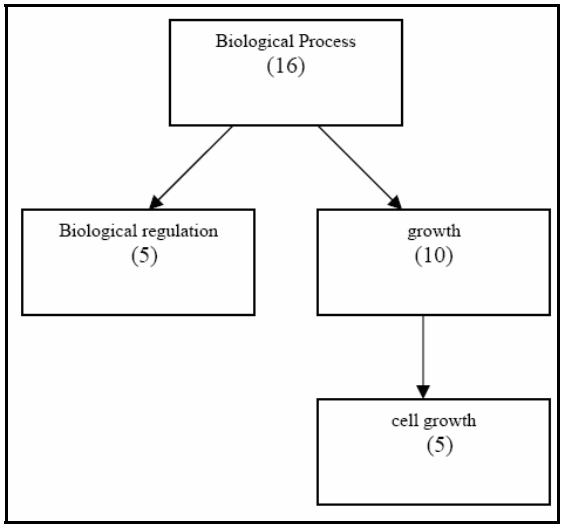

圖 5. 各 GO Term 累加後的次數

得到各 GO term 之註解個數之後，將註解個數轉換成機率，而每一個 term 的機率值，我們以 p(t)來表示，在此我們將 GO 裡的“Molecular Function”、“Biological Process”和“Cellular Component”分開處理，而計算每個 GO term 的機率值的分母為註解此 GO Term 所在的類別的總頻率。如 BP，而分子為註解 GO term 的頻率，例如假設“biological process”被用來註解的頻率為 16，若“cell growth”被用來註解的頻率為 5，則“cell growth”被用來註解的機率為 0.3125，圖 6 顯示圖 5 的例子計算後的結果。

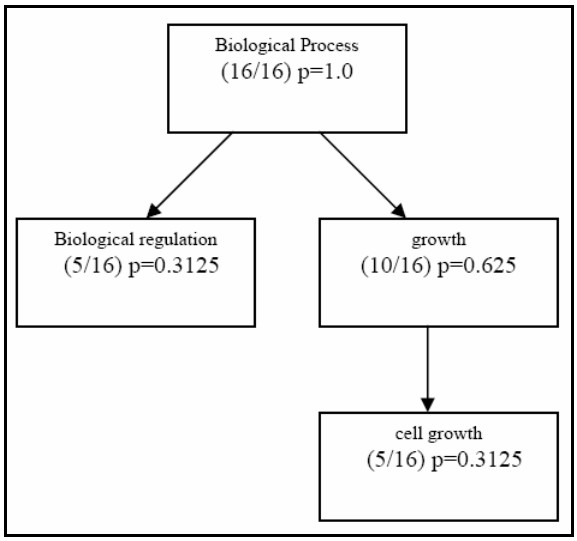

圖 6. 計算後各 GO Term 之機率

得到各 GO Term 之機率 p(t)將其代入公式(1)，經過轉換後即可到各 GO Term 之權重值。

$$w(t) = -\ln(p(t)) \qquad (1)$$

3.4 基因語意相似度計算

在本論文中，我們假設每一個基因至少都會有一個或一個以上的 term 來註解此基因，若要去計算不同基因之間的兩個 term 的在 GO 上的語意相似度時，我們使用 Information Content 的方法去做交叉比對，如圖 7 所示，假設 g_1 分別被 t_1、t_2

和 t_3 所註解，而 g_2 分別被 t_4、t_5 和 t_6 所註解，要計算 t_1 和 t_4 在 GO 上的語意相似度，則分別找出這兩個 term 的所有共同父節點(share parents)出來。如圖 8 所示，然後再去比較找出來的這些父節點的權值的大小，將具有最大權值的父節點來當作此兩個 term 的語意相似值。

