建構 cDNA 生物晶片之二元資料挖礦模式及其實證研究

簡禎富
清華大學工業工程與工程管理學系

林國勝
清華大學工業工程與工程管理學系

摘要

生物晶片之基因微陣列技術（microarray）及基因選殖技術（gene cloning）的突破，加上資訊科技的進展能力，使得生物科技和生物晶片的研究和應用在過去十年有非常蓬勃的發展，也因而衍生許多資料處理和分析的問題亟待克服，特別是生物晶片資料變數多而樣本數少的問題。本研究目的係針對 cDNA 生物晶片之二元資料的特性，發展生物晶片資料挖礦（Data Mining）方法和模式藉以探索與尋找疾病與特定基因的關係，並建構其規則以作為醫療診斷決策支援參考。本研究並採用史丹佛大學晶片資料庫中乳癌晶片資料以驗證研究效度，從四萬多個基因與 64 個樣本當中，使用顯著性分析（Significant Analysis of Microarray）與決策樹（Decision Tree）挖掘出具影響力的基因及診斷決策規則，從中萃取有價值之資訊，研究結果可以驗證本研究所提出之方法的效度。

關鍵字：生物晶片、資料挖礦、決策樹、微陣列技術、顯著性分析

A data mining framework for binary cDNA bio-chip data analysis and its validation

Chen-Fu Chien
Department of Industrial Engineering and Engineering Management, National Tsing Hua University

Kuo-Sheng Lin
Department of Industrial Engineering and Engineering Management, National Tsing Hua University

ABSTRACT

Owing to increasing breakthroughs for microarray in biochips and gene cloning technologies, biotechnology is now an emergent and promising industry worldwide. Although information technology advancements enable complex calculation and comprehensive data storage involved in biotechnology, a number of critical issues need to be addressed for both practice and research needs. This study aims to develop a data mining framework for analyzing huge bio-chip data that are different from the data addressed in manufacturing and service industries. In particular, specific genes between normal and abnormal individuals were extracted in decision rules to clarify the relationships among genes, and diseases. We adopt the breast cancer patient cDNA microarray dataset for validating the proposed approach. We firstly extracted significant genes from more than 44,000 genes and then use decision tree to derive classification rules to support medical diagnosis. The results showed practical viability of this approach.

Key words: Biochip, Data mining, Decision Tree, Microarray, Significant Analysis of Microarray

壹、緒論

生物晶片一次就能夠紀錄成千上萬個基因表現的樣型（gene expression pattern），再以光學儀器掃描檢測結果轉換成數據（numerical data）加以分析，已經根本地改變了醫療檢測的方式。在醫學上的應用，若能將人類基因機能加以推測，並將各個族群加以分類的話，必能更準確地得知各項機能；或者不同的基因表現用來區別不同的腫瘤型態；亦可藉此研發出和以往化學結構與作用機制完全不同的新藥，觀察或預測用藥後的成果及反應；並且可與臨床診斷結合發展特定的療法。目前同一種疾病的醫療方法及藥物都是相同，所以有時會發生同種症狀的不同病人產生不一樣的結果。主要原因是因為每個人身上的基因結構並非完全相同，未來基因的謎團逐漸解開之後，醫生就可依據每個人的基因表現或特色進行診斷及醫療行為決策。

目前正在發展將資訊和生物統計理論有效應用於分析生物晶片資料上，根據美國國家衛生院（NIH）定義：生物資訊是一門資料蒐集與運用的學科，它是一門將數學、統計、資訊科學、人工智慧等理論與技術運用於生物科學研究所取得之資料中，以有效地得到顯著成果的學科。生物基因晶片之基因微陣列技術的突破與電腦資料處理和計算的能力，加速了基因研究的速度，也因而衍生許多資料處理和分析的問題亟待克服。從生物晶片上記錄下來的龐大基因資料，使得生物資訊的資料取得、收集、整理、除錯、建檔及分析成為相當複雜的工作，要從中找出有意義的規則或樣型，更需要整合生物統計、資料挖礦分析與資料處理技術，將原始序列資料或晶片資料轉為有用的知識，以突破資料分析理論限制，並更有效且快速的做有系統的整理解讀和分析。

本研究歸納進行基因晶片資料挖礦（data mining）時，所面臨的挑戰主要包括：1.基因挑選（Gene Selection）：對於某特定類別挑選出最相關的基因的過程；2.分類（Classification）：從基因表現的樣型來分類疾病或預測結果用以進行最佳的治療方式，由於晶片資料是變數遠大於樣本數的型態，一般統計及人工智慧的方法極易因隨機變異的關係導致錯誤的發現（Bergeron, 2002）。

本研究目的係研究生物晶片資料挖礦方法與技術發展，並以史丹佛大學生物晶片資料庫中的乳癌資料集進行實證研究，以驗證本研究模式之效度。因此，本研究擬從問題定義與架構、資料預處理、基因分析模式、模式驗證及詮釋的完整分析過程，建立 cDNA 生物晶片之二元資料挖礦模式，從生物晶片資料進行基因分析解讀，探索具價值之資訊、判定危險因子關係與發展分析模式，提供生物晶片數字化、系統化的分析，以作為後續醫療檢測與診斷之依據。

本文架構說明如下：第一節敘述分子生物學與基因工程當今的發展、所面臨的問題與其特殊的資料型態。第二節介紹生物晶片與生物資訊，並回顧其相關研究與資料挖礦理論基礎。第三節提出 cDNA 生物晶片之次元資料挖礦模式。第四節以史丹福大學生物晶片資料庫中的乳癌資料集進行實證研究，以檢驗本研究所建構的模式的效度。第五節討論本研究貢獻和限制，並探討未來研究方向。

貳、理論基礎

一、生物晶片與生物資訊

生物科技在世界各國均被視為前瞻產業而大力推動，我國亦列為「兩兆雙星」重點發展產業之一，但在國內仍屬於萌芽階段。分子生物學是對於細胞及其組成分子彼此間的研究和操作運用，也是一門提供新物理及新化學的應用在生物學上的學科（何國傑等，2001）。以下從二個方向簡示遺傳物質的結構及功能。若從生物細胞分子構造由大至小來看分別是，細胞（cell）→細胞核（nucleus）→基因組（genome）→染色體（chromosomes）→基因（gene）→去氧核醣核酸（DNA）→鹼基（pair bases）。圖 1 說明從遺傳訊息傳遞的過程及功能。生物體細胞中的分子鏈（chain molecules）負責生物組織的功能表現及演化，所幸這些由核苷酸（nucleotide）及胺基酸體（amino acid monomer）組成的 DNA、RNA 及 proteins 能被清楚地轉換成數位符號的序列。DNA 即是 A、T、G、C 四種鹼基排列組合的序列。這些基因資料的數位符號有別於一般的科學資料較無不確定性且明確，然而，其相關性、位置及經濟等因素卻影響著基因資料的品質。另外，由於基因表現（gene expression）會在不同時間和不同情況下有著不同的表現，此外尚須考慮基因序列在不同樣本（species）中的內部雜訊（noise），這些因素大幅增加基因資料分析的困難性（Baldi and Brunak, 2004）。

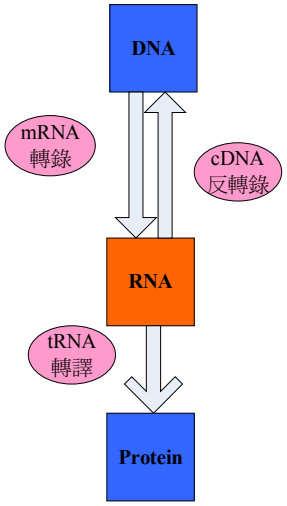

圖 1：遺傳訊息傳遞示意圖（資料來源：本研究整理）

生物晶片能夠在數小時內一次判別上萬的基因表現並且透過正常與生病的組織中取得的 mRNA 和 cDNA，運用生物晶片的特性將能從中判別是哪些基因、關連或環境狀況導致疾病發生。由於基因選殖（gene cloning）技術縮短基因複製的時間及正確

性，拓展快速有效的 DNA 定序技術，以及生物晶片的產生使研究者得以同時觀察數以萬計的基因表現，加速基因研究的速度及範圍。

生物資訊所涵蓋的對象包含 DNA、RNA、mRNA、cDNA、Protein secondary structure 等，從標的萃取、複製技術到晶片製程品質、雜交實驗的 DNA 濃度，到最後判讀數據的分析等皆影響資料分析的結果。

半導體設計與製造技術已成功引進到生物晶片上，發展出包括目前最廣泛的 Affymetrix，使得微陣列晶片更加縮小且涵蓋更多的點。針對不同半導體的問題，已發展具體有效的資料挖礦方法，包括事故診斷（Chien *et al.*, 2002）、良率提升（簡禎富等，2003）、生產效率（簡禎富等，2004），及人力資源管理（簡禎富等，2005）等相關分析。