$$sim(t_i, t_j) = \max(w_{parent\_of\_ij}(t)) \quad (2)$$

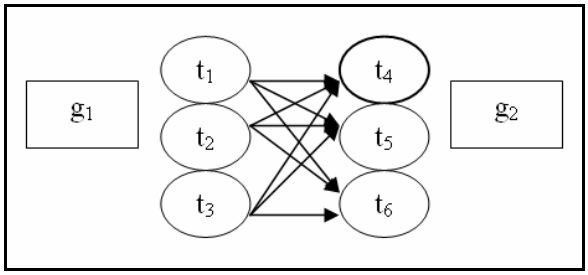

圖 7. 不同註解的 GO Term 交叉比對

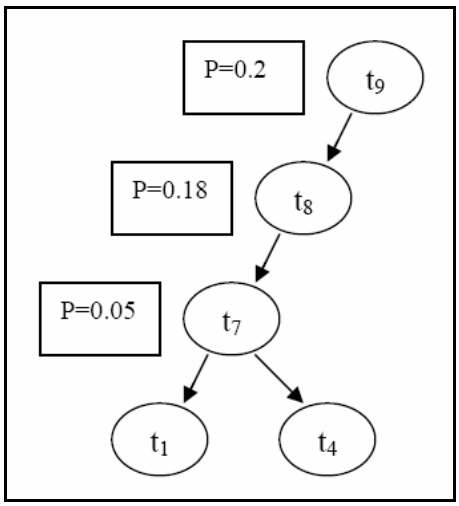

圖 8. t_1 與 t_2 之所有共同父節點的機率值

其中假設 g_1 與 g_2 分別被 $C_1=\{t_i\}$, i=0,...,m 和 $C_2=\{t_j\}$, j=0,...,n，這兩個 term sets 所註解，而 C_1 和 C_2 的 term 的個數分別為 m 和 n 個，而以圖 8 為例的話，t_1 和 t_4 的所有共同父節點為 t_7、t_8 和 t_9，所以其在 GO 上的語意相似度分別為:

$$sim(t_1, t_4) = \max(w(t_7), w(t_8), w(t_9))$$
$$= \max(-\ln(0.05), -\ln(0.18), -\ln(0.2))$$
$$= \max(\ln(500), \ln(18), \ln(2)) = \ln(500)$$

而其最大值為 ln(500)，所以 t_1 和 t_4 在 GO 上的語意相似度為 ln(500)。而接著在去計算 t_1 和 t_5、t_1 和 t_6...等兩兩交集的結果，若要計算 g_1 和 g_2 這兩個基因在 GO 上的語意相似強度，將分別註解 g_1 與 g_2 terms 經由兩兩配對計算出來的最大值來當作兩個基因在 GO 上的語意相似度。

四、實驗結果

這一章節，我們對於一些會影響我們演算法的變數做一些探討，另外對於所找出來的規則，會利用 Information Content 計算其在 GO 上的關聯性，是否與間接關聯規則的概念一致，做初步的驗證工作。

我們實驗的資料來自於和[8]相同的酵母菌的資料，主要是用了 300 種藥物對於酵母菌做測試。而有關這些酵母菌基因的註解資料是由 Gene Ontology（http://www.geneontology.org/）的網站下載而來。

本階段的實驗，無論是項目對或是中介皆為單一基因。而於挖掘間接關聯規則的第一階段，利用 Apriori 演算法產生所有的大項目集合，其支持門檻設為 0.1，這個相同於[8]中所使用的標準。然後再觀察，在不同項目對支持門檻與中介信賴門檻下，所得的規則數及相對應在 GO 關聯性上的變化。

從表 2 中我們可以看的出來，在中介支持門檻設為 0.1 的狀況下(因為大項目集合支持門檻設為 0.1)，中介信賴度因必需高於門檻值，故門檻值設越大，則所得的規則數目相對就越少，而項目對支持度因必需低於門檻值，故門檻值設越小，則所得的規則數目相對就越少。所產生的規則數目，符合間接關聯規則的概念。

表 3 代表的是不同中介信賴門檻與項目對支持門檻下，所產生間接關聯規

則，其在 GO 的 MF 分支下，相關的關聯性強度。每一組門檻值會有兩組 Information Content 分數，上方為項目對各別與中介的 Information Content 的平均分數，下方為項目對的 Information Content 分數。例如，在中介信賴門檻設為 0.60 時而項目對支持門檻設為 0.100 時，項目對各別與中介的 Information Content 的平均分數為 0.9798，項目對的 Information Content 分數為 0.8939，代表在這組門檻參數下，項目對之間在 GO 上的關聯性強度，小於項目對各別與中介在 GO 上的關聯性強度的平均，這與間接關聯規則的概念是一致的。

表 2. 不同門檻值下所產生間接關聯規則的數目

中介支持門檻 0.1		中介信賴門檻									
		0.60	0.62	0.64	0.66	0.68	0.70	0.72	0.74	0.76	0.78
項目對支持門檻	0.100	8863	6977	4852	3134	1837	1058	481	176	59	16
	0.095	5383	4040	2587	1520	790	422	161	54	10	2
	0.090	4071	3026	1881	1076	540	277	96	32	5	0
	0.085	2037	1421	783	386	151	52	15	6	0	0
	0.080	1329	928	493	233	84	27	8	3	0	0
	0.075	450	299	137	60	11	2	0	0	0	0
	0.070	228	149	63	24	2	0	0	0	0	0
	0.065	88	58	19	7	0	0	0	0	0	0

表 3. 不同門檻值下所產生間接關聯規則，在 GO 的 MF 分支下，相關的關聯性強度；每一組門檻值會有兩組 Information Content 分數，上方為項目對各別與中介的 Information Content 的平均分數，下方為項目對的 Information Content 分數。

中介支持門檻 0.1		中介信賴門檻									
		0.60	0.62	0.64	0.66	0.68	0.70	0.72	0.74	0.76	0.78
項目對支持門檻	0.100	0.9798	1.0107	0.9589	0.9537	0.9394	0.9850	0.9357	1.1028	1.0025	0.7819
		0.8939	0.8710	0.8879	0.8519	0.8213	0.8644	0.9052	0.7664	0.7283	0.2590
	0.095	0.9745	1.0135	0.9465	0.9548	0.9802	1.0295	*0.8812*	1.1841	1.6243	2.3353
		0.8456	0.8360	0.8554	0.8275	0.8100	0.8884	*0.9461*	0.7885	0.9674	0.0000
	0.090	0.9352	0.9635	0.8795	0.8853	0.9113	0.9842	0.7947	0.9875	1.1302	
		0.7362	0.7261	0.7166	0.6713	0.6019	0.6579	0.6370	0.6150	0.8339	
	0.085	0.9370	0.9601	0.8387	0.8789	0.8965	1.0688	0.4240	*0.0695*		
		0.8071	0.7805	0.7633	0.6844	0.5801	0.5527	0.3809	*0.2780*		
	0.080	0.9315	0.9611	0.7913	0.8440	0.6780	0.8039	*0.2085*	*0.0000*		
		0.8315	0.8085	0.8002	0.6929	0.5905	0.4896	*0.5057*	*0.2780*		
	0.075	0.9825	0.9576	*0.7097*	0.7656	*0.2807*	*0.0000*				
		0.7763	0.7013	*0.7189*	0.6405	*0.4285*	*0.8339*				
	0.070	1.0802	0.9292	0.5978	0.9134	1.1677					
		0.8248	0.6672	0.5915	0.5428	0.0000					
	0.065	1.4238	1.1410	*0.7813*	1.3344						
		1.0469	0.8071	*0.8168*	0.7925						

若某些組門檻參數下所得的 Information Content 分數與間接關聯法則的概念不一致，違反了項目對各別與中介的 Information Content 的平均分數必需大

於項目對的 Information Content 分數原則，在表 3 中我們將以斜體的字體標示其 Information Content 分數,並以灰色底凸顯。而這些大致是發生在中介信賴門檻設定過高，導致所挖掘出來的規則於項目對支持度相對較高，其間接關聯的意義較弱。或是發生在項目對支持門檻設定過低，導致所挖掘出來的規則於中介信賴度相對較低，其間接關聯的意義較弱。在一般狀況下所找出來的間接關聯規則，在 GO 上的 MF 分支，皆能驗證其關聯性符合間接關聯規則概念。

因為 BP 所描述的動作必定是包含一個以上的步驟，不像 MF 只描述單一的生化活動，故其代表的生化活動是較為廣泛的，故並不適合用於驗證間接關聯規則。而基因於 CC 的 GO Term 由於註解過少，參考的訊息不足，故亦不適合用於驗證間接關聯規則。故在利用 GO 做驗證的部份，僅採用 MF 分支驗證關聯性強度。

五、結論

本論文應用了 KDD 當中的間接關聯規則演算法來挖掘基因表現資料。此演算法結合了關聯規則，型成了一個更結構化的關聯性架構，能夠同時挖掘出關聯規則與使用者感興趣的負關聯規則。而於實驗中，我們運用的酵母菌的資料，主要是用了 300 種藥物對於酵母菌做測試，我們以此資料研究在不同藥物測試下基因表現的關聯性，再以 GO 架構下的 Information Content 強度，來驗證所找出來的間接關聯規則的正確性。實驗結果，利用間接關聯規則應用於基因表現資料上，所找出來的規則符合在 GO 上基因之間的關聯性。所以，本論文所提出的方法架構，有助於在不同生物反應下找尋共同的關聯性研究。

在未來的研究中，我們更期望以具體的生物反應功能上的分類研究，提出更具體的驗證，來進一步證明方法架構的正確性及實用性。雖然本論文所提出的方法架構可以發現生物反應上的關係，但生物上面基因的關係往往是非常複雜的模式，在此方面本論文即可當作一個起始點，藉由探討這樣的關係，可以進一步研究更複雜的生物反應關係。

致謝

本研究由中華民國國家科學委員會所補助，計畫編號NSC 95-2221-E-006 -372。

六、參考文獻

[1] R. Agrawal, T. Imielinski and A. Swami, “Mining Association Rules Between Sets in LargeDatabases,” Pro. Of ACM SIGMOD Conference on Management of Data, pp 207-216. 1993.