二、生物晶片資料分析相關研究

生物晶片資料的特性就是高維度而樣本數卻很少。因此基因挑選可視為一個困難的變數挑選（Feature selection）的問題。Tusher *et al.*（2001）使用實驗設計的方式以不同細胞、基因、實驗組及對照組觀察基因的檢測值是否會因輻射而變化，使用顯著性分析法(Significant Analysis of Microarray; SAM)，SAM 法類似統計 T 檢定以相對差異是否顯著，說明依照不同準則分組下，何種準則具有顯著差異，以證明輻射會影響細胞及基因的量測值，並且以錯判率驗證 SAM 法表現較傳統的倍差法(Fold change method)有效。Mukherjee *et al.*（2004）提出拔靴法的基因排列演算法解決小樣本的 T-統計量變異的有效方法，用來尋找基因與發病的關係，由於本類問題的特性即是樣本數小，因此作者採用拔靴法創造資料量，再以此人工資料集進行 T 檢定並將具影響力之基因排序。同樣 Chen（2003）提出一個系統化的方法用來選取癌症分類的基因集合，利用基因演算法及支援向量機演算法(Support Vector Machines; SVM)進行基因分類，目標是選出能判別不同的樣本來自於不同類別的基因並藉由 SVM 來評估其判別力；基因演算法則是用來尋找最佳的子集；拔靴法則是用來創造資料以克服小樣本的難題，最後利用基因演算法得到的候選基因集合計算個別基因出現的次數當作顯著基因的說明。Wu 和 Zhang（2004）提出一個有效率的變數選取方法，稱為因素重要度。這個方法考量了兩個部分：資訊增益（information gain）來衡量每個因素的貢獻度；與相關係數（correlation）來表示因素間的相關性。因為這兩個部分都會提供某種程度的資訊，藉由這兩者結合的最佳資訊量，找出最重要的因素，以降低高維度數值分類的問題，藉以選取具影響力的因素。Miyamoto *et al.*（2003）將變數選擇的問題定義為從候選變數集合中，挑選出在某一分類系統下表現最佳的子集合。這個過程不僅可因減少變數量而降低成本，同時能提供較佳的分類正確性。首先採用倍差法找出基因表現值在 A、B 群差異大於二倍者；依照 Fisher criterion 選取最佳的 50 個；從這 50 個中反覆搜尋選出最理想的 d 個（C_d^{50} 次）。Chuang *et al.*（2004）整理變數選取的方法，區分為：(1)參數法：利用統計方法辨別基因對於類別間的差異的影響，並給予分數。這些方法皆奠基於假設這些資料都具備同樣的分配；(2)無母數法：使用排序結果進行分

析，並設定懲罰函數。作者結合 T-test、Fisher criterion、Golub criterion、Threshold Number of Misclassification (TNoM)、Minimum Distance to Modal Ranking (MDMR)及 WEPO 等 6 個方法並提出一個新的研究架構，稱之為「Ranking and Combination analysis, RAC」。RAC 包含兩個階段。第一階段是收集經由不同方法所得到的基因排序結果。第二階段是藉由作者提出的轉換函數，將以上 6 種方法的各別排序倒數相加後轉換成一個最終排序結果，並將基因排序表現當做基因選取的指標。

表 1：生物晶片資料分析相關研究（本研究整理）

作者（年代）	使用方法	理論概念	功能
Chen(2003)	基因演算法、support vector machines(SVM)	創造資料量	基因排序
Mukherjee(2004)	拔靴法、基因演算法	創造資料量	基因排序
Tusher *et al.*(2001)	SAM	降低變數維度	基因選取
Miyamoto(2003)	Fold change method and Fisher criterion	降低變數維度	基因選取
Wu and Zhang(2004)	資訊增益及相關係數	降低變數維度	基因選取
Chuang(2004)	Ranking and Combination analysis(RAC)	降低變數維度	基因選取
SØrlie *et al.*(2001)	分群分析	相關性	關係建立
Troyanskaya(2003)	貝氏網路	相關性	關係建立
Khabzaoui *et al.*(2004)	關聯法則、基因演算法	相關性	關係建立

Khabzaoui *et al.*（2004）使用關聯性法則（association rule）探索基因尋找樣型（pattern）。關聯性法則也許能被塑造成一種最佳化問題。本篇提出一個多屬性模式的關聯法則問題並且結合基因演算法設計及關聯性法則來處理生物晶片資料。使用分群或分類的方法將每個實驗結果的基因分群，則分群後的基因就如同購物籃法則的項目而各個實驗則視為交易紀錄。進行方式為使用遺傳演算法產生資料（候選集合），將其依照形成關聯法則的形式分群，並從以上關聯法則的評估原則折衷評估選取出規則。Troyanskaya *et al.* （2003）提出一種使用貝式網路整合生物資料用來做基因功能預測的網路架構（Multisource Association of Genes by Integration of Clusters, MAGIC），結合專家知識架構的系統、基因表現、物理上的交互作用與生物狀況，然後根據輸出值提供一個信心水準（belief level）並允許使用者依其需要而變化。MAGIC 計算該群內與生物狀況相關性的正確性。SØrlie *et al.*（2001）使用晶片資料上的基因表現進行階層式分群（hierarchical clustering），並使用 SAM 法萃取出顯著基因，再結合病歷資料觀察分群基因與病歷資料的關係，或觀察病歷資料（如存活時間）依分群狀況而改變的情形。本研究將相關文獻依其理論概念及提供之功能整理如表 1。

三、資料挖礦技術

資料挖礦是一種不斷循環的資料分析與決策支援過程，主要是以自動或是半自動的方式如：統計分析方法、決策樹、貝氏網路（Bayesian Network）、集群分析（Clustering）、機器學習、類神經網路（Neural Network）等，從大量資料中探索和分析，以發現出有意義的樣型（pattern）或規則（rule），並將其整理成有價值的知識（簡禎富等，2001）。因此在整個分析過程當中，資料挖礦是屬於探索導向（Discovery-Driven），並非假設導向（Assumption-Driven）；資料挖礦模式是由相關實證資料推導的（Empirically Derived）而非理論架構。首先了解應用領域，包含相關且具有價值的知識以及應用的目標。建立目標資料集，選擇一個適合分析的資料集或是相關變數之子集。然後進行資料預處理作業以去除混雜其中的不相關資料及清理、合併、分箱等去蕪存菁的方式轉化處理，以確保分析資料的品質和分析結果的正確性，或是將資料投射和簡化以轉換成適於應用分析目標的格式。接著選擇合適的資料挖礦的工具，推導資料挖礦演算法及分析模式。最後與領域專家合作透過闡釋將探索出的資訊以一種可以被確認、觀察和再使用的形式呈現，而使決策者能夠理解，並將這些所得有價值的資訊實際應用於現實系統上，以確認這些知識的價值性。

資料挖礦在銀行業、商業行為及工業等有許多成功案例，如製程良率提升（Yield Enhancement）或錯誤偵測（Fault Detect & Classification）等均有良好的成效（Chien *et al.*, 2007）。然而這些案例一般來說都是具有大量的資料（或紀錄）及遠小於資料量的變數（或屬性）；相反地，生物晶片資料由於樣本蒐集的困難性和實驗成本的關係，卻是極少量的資料（大多是數十到數百筆）卻擁有大量的變數（即由數千到上萬基因所構成）。既有的資料挖礦方法並不能直接用來分析生物晶片的資料。因此，本研究針對資料特性發展更適合分析生物晶片資料的架構和模式。

由於挖掘目標和問題不同，配合所使用的挖掘工具，可得到不同類型的挖掘結果，資料挖礦結果的類型包括：關連規則（Association Rule）、分類（Classification）、聚集（Clustering）、預測（Prediction）。資料挖礦藉由使用不同的工具來執行，如：貝氏網路、決策樹、統計方法、類神經網路、遺傳演算法等。其中，決策樹（Decision Tree）在資料挖礦的領域裡通常扮演監督式特徵萃取與描述的角色，解決分類型態的問題。透過變數的選擇與目標（target）的指定對資料進行分類而成樹枝狀的架構。對於資料挖礦之主要用途是能夠將輸入變數依據某種規則或資料特性對資料進行分類，並以樹枝狀方式表現類別之間由輸入變數所造成的區別，故藉由決策樹的分析規則可對資料進行層級架構的分類。利用訓練完成的決策樹架構，亦可對資料進行比較或預測分析。決策樹是一個廣泛應用來架構決策問題的工具，應用於資料挖礦之主要功能在於能夠將輸入變數依據某種規則或方法對資料進行分類，並以樹枝狀方式表現出來，進而挖掘出顯著影響結果的因子，故可以藉由決策樹的分析規則對資料進行層級架構的分類。

目前相關研究採用諸多啟發式演算法進行資料分析，生醫工作者在判定某基因是否顯著的典型簡易處理方式如公式(1)，將實驗組病人在基因 i 的表現平均值除以對照

組在基因 i 的表現平均值，取對數函數僅為壓縮數值表現及數值標準化的意義，通常以 2 為底。

$$r_i = \log_2 \left| \frac{\overline{R}_i}{\overline{G}_i} \right|, |r_i| \text{ or } \frac{1}{|r_i|} \geq threshold \tag{1}$$

其中 R 表病人；G 表正常人的基因表現值。舉例而言，假設已知基因 i 與 j 在病人與對照組正常人的表現(i, j)分別為{(8080,319), (14117,61), (2861,44), (3197,8), (7625,116)}、{(410,552), (170,115), (432,38), (306,65), (325,287)}，並設定門檻值為 2。計算結果如下，則可判定基因 i 為顯著基因，基因 j 為否。

$$r_i = \log_2 \left| \frac{\overline{R}_i}{\overline{G}_i} \right| = \log_2 \left| \frac{7176}{328.6} \right| = 4.45 > 2,$$

$$r_j = \log_2 \left| \frac{\overline{R}_j}{\overline{G}_j} \right| = \log_2 \left| \frac{109.6}{211.4} \right| = -0.95, \frac{1}{|r_j|} < 2$$