[2] R. Agrawal and R. Srikant. “Fast Algorithms for Mining Association Rules,” Pro. 20th Very Large Databases (VLDB) Conference, pp 487-499, Santiage, Chile. 1994.

[3] Manoj Bhasin and G. P. S. Raghava, “SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence,” Bioinformatics 20: 421 - 423. 2004.

[4] Mehmet Bilgen, Mehmet Karaca, A. Naci Onus, and Ayse Gu"l Ince, “A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences,” Bioinformatics, Dec 2004; 20: 3379 - 3386.

[5] Volker Brendel, Liqun Xing, and Wei Zhu, “Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus,” Bioinformatics, May 2004; 20: 1157 - 1169.

[6] S. Brin, R. Motwani, J. Ullman, and S. Tsur. “Dynamic itemset counting and implication rules for market basket data.” In Proceedings of the International ACM SIGMOD Conference, pages 255–264, Tucson, Arizona, USA, May 1997.

[7] R. Chen, Q. Jiang, H. Yuan and L. Gruenwald. “Mining Association Rules in Analysis of Transcription Factors Essential to Gene Expressions,” Atlantic Symposium on Computational Biology, and Genome Information Systems & Technology. 2001.

[8] C. Creighton and S. Hanash. “Mining Gene Expression Databases for AssociationRules,” Bioinformatics Vol19 no. 1, pp. 79-86, 2003.

[9] M.B. Eisen, P.T. Spellman, P.O. Brown, and Botstein, D. “Cluster analysis and display of genome-wide expression patterns,” Proc. Natl Acad.

Sci. USA, 14863-14868, 1998.

[10] L. Kaufrnan and P.J. Rousseeuw. “Finding Groups in Data: An Introduction to Cluster Analysis.” New York: John Wiley & Sons, 1990.

[11] P. Kotala, P. Zhou, S. Mudivarthy, W. Perrizo and E. Deckard. “Gene Expression Profiling of DNA Microarray Data using Peano Count Trees (P-trees),” Online Proceedings of the First Virtual Conference on Genomics and Bioinformatics. 2001.

[12] J. MacQueen. “Some methods for classification and anlysis of multivariate observations.” Proc. 5th Berkeley Symp. Math. Statist, Prob., 1:281-297, 1967

[13] Ritu Pandey, Raghavendra K. Guru, and David W. Mount “Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data,” Bioinformatics, Sep 2004; 20: 2156 - 2158.

[14] I. Res, I. Mihalek, and O. Lichtarge, “An evolution based classifier for prediction of protein interfaces without using protein structures,” Bioinformatics, May 2005; 21: 2496 - 2501.

[15] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” in Proc. of the 14th International Joint Conference on Artificial Intelligence, Montreal, pp. 448-453, 1995.

[16] A. Savasere, E. Omiecinski, and S. Navathe. “Mining for strong negative associations in a large database of customer transactions.” In Proceedings of the 14th International Conference on Data Engineering, pages 494–502, Orlando, Florida, February 1998.

[17] Tamayo, P. Slonim, D. Mesirov, J. Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, T. Golub, “Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.” Proc. Natl Acad. Sci. USA, 2907-2912, 1999.

[18] P. Tan and V. Kumar. “Interestingness measures for association patterns: A perspective.” In KDD 2000 Workshop on Postprocessing in Machine Learning and Data Mining, Boston, MA, August 2000.

[19] P. Tan and V. Kumar. “Mining indirect associations in web data.” In Proc of WebKDD2001: Mining Log Data Across All Customer TouchPoints,” August 2001.

[20] P. Tan, V. Kumar, and J. Srivastava. “Indirect association: mining higher order dependencies in data.” In Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pages 632–637, Lyon, France, 2000.

[21] X. Wu, C. Zhang, and S. Zhang. “Mining both positive and negative association rules.” In Proceedings of the 19th International Conference on Machine Learning (ICML-2002), pages 658–665, Sydney, Australia, July 2002.

[22] The Gene Ontology (GO) Consortium, “Creating the Gene Ontology Resource: Design and Implementation,” Genome Res.Vol. 11, pp. 1425-1433, 2001.

[23] The Gene Ontology Consortium,“The Gene Ontology (GO) database and informatics resource,” Nuclide Acids Research, 32:D258-D261, 2004.