表 2：決策樹整理表（簡禎富等，2005）

決策樹演算法	作者	可處理資料型態	分支法則	演算公式
CHAID	Hartigan(1975)	離散	卡方檢定	p-value
CART	Breimen *et al.* (1984)	離散、連續	Gini value	$1-\sum_i P_i^2$
ID3	Quinlan(1986)	離散、連續	Entropy	$-\sum_i P_i \log P_i$
C4.5	Quinlan(1993)	離散、連續	Gain-ratio	$\frac{gain(x) = \inf(t) - \inf x(t)}{\text{split inf}(x) = -\sum \frac{N(t_i)}{N(t)} \log \frac{N(t_i)}{N(t)}}$

雖然使用公式(1)可簡易判斷基因是否顯著，但卻難以決定其門檻值；尤其生物晶片資料的特性除了具有變數維度遠大於樣本數的問題，導致自由度不足之外，且往往不服從常態分配，亦無法進行統計 Z 檢定或 T 檢定，因此不論是進行基因重要性排序或分群建立關係，皆無法有效提供建設性的參考閥值。而決策樹分析不但可建立基因與疾病的網路關係，辨別具影響力之基因，更可有效提供參考閥值。

決策樹的演算步驟，首先依照選取的分支法則選取最佳的分支變數當作起始節點（node）依序建構完整的決策樹，建立之路徑（path）或規則稱之為樹枝（branch）；再將未達到分支標準門檻值的樹枝予以刪除，包含分支水準與樣本數，謂之修剪（prune）。影響決策樹分類規則的因素，除了資料本身之外，包括分支輸入變數以及分支法則演算法。決策樹演算法發展至今大致可分為 CHAID（Chi-squared automatic interaction detection）(Kass, 1980)、CART(Breiman *et al.*, 1984)、ID3 及 C4.5(Quinlan,

1993)等，如表 2 所示。CHAID 採用統計的卡方檢定作為分支依據，因此用作處理非連續型變數資料是一個有效的演算法，所以若是資料型態為連續型變數則須先行處理將資料離散化；CART 是二元分類法則的決策樹，其分類法則為最小化錯分率（Gini-index），可處理連續型或離散型態的資料；ID3 則以熵度（Entropy）為分類法則，可處理連續型或離散型態的資料；而 C4.5 則是採用 Gain-ratio 為其分支法則，可處理連續型或離散型態的資料，進一步之整理比較可參考表 2。假設有癌症病人 15 筆基因資料紀錄以及對照組正常人資料 15 筆，經決策樹分類後發現當 Gene 1>2000 時可清楚區分出病人/正常人，如圖 2 所示。

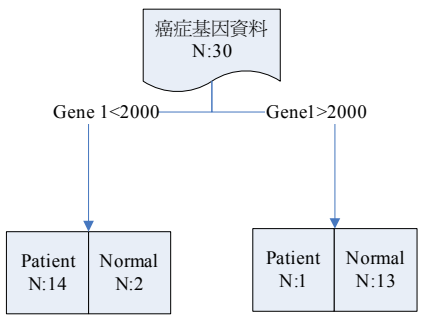

圖 2：決策樹示意圖

為了因應生物醫藥的資料型態，各種新的資料庫結構也相繼出現，這些資訊系統被設計來管理並收集生物晶片的資料。生物晶片所產生的大規模實驗資料，衍生出複雜的資料處理問題。對於這些大量且複雜的資料，資料挖礦將可提供快速且正確的分析方法。

參、生物晶片資料挖礦架構

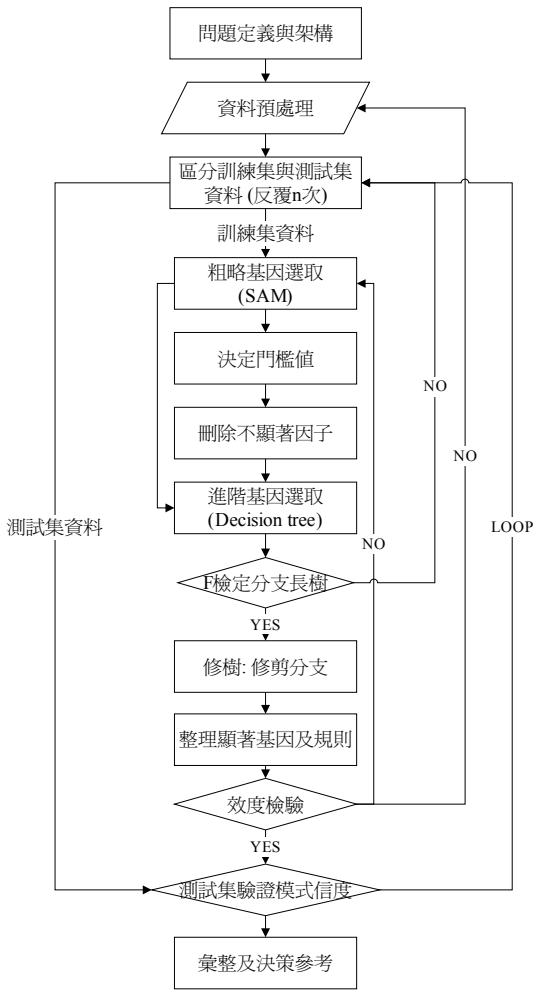

圖 3：研究架構

本研究針對生物晶片資料之特性與問題需求，特別是生物晶片資料變數多而樣本數少的問題，根據資料挖礦的循環過程，發展 cDNA 生物晶片之二元資料挖礦架構，主要包括兩大階段，第一階段為利用 SAM 法初步先篩選顯著的基因；第二階段則利用決策樹分析進一步篩選基因，並進而萃取所建立的決策規則，然後結合領域專家的經驗和知識對結果做詮釋和解讀，以整理發掘有意義的資訊，提供相關決策和判斷之根據。本研究所提出的資料挖礦架構，主要步驟包括：問題定義與架構、資料預處理、生物晶片資料之決策樹建構、模式驗證、結果詮釋與說明等五個步驟（如圖 3）。本模式為資料導向(data-oriented)的模式，利用結合資料特性之演算法則，不僅取決於所獲得的樣本多寡、資料完整性，亦受到生物特性變異所影響。具體步驟說明如下：

一、問題定義與架構

生物晶片萃取基因表現的方式，是以微陣列製備儀（Arrayer）放置人類或生物的單股基因探針，由實驗組抽取出標地核酸，以紅色螢光劑標示，同時將標地核酸對照組以綠色螢光劑標示，將兩組標地核酸均勻混合後，再與已放置基因探針的玻璃片進行雜交反應，經螢光掃瞄儀分析後，若玻璃片上某些探針點位置只出現紅色亮點，則表示該基因只有在實驗組會表現，而正常人並沒有表現；反之，則表示該基因只有在正常人會表現，而在病人並沒有表現。各個基因表現從頻譜變化顏色轉換後的數據顯示之（如圖 4 所示），然而實驗數據的變化除了受個別基因的影響，尚受到其他干擾因素的影響，本研究整理干擾實驗數據的因素，以實驗準備面、設備面及數據面將變異區分如表 3。因此，一次的實驗中，即可獲得實驗組與對照組的所有基因表現報告。

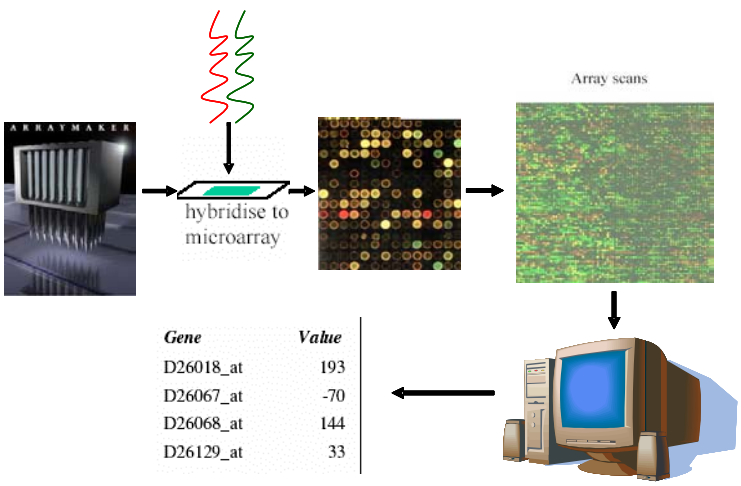

圖 4：生物晶片分析示意圖（資料來源：本研究整理）

一片生物晶片可同時解析出上千種基因表現，龐大的資料和數據，若未經更進一步的資料的處理和分析，將無法從中發現哪些基因將會導致致病因子的產生。目前一片晶片的價格可能高達 500 美金，使得資料蒐集不易，受測者的樣本數往往遠小於實驗變數個數，不僅增加了生物晶片分析的困難度，更遑論檢視各個基因彼此之間的交互作用。本研究針對生物晶片上 Cdna 之二元資料發展資料挖礦理論模式，從中刪除不敏感之基因並擷取可能具影響力之基因，並從分析過程中獲得有意義之樣型或歸納出規則，搜尋出基因在正常人與病人不同的表徵，以及藉由了解基因與致病因子之間的關聯，以提供醫學研究者針對特定的疾病或症狀下判斷之依據。

二、資料預處理

為提高分類結果的準確性與效率，需對資料進行預處理的動作，包括資料整合、資料清理、資料轉換。由於資料來源在同一次實驗可能來自於不同的資料集，因此在進行分析前必須先經過整理，此外由於每一樣本的檔案高達數 10MBytes，整合數十樣本動輒數百 MB 甚至一個 GBytes 以上，因此需要透過資料庫進行合併、清除多餘的紀錄與資料。在分子生物學發展的初期由於對基因的了解尚不健全加上命名法尚未統一，造成同一基因（或片段）有多個命名，然格式統一之後仍有其存在的意義，因此要進行欄位名稱的整理，並去除不必要分析之欄位。資料清理過程則對遺漏值與空白值進行處理，包含刪除空白值過多的變數、進行遺漏值的補值、刪除無效的屬性。針對無法控制（或判別）之潛在變異，即基因 i 非某特定疾病之顯著基因，若某些正常人基因 i 表現異常時，則可能造成誤判。為避免離群值對模式造成偏誤，在資料清理時將正常人基因表現離群值予以刪除。另外由於數據的變異來源可能來及不同機器，機器在各個基因判讀出來的數據受雜交情形、環境影響或探針位置等等原因的影響，造成數據的不一致性(即並非某個基因表現數據特別高便是顯著的)，因此需要將數據資料標準化。判讀出來的數據受到以下變異的影響，並假設同一批實驗下基因 i 在病人/正常人來自機器及實驗的變異是相同的。

$$\sigma(i)_N^2 = \sigma(i)_{machine}^2 + \sigma(i)_{batch}^2 + \sigma(i)_{N-inner}^2 \tag{2}$$

$$\sigma(i)_A^2 = \sigma(i)_{machine}^2 + \sigma(i)_{batch}^2 + \sigma(i)_{A-inner}^2 \tag{3}$$

其中，N 代表正常（normal）而 A 代表異常（abnormal）。

最後，將經過預處理過後的資料採用比例原則將資料集隨機區分為訓練資料集(80%)與測試資料集(20%)。並採用隨機重複抽樣（re-sampling）的方式進行，並藉用交互驗證（cross validation）的概念，訓練集與測試集為 80％：20％，重複抽樣 n 次以計算平均正確率，各次訓練資料集主要用以建構生物晶片資料之決策樹模式與規則，測試集資料則用以衡量模式之效度。

表 3：生物晶片變異來源（本研究整理）

實驗準備面	數據面	設備面
cDNA 的濃度	實驗設計	點陣位置
試劑的準備	點陣數值化的強度	設備問題
cDNA 的粘附狀況	雜交情形	機器手臂的正確性
微陣列表面的化學反應	基因的品質	
	環境變化（溫度等）	

三、生物晶片資料之決策樹建構

針對生物晶片資料不易蒐集，量測基因眾多的特性，一般變數投影轉軸方式無法處理，因此，本研究整合 SAM 法及決策樹分析，建構生物晶片資料基因之決策樹架構。先利用 SAM 法進行以初步挖掘較顯著的基因；接下來採用資料挖礦的決策樹方法找出對模式具顯著影響的基因，並進一步建立決策規則。具體步驟敘述如下：

步驟一：以 SAM 法粗略篩選敏感（sense）基因

根據基因表現值在不同狀態下之差異，利用重複量測及差異標準化的方式，給定各基因相對評分，超過類似 T 檢定值彈性調整接受範圍之基因，判定該基因變化具統計顯著差異，最後利用不同樣本之排列與重複量測的方式評估各基因之鑑別率(FDR)。由於基因表現值的變動對基因具專一性，首先定義各基因之相對差 d_i：

$$d_i = \frac{r_i}{s_i + s_0}, i = 1,2,...p \tag{4}$$

$$r_i = \frac{\sum_j y_i (x_{ij} - \bar{x}_i)}{\sum_j (y_j - \bar{y}_j)^2} \tag{5}$$

x_{ij} 為樣本 j 的基因 i 表現值；y_j 為樣本 j 之依變數；r_i 為基因 i 之線性迴歸係數

在本研究模式中，y_j 為正常人/病人成對資料，即依變數-k 與依變數 k 成對，$y_j \in \{-1,1,-2,2,...-K,K\}$

$$s_i = \frac{\hat{\sigma}_i}{\sqrt{\sum_j (y_j - \bar{y}_j)^2}} \tag{6}$$

$$\hat{\sigma}_i = \sqrt{\frac{\sum_j (x_{ij} - \hat{x}_{ij})^2}{n-2}} \tag{7}$$

$$\hat{x}_{ij} = \hat{\beta}_{i0} + r_i y_j \tag{8}$$

$$\hat{\beta}_{i0} = \bar{x}_j - r_i \bar{y}_j \tag{9}$$

s_i 為 r_i 之標準誤；$\hat{\sigma}_i$ 為殘差均方誤

$d_{(i)}$ 為由基因的相對差異（relative difference）順序統計量（order statistic），而 $\bar{d}_{(i)}$ 為 $d_{(i)}$ 平均。選定適合之門檻值△。對大部分基因 $d_i \cong \bar{d}_i$，若基因兩者間差異大於設定之門檻者，則判定為具顯著差異。為了預估 SAM 可能誤認之顯著基因數，以所有判定顯著基因之最小 $d_{(i)}$ 與所有被抑制基因之最大之負 $d_{(i)}$ 為水平界線，計算對於各排列所得超出界線之基因數之誤認數，最後以各組排列之平均為預估之誤認數。SAM 法可有效而簡易的過濾在各生物狀態具差異之基因，並提供篩選出之各基因其鑑別率資訊，以彈性縮小進行深入分析的範圍。

步驟二：以降低變異為分支準則進階篩選敏感（sense）基因

為了要最小化決策樹葉節點中目標變數的變異，使用降低變異準則作為分支準則，透過卡方自動檢定對基因作進階選取，以進一步萃取有意義的基因規則。目標變數定義為病人(0)或正常人(1)，以 SAM 篩選後的顯著基因為分支變數進行長樹，若基因為連續尺度，則對基因區間進行分組，並計算各組檢定 p-value 值，若大於顯著水準α則將二組合併，直到所有組別檢定 p-value 均小於α。將所有顯著基因作為候選分支變數，再從中選出可降低最多變異的基因作為分支變數。

步驟三：長樹與規則萃取

若有分支變數可降低目標變數變異，則繼續進行分支，否則結束為葉節點，反覆生長決策樹，直到所有的節點無法符合分支準則或都是葉節點。根據決策樹分支架構與葉節點所代表的意義，每一條由根節點到葉節點的路徑可萃取出 If-Then 規則。

四、模式驗證

利用反覆 n 次之測試集資料對萃取之規則進行驗證與評估模式效度，將挖掘出的規則套用在測驗集上，測試其平均正確率。考量在小樣本下，採用重複抽樣方法反覆進行交互驗證評估建構模式之效度，以求取最佳決策規則及顯著基因。驗證作法為依研究個案分別訂定決策規則正確率、偽陰性（false negative）及偽陽性（false positive）門檻作為驗證標準（在醫學上偽陰性，即未將病人檢驗出來較偽陽性顯得重要）。萃取通過驗證的模式之規則，解釋分析結果與規則，提供醫學研究者參考依據；若模式驗證未通過，則刪除該筆規則或返回資料蒐集、資料預處理步驟，重新進行資料準備，或修正分析模式及其參數，並反覆循環此一步驟，直到模式驗證通過。

五、規則解釋與評估

資料挖礦並不是資料打撈，也不是把資料利用資料挖礦工具和技術來分析就可以達到目的。最後必須結合領域專家對分析結果進行詮釋與說明，歸納對某特定疾病具影響力之基因樣型特徵，提出數字化檢測數據與醫療檢測決策法則供醫學研究者參考。整個分析過程至挖掘的結果，應不斷反覆重複上述研究步驟並與領域專家討論，以擷取其經驗與意見改良模式，使得研究模式與資料挖礦結果更能達到目標。

肆、實證研究

一、問題定義與架構

根據本研究架構，選用史丹佛大學的生物晶片資料庫(Stanford Microarray Database; SMD)中乳癌實驗晶片 cDNA 資料進行研究，各晶片約包含 45,696 個基因(探

針點）與病人、非病人各一位樣本，反應後所得之表現值，總計 64 筆晶片資料分別儲存於 64 個 EXCEL 檔中，每個原始檔案約包含 46,000 乘 64 個儲存格，部份原始資料如表 4，表 4 為編號 18195 晶片資料，每列為各個不同基因，每欄表示各個基因不同表現值，包含基因名稱、座標、基因強度表現等等，共 128 筆樣本生病與非生病各半，如 Spot 為探針流水編號、Accession 為基因名稱，而 Ch1/Ch2 Net 之數值為各基因相對應之正常人/病人基因強度表現。因此可獲得 64 片晶片資料的 64 位病人在 45,696 個基因的晶片表現值；及其對照組 64 位正常人在 45,696 個基因的晶片表現值。

針對本組基因晶片的資料，從中刪除不敏感之基因並擷取可能具影響力之基因，並從分析過程中獲得有意義之樣型或歸納出規則，以提供醫學研究者對於先期判斷乳癌之依據或後續研究之參考。因此首先決定須分析變數為表 4 中基因名稱(Accession No.)、CH1 Net、CH2 Net 分別為正常人與病人之基因表現，並根據本資料特性和問題特徵擬定假設以進行分析，假設實驗來自同一批製程下之晶片且實驗進行為同樣的時間及環境，並假設基因 i 在病人/正常人來自機器及實驗的變異是隨機變異且相同，再進行資料預處理動作。

表 4：生物晶片原始資料表

	A	B	C	D	E	F	AN	AO	AP	AQ	AR	AS	AT
12	!Country=USA												
13	!SlideName=shbg054												
14	!Printname=SHBG												
15	!Channel 1 Description=Stratagene_aT												
16	!Channel 2 Description=BC-D-091_aT												
17	!Scanning Software=GenePix												
18	!Software version=Pro												
19	!Scanning parameters=PMTVolts=630650LaserPower=5.294.05												
20	Spot	Clone ID	Gene Symbol	Gene Name	Cluster ID	Accession	Channel 1	Ch1 Backg	Std Dev of	Ch1 Net (M	Ch2 Intens	Ch2 Net (M	Ch2 Net
21	1	IMAGE:34	ITGB2	integrin, beta	Hs.375957	W68291	252	235	102	3385	2488	2132	212
22	2	IMAGE:233721				H78560	292	256	138	5238	2820	2478	236
23	3	IMAGE:50	MTIF2	mitochondria	Hs.149894	H18070	271	252	106	10591	6245	5894	605
24	4	IMAGE:12	HMBS	hydroxymeth	Hs.82609	R06263	275	255	83	5514	3385	3034	283
25	5	IMAGE:23	GATA6	GATA bindit	Hs.50924	H77651	265	242	99	19392	9912	9564	992
26	6	IMAGE:18	ICAM5	intercellular	Hs.151250	R87763	293	267	137	4249	3498	3147	314
27	7	IMAGE:80	SEMA3B	sema domain	Hs.82222	AA455145	315	235	624	34020	22669	22333	2104
28	8	IMAGE:18	RIN2	Ras and Rab	Hs.446304	R83223	268	173	688	4744	5723	5462	570
29	9	IMAGE:29	CSE1L	CSE1 chrom	Hs.90073	N69204	139	70	109	1348	1177	1087	43
30	10	IMAGE:119384				T94272	180	188	124	419	1427	1155	6
45713	46069	IMAGE:36	CENTB5	centaurin, be	Hs.21446	AA024391	229	222	77	126	607	254	17
45714	46070	IMAGE:81	PIK3R2	phosphoinos	Hs.211586	AA485731	261	219	584	2029	2428	2059	203
45715	46071	IMAGE:78	MGC33630	hypothetical	Hs.359981	AA448270	277	215	682	300	1182	807	59
45716	46072	IMAGE:81	LOH12CR1	loss of heterc	Hs.105040	AA459384	235	218	143	1378	1351	952	85

18195

二、資料預處理

資料的預先處理包含無效屬性刪除、資料整合或解構、遺失值的填補等，可以提高分類過程的準確性和效率性，本研究並不盲目的將所有的資料皆放進模式分析，而事先將資料以去蕪存菁的方式轉化處理，以確保分析資料的品質和結果的正確度。首先整理各別晶片資料(subset)將冗餘及不需要之名目（nominal）欄位去除，如 spot、gene name、gene symbol、gene ID，僅保留 Accession No.（即唯一且統一的完整基因編碼），再將不需要及無效之數值欄位去除，如 Accession No.名稱遺失。再將個別 Accession No. 遺失值過多者去除（20%遺失值，本資料即遺失值超過 25 個），不予做之後的分析。

在資料預處理中，除冗餘及不需要之名目（nominal）欄位去除工作以手工執行外，其餘由於生物晶片資料量龐大，包含撰寫資料庫程式以彙整 64 個各別晶片所需之資料表、去除重複紀錄、遺失值過多者以 Visual Basic 撰寫程式協助處理，完成資料預處理之資料集，即以 Accession No.為主要分析變數、CH1 Net 與 CH2 Net 在個別基因的表現為依變數，共計 41,681 個基因(列)與 128 筆樣本(欄)，再將遺失值未達刪除標準者以 K-nearest neighbor 法進行補值（如表 5）。配合步驟一 SAM 法，表 5 中第一列標記為第 i 個樣本，其中負號者表示正常人，並且資料為成對樣本，如第 3 個正常人在 H71721 的基因表現值為 67。為更清楚呈現資料轉換邏輯，將原始資料與可分析表格之資料轉換示意如圖 5，然實際之資料並非如圖 5 所示排序完美，如 C1 晶片之基因排序依序為 g1、g2、g3 而 C2 晶片則為 g3、g2、g4、g5。

接著將整理過後之資料集採用隨機重複抽樣的方式，分別區分五次個別之訓練集與測試集，各包含 100 筆與 28 筆樣本。

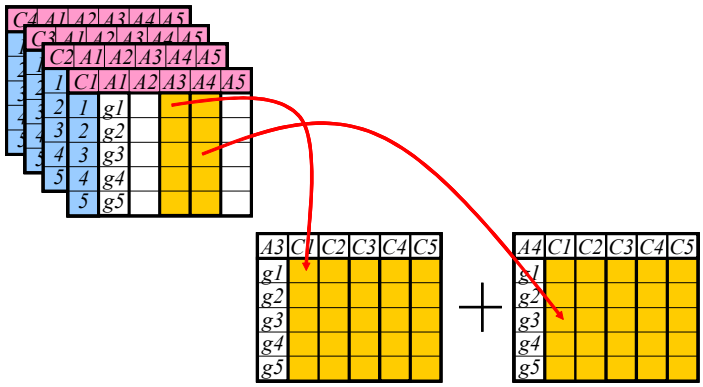

圖 5：資料轉換示意圖

表 5：部分整理後資料表

Accession No.	Item	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
H71697	33365	1690	552	328	512	205	937	218	1172	682	738	1273
H71703	33366	5010	494	390	305	192	3239	174	4002	1295	711	1479
H71713	33367	5511	254	247	638	288	726	502	695	209	165	290
H71719	33368	1140.5	246.5	142.5	269.8	105.9	836	166.3	491	292	248	588
H71721	33369	390	91	67	111	24	776	18	519	356	336	419
H71722	33370	3752	367	234	732	27	2019	77	2719	1889	2032	2864
H71725	33371	3789	586	422	753	335	1483	249	1390	1003	633	989
H71752	33372	2599	281	237	643	314	1225	129	1092	746	699	1050
H71756	33373	2345	370	73	203	108	163	106	61	83	65	139
H71759	33374	3266	11	29	7	20	220	13	263	227	143	281
H71812	33375	1660	227	205	268	106	1010	105	1123	640	416	723
H71822	33376	1751	209	153	53	66	2039	88	2094	1020	735	1371
H71824	33377	3394	463	371	480	372	4269	250	2699	1550	1745	2423
H71847	33378	3489	272	161	349	149	627	162	731	634	528	366
H71850	33379	780.667	224.556	146.222	303.111	111.556	1178	131.333	716	701	248	622
H71853	33380	1137	295	223	368	234	1050	216	824	585	621	624
H71940	33381	264	75	38	148	82	5037	54	4511	2359	2242	2202
H72027	33382	14005	2668	2444	1884	1361	12054	1854	5306	3543	4270	3928
H72029	33383	6636	126	97	255	7	1587	100	3550	7164	3110	4692
H72080	33384	5083	22	76	191	11	142	22	115	45	41	83
H72083	33385	1198	268	137	333	64	-6	126	528	72	306	785
H72086	33386	9288	4973	3538	4679	1256	8	2074	6655	6821	2802	6425
H72089	33387	374	38	58	161	20	527	25	339	54	152	400

Sheet1 / Sheet1 (Imputed) / SAM Work (Do not edit!) / SAM Sample Size Plotsheet / SAM Plot / SAM Output

三、生物晶片資料之決策樹建構

步驟一：將處理過之病人與非病人各 64 筆資料所彙整之資料表，任意成對挑選出各 50 筆共 100 筆作為訓練資料，剩餘各 14 筆共 28 筆做為最後驗證用資料，並重複抽樣五次。設定門檻值△=3。分析後分別得到 11,104、12,829、13,219、12,770、13,745 個較顯著基因，部分分析結果如表 6 及圖 6。表 6 為經 SAM 分析後顯著基因列表，紅色顯示為超過管制線之基因，反之為綠色，Score(d)、Numerator、Denominator 則分別為公式(3)之 d 值、分子、分母值。圖 6 為基因表現散佈圖，超過上下二條管制線者判定為顯著基因。表 6 則列出各個顯著基因及其檢定值。

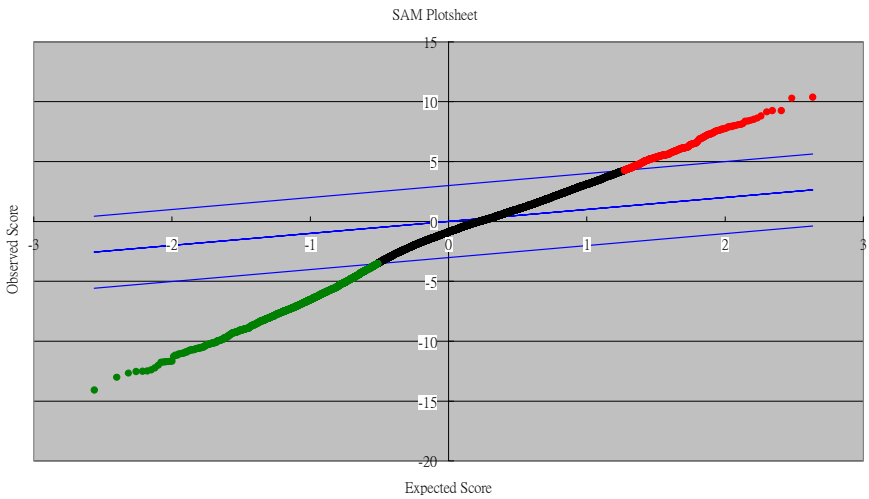

圖 6：部分 SAM 之 $d(i)$ vs. $d_E(i)$ 散佈圖

表 6：SAM 分析結果（部份）

List of Significant Genes for Delta = 3

Positive genes (1295)

Row	Gene ID	Gene Name	Score(d)	Numera	Denominato	Fold Change	q-value	localfdr(%)
9249	AA490471	9248	11.68782239	7929.85	678.471124	-14.39202537	0	0.00618
990	AA044829	989	10.89221312	21239.1	1949.93178	-41.13201974	0	0.05313
3335	AA232837	3334	10.65158735	2309.83	216.853125	-2.35542618	0	0.06679
9465	AA496741	9464	10.34652123	3194.67	308.767549	-9.347879281	0	0.08293
2441	AA155913	2440	10.24891263	15161.8	1479.35986	65.78418633	0	0.08765
8758	AA486362	8757	10.11791467	7443.35	735.660483	-18.18078571	0	0.09349
8207	AA476918	8206	9.983305843	18902.7	1893.43393	-10.37560848	0	0.09873
1474	AA082747	1473	9.637078379	4599.03	477.222434	3.042728346	0	0.10741
2461	AA156571	2460	9.574580254	4453.19	465.105507	2.088080434	0	0.10801
3362	AA233805	3361	9.534615191	981.73	102.964827	-1.497639757	0	0.1082
5126	AA421258	5125	9.485455034	1667.21	175.764894	-4.076386858	0	0.10821
####	AA598653	10309	9.415108893	13221.5	1404.28859	68.6539458	0	0.10779
6107	AA442984	6106	9.17529274	4425.43	482.320306	-44.4454479	0	0.1018
8646	AA485739	8645	8.960838471	10783.1	1203.35279	-51.21086673	0	0.08924
4435	AA402920	4434	8.857941956	10553.3	1191.39751	-54.64247389	0	0.08029
733	AA031287	732	8.85792581	4150.83	468.600222	13.18695136	0	0.08028
7931	AA464246	7930	8.854319481	11811	1333.92408	3.847989778	0	0.07993
2516	AA158584	2515	8.72843308	6137.33	703.142241	3.928536536	0	0.0659
7034	AA455026	7033	8.647998638	2550.51	294.924885	11.41879499	0	0.05505
8783	AA486532	8782	8.580556798	7848.83	914.72269	2.332697647	0	0.04473
898	AA041382	897	8.552551284	4799.59	561.188099	-34.76479947	0	0.0401

Step 1 (Imputed) | SAM Work (Do not edit) | SAM Sample Size Plotsheet | SAM Plot | SAM Output

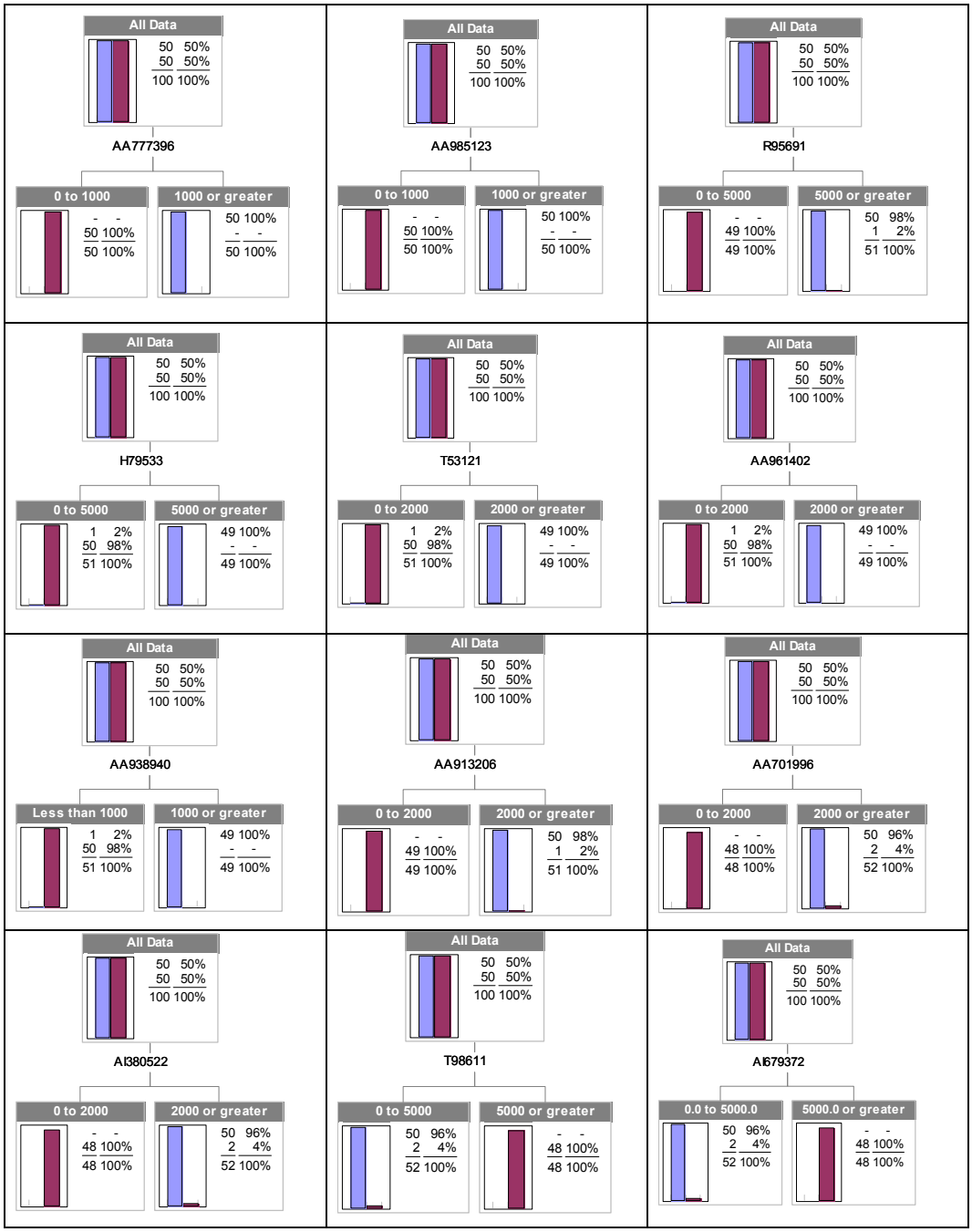
All Data
50 50%
50 50%
100 100%
AA777396
0 to 1000
1000 or greater
- -
50 100%
50 100%
50 100%
- -
50 100%
All Data
50 50%
50 50%
100 100%
AA985123
0 to 1000
1000 or greater
- -
50 100%
50 100%
50 100%
- -
50 100%
All Data
50 50%
50 50%
100 100%
R95691
0 to 5000
5000 or greater
- -
49 100%
49 100%
50 98%
1 2%
51 100%
All Data
50 50%
50 50%
100 100%
H79533
0 to 5000
5000 or greater
1 2%
50 98%
51 100%
49 100%
- -
49 100%
All Data
50 50%
50 50%
100 100%
T53121
0 to 2000
2000 or greater
1 2%
50 98%
51 100%
49 100%
- -
49 100%
All Data
50 50%
50 50%
100 100%
AA961402
0 to 2000
2000 or greater
1 2%
50 98%
51 100%
49 100%
- -
49 100%
All Data
50 50%
50 50%
100 100%
AA938940
Less than 1000
1000 or greater
1 2%
50 98%
51 100%
49 100%
- -
49 100%
All Data
50 50%
50 50%
100 100%
AA913206
0 to 2000
2000 or greater
- -
49 100%
49 100%
50 98%
1 2%
51 100%
All Data
50 50%
50 50%
100 100%
AA701996
0 to 2000
2000 or greater
- -
48 100%
48 100%
50 96%
2 4%
52 100%
All Data
50 50%
50 50%
100 100%
AI380522
0 to 2000
2000 or greater
- -
48 100%
48 100%
50 96%
2 4%
52 100%
All Data
50 50%
50 50%
100 100%
T98611
0 to 5000
5000 or greater
50 96%
2 4%
52 100%
- -
48 100%
48 100%
All Data
50 50%
50 50%
100 100%
AI679372
0.0 to 5000.0
5000.0 or greater
50 96%
2 4%
52 100%
- -
48 100%
48 100%

圖 7：決策樹規則（部份）

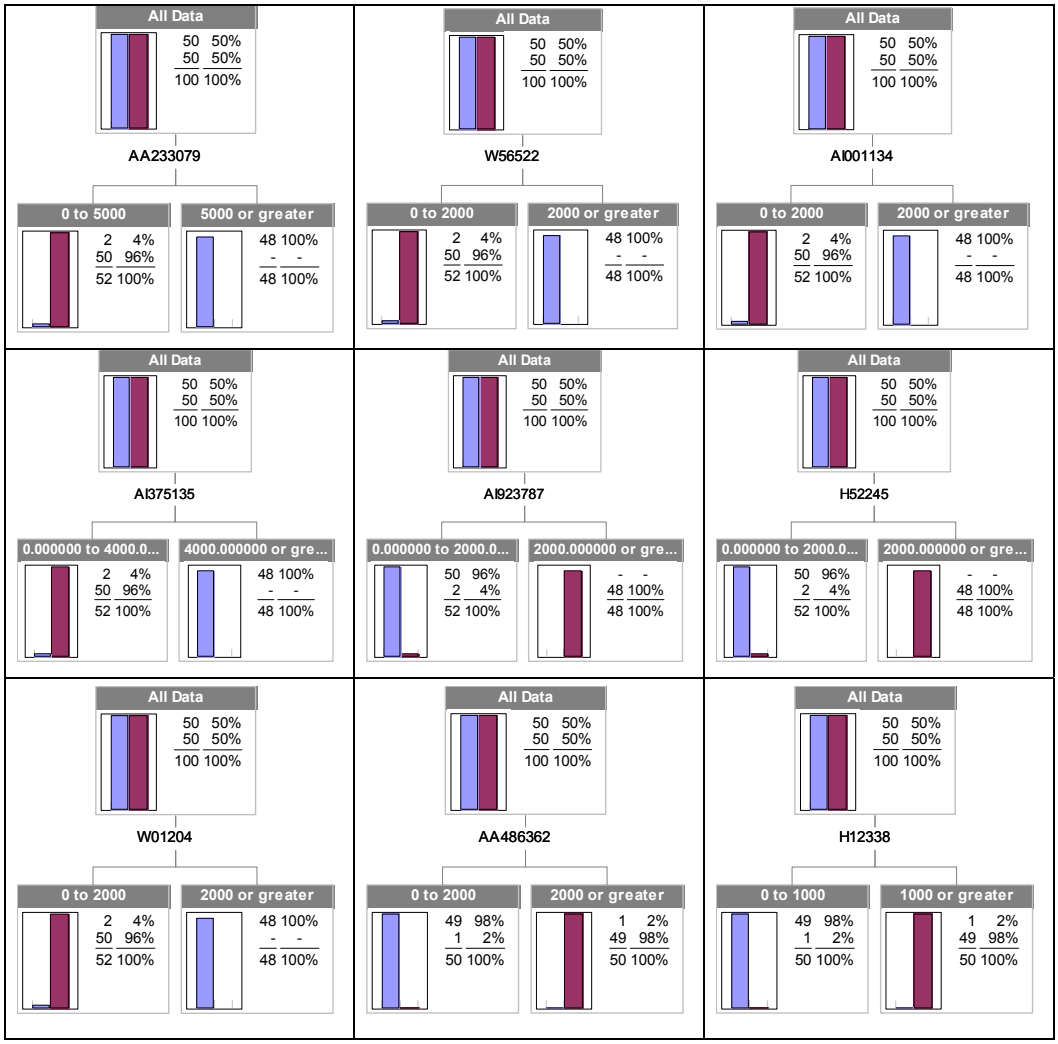

圖 7：決策樹規則（部份）(續)

步驟二：在此由於變數屬性為連續變數，利用決策樹分析，使用 F 分配檢定為分支法則，接著從步驟一挖掘出的較顯著基因當做乳癌決策樹的分類子（classifier）進行決策樹分析後，五次重複抽樣後分別得到 12、14、18、14、16 個分支，其解釋率達到 90%以上之決策樹，聯集共得到 21 個分支解釋率 90%以上之決策樹(如圖 7)與顯著影響乳癌的基因及其規則，本文彙整 IF THEN 規則及其分支正確率（判定為乳癌病人之分支）、平均正確率及模式解釋力等資訊，其中分支解釋率以分數型式表示能更清楚顯示分支情形，括弧內之數字為該決策規則在五次重複抽樣分析中出現之次數，如 50/50(5)為此規則 50 人為判定乳癌患者，實際患者亦為 50 人，此規則在五次分析中出現五次；整體正確率為計算所有正確判別情形；平均模式解釋力為單一規則在各次模式解釋力的平均表現，詳如表 7：

Rule 1: IF (AA777396 < 1,000) THEN patients （若基因 AA777396 < 1,000，則判定為患有乳癌）
Rule 2: IF (AA985123<1,000) THEN patients （若基因 AA985123<1,000，則判定為患有乳癌）
Rule 3: IF (AA961402<2,000) THEN patients （若基因 AA961402<2,000，則判定為患有乳癌）

表 7：決策樹規則整理表（部分）

項次	決策規則	分支正確分類率(次數)	平均整體正確率	平均模式解釋力
1	IF (AA777396 < 1,000) THEN patients	50/50(5)	1.00	100%
2	IF (AA985123 < 1,000) THEN patients	50/50(5)	1.00	100%
3	IF (R95691 < 5,000) THEN patients	49/49(4) 50/50(1)	0.99	97%
4	IF (H79533 < 5,000) THEN patients	50/51(5)	0.99	96%
5	IF (T53121 < 2,000) THEN patients	50/51(5)	0.99	96%
6	IF (AA961402 < 2,000) THEN patients	50/51(1)	0.99	96%
7	IF (AA938940 < 1,000) THEN patients	50/51(1)	0.99	96%
8	IF (AA913206 < 5,000) THEN patients	49/49(2) 49/50(2)	0.99	94%
9	IF (AA701996 < 2,000) THEN patients	48/48(4) 49/49(1)	0.98	93%
10	IF (AI380522 < 2,000) THEN patients	48/48(4) 49/49(1)	0.98	93%
11	IF (T98611 > 5,000) THEN patients	48/48(5)	0.98	92%
12	IF (AI679372 > 5,000) THEN patients	48/48(1)	0.98	92%
13	IF (AA233079 < 5,000) THEN patients	50/52(3)	0.98	92%
14	IF (W56522 < 2,000) THEN patients	50/52(5)	0.98	92%
15	IF (AI001134 < 2,000) THEN patients	50/52(4) 48/48(1)	0.98	92%
16	IF (AI375135 < 4,000) THEN patients	50/52(2)	0.98	92%
17	IF (AI923787 > 2,000) THEN patients	49/50(4) 48/48(1)	0.98(5)	92%
18	IF (H52245 > 2,000) THEN patients	48/48(2)	0.98(2)	92%
19	IF (W01204 < 2,000) THEN patients	50/52(3)	0.98(3)	92%
20	IF (AA486362> 2,000) THEN patients	49/50(1)	0.98	92%
21	IF (H12338> 1,000) THEN patients	49/50(1)	0.98	92%

四、模式驗證

最後將各次重複抽樣所剩餘的 28 筆資料當作測試集進行模式規則驗證，在醫學上偽陰性（false negative），即未將病人檢驗出來比偽陽性（false positive）顯得重要，根據生物晶片與生物資訊領域知識，在模式驗證時，若偽陰性高於 10％則該規則予以刪除，若偽陽性高於 20％時與以刪除。表 8 為將測試集資料分別帶入各次分析中所挖掘出的決策規則中，Category 為正常人/病人(0/1)之分類，藉以驗證，紅色表決策規則判斷錯誤者，黃色表遺失值。從表 8-1 中可看出決策規則僅規則 6:IF (AA961402<2,000) THEN patients，在正常人中誤判三人（3/14），偽陰性為 21.4％，應將此規則刪除；而其餘 11 條決策規則均沒有錯誤分類的狀況，正確率 100％，在 336 筆檢測值當中，正確率高達 99.1％；在表 8-2 中，規則 8：IF (AA913206 < 5,000) THEN patients，在正常人中誤判二人（2/14），偽陰性為 14.3％，應予以刪除，其餘 13 條規則僅錯判 2 筆，在 390 筆檢測值當中，正確率高達 98.0％；在表 8-3 中，規則 5：IF (T53121 < 2,000) THEN patients、規則 7：IF (AA938940 < 1,000) THEN patients、規則 8：IF (AA913206 < 5,000) THEN patients 等三條規則，偽陰性(false negative)超過 10％，應予以刪除，在 488 筆檢測值當中，正確率高達 97.1％；在表 8-4 中，均無規則偽陰性(false negative)超過 10％，在 390 筆檢測值當中，正確率高達 99.2％；在表 8-5 中，IF (AA913206 < 5,000) THEN patients，在正常人中誤判三人，已於前次驗證時刪除，在 430 筆檢測值當中，正確率高達 98.8％，五次整體平均正確率為 98.62%(詳如表 8)。

綜上所述，在表 7 所建立的 21 條決策規則，在經過驗證之後刪除項次 5、6、7、8，IF (T53121 < 2,000) THEN patients、IF (AA961402 < 2,000) THEN patients 、IF (AA938940 < 1,000) THEN patients、IF (AA913206 < 5,000) THEN patients 等四條規則，五次驗證結果正確率均達 97％以上。

表 8-1：決策規則驗證結果(第一次)

AA777396 <1000	AA985123 <1000	AA961402 <2000	H79533 <5000	T53121 <2000	R95691 <5000	AA701996 <2000	T98611 >5000	AI380522 <2000	W56522 <2000	AI001134 <2000	AI923787 >2000	Category
6646	3270	6919	19470	7150	10679	9590	462	5181	21011	5169	194	0
3635	6582	559	11294	8057	21355	9877	280	5688	19294	8149	199	0
7611	3548	1298	12609	12873	18847	10252	390	4537	18459	3109	274	0
6375	5971	2626	11609	12154	34794	9945	377	4028	4844	6719	325	0
8428	13564	3968	23179	16591	45387	20716	515	7210	16203	10294	558	0
4311	8384	4765	29194	17105	37955	22682	965	11827	26230	9488	412	0
4678	4479	2901	14241	7664	27040	12798	649	7653	12416	8884	243	0
5235	7010	661	16918	6467	38116	25514	638	11762	20313	8371	776	0
1638	6245	1822	12946	4927	34315	17677	345	13167	15963	6345	286	0
1255	5606	561	13747	7625	29581	15286	317	10496	16971	6195	221	0
1129	5952	1189	11276	8080	24317	17162	358	10774	13379	7049	276	0
3044	5519	3754	15935	14117	43556	14981	718	19371	21673	7015	682	0
5075	1799	8528	33827	2861	24515	3565	297	10760	25192	3157	264	0
2723	2033	10423	36631	3197	20740	5154	360	9620	25991	3554	329	0
157	131	354	355	410	385	291	32638	433	362	248	3957	1
-164	-27	-25	800	177	727	103	18508	466	241	-34	11488	1
366	36	100	291	432	800	279	29981	128	240	4	7377	1
-34	79	7	86	306	1360	149	36920	25	15	51	7616	1
-391	-147	-80	40	325	1943	54	40426	302	31	-149	21041	1
43	-50	-22	323	216	791	121	21914	268	56	40	7864	1
36	-5	306	174	190	1038	202	24192	79	19	88	4853	1
-78	20	-26	221	145	1544	389	40903	299	83	11	22389	1
-165	-10	-24	14	-145	1565	92	30584	208	52	-35	9050	1
-40	-28	12	133	59	989	122	19105	342	109	27	8084	1
-105	256	-88	212	162	876	422	21235	166	92	158	6685	1
-13	239	365	589	631	1390	553	37699	1074	236	415	11384	1
470	655	835	2722	457	1802	269	14051	864	694	785	6378	1
221	709	818	3309	396	1297	439	11675	735	742	557	10540	2

表 8-2：決策規則驗證結果(第二次)

AA777396 <1000	AA985123 <1000	H79533 <5000	T53121 <2000	R95691 <5000	AA913206 <5000	AA701996 <2000	T98611> 5000	A1380522< 2000	AA233079 <5000	W56522< 2000	A1001134 <2000	W01204< 2000	A1923787 >2000	category
4027	2170	17514	3521	10720	3675	6718	648	3495	18880	6100	2246	4163		0
4778	3706	21739	9178	21024	9435	19280	306	11679	10742	12956	4649	4717	264	0
6359	5013	21528	17166	48064	6662	15675	694	11059	21758	19626	5317	4044	959	0
5275	2201	21123	8934	24946	9246	15136	689	7885	30122	14549	6647	5317	479	0
4395	1561	8858	2212	12427	1918	8482	338	4301	3963	6348	2281	1597	323	0
5719	3985	26763	11701	34569	12050	17912	526	15987	31683	16493	6708	9523	258	0
5129	2129	14009	6435	18864	7483	6041	310	9428	25789	14229	3744	4540	224	0
3266	3503	9302	6071	13567	7000	6073	290	8595	9692	8950	5106	5346	94	0
3635	6582	11294	8057	21355	9878	9877	280	5688	22425	19294	8149	6511	199	0
6375	5971	11609	12154	34794	12759	9945	377	4028	21827	4844	6719	9474	325	0
8428	13564	23179	16591	45387	15232	20716	515	7210	45318	16203	10294	11482	558	0
4678	4479	14241	7664	27040	11581	12798	649	7653	28371	12416	8884	9300	243	0
1129	5952	11276	8080	24317	8460	17162	358	10774	15460	13379	7049	7784	276	0
5075	1799	33827	2861	24515	9394	3565	297	10760	27529	25192	3157	7787	264	0
216	201	498	435	1054	193	490	23113	148	1413	174	211	293		1
262	367	1187	800	2620	720	1091	45761	1039	1062	735	462	372	16380	1
401	581	1189	1261	2406	4160	1048	39308	803	1577	736	653	461	15037	1
239	148	761	807	1332	518	428	41833	384	962	562	526	197	21147	1
338	373	483	466	739	367	625	20485	389	621	241	331	187	5423	1
91	228	270	441	1886	404	410	52435	744	749	318	196	778	22668	1
236	92	337	457	1396	573	891	16406	430	1262	316	177	696	5179	1
55	56	131	165	518	208	230	21178	292	265	127	84	197	7757	1
-164	-27	800	177	727	248	103	18508	466	486	241	-34	145	11488	1
-34	79	86	306	1360	304	149	36920	25	245	15	51	178	7616	1
-391	-147	40	325	1943	146	54	40426	302	546	31	-149	175	21041	1
36	-5	174	190	1038	340	202	24192	79	1085	19	88	228	4853	1
-105	256	212	162	876	346	422	21235	166	669	92	158	320	6685	1
470	655	2722	457	1802	1167	269	14051	864	2771	694	785	995	6378	1

表 8-3：決策規則驗證結果(第三次)

AA777396 <1000	R95691 <5000	AA985123 <1000	H79533 <5000	T53121 <2000	AA938940 <1000	AA701996 <2000	A1380522 <2000	T98611> 5000	A879372> 5000	A1001134 <2000	H52245<2 000	W56522< 2000	A1375135 <4000	W01204 <2000	AA913206 <5000	AA486362 >2000	A1923787 >2000	category
2373	17018	1642	26305	2354	3593	5422	3622	301		3704		5904		3733	4518	682		0
1118	19972	1217	17659	2052	1859	5305	3629	185		2375		3695		2098	2985	503		0
6709	39173	3238	34259	12088	6050	16188	9922	379	455	6005	453	12262	13927	6768	12729	1064	646	0
8634	48311	4751	30709	17916	9379	23200	10818	685	546	7726	465	30041	21864	9228	17961	864	760	0
12718	40623	3588	31964	13997	5965	16060	13020	468	498	7667	633	12593	16493	7006	12256	1184	589	0
3006	17644	1802	5819	4173	607	8737	5642	327	312	3029	376	4635	6161	4075	6495	776	216	0
3849	20787	3832	8974	4728	686	17422	9430	548	375	6416	313	11553	9191	6154	9685	725	481	0
7017	25445	2461	20275	9679	4501	9364	10286	277	470	6997	361	18666	15719	5398	6690	1836	329	0
1445	7918	1694	6358	5334	2447	5349	5628	245	145	2962	97	6459	5302	1423	3738	167	52	0
3266	13567	3503	9302	6071	1957	6073	8595	290	162	5106	107	8950	12464	5346	7000	218	94	0
2342	11088	2209	8383	6774	1631	5933	6292	266	183	4518	104	5215	7628	3000	4981	233	78	0
1928	14657	1651	9774	4749	1862	4862	6694	184	127	4065	92	9217	7669	3529	4807	3	84	0
1638	34315	6245	12946	4927	4340	17677	13167	345	341	6345	166	15963	13382	8732	10258	725	286	0
2723	20740	2033	36631	3197	4642	5154	9620	360	153	3554	126	25991	10469	9105	10394	625	329	0
103	897	53	344	191	152	111	462	13004		208		135		127	121	10945		1
47	533	48	118	59	74	79	143	5323		155		45		45	77	3659		1
545	2505	566	2460	1713	666	1330	890	29151	27764	874	11633	565	1425	713	1527	19736	16139	1
819	2763	703	2105	1659	1073	2204	625	42398	46909	1023	24198	1167	2023	1056	2090	20883	34832	1
875	11469	584	1631	1748	634	1191	1124	36184	34461	1015	13215	697	1986	649	1253	12252	14972	1
260	907	275	478	405	153	667	413	14756	25153	575	4728	303	526	122	807	9903	4571	1
230	2099	938	508	824	127	743	1268	30951	21233	1040	3760	342	1011	597	925	14728	6139	1
857	1125	649	1376	1457	460	1093	2024	18445	31060	1509	8902	852	1423	686	1155	5680	11554	1
30	221	19	74	90	35	119	243	14978	7269	51	3525	65	60	66	134	3062	4254	1
55	518	56	131	165	38	230	292	21178	15747	84	3908	127	237	197	208	3257	7757	1
68	344	49	146	274	43	166	279	8155	6163	124	4609	99	158	87	198	3436	5680	1
42	1144	21	128	172	44	144	247	15469	9143	68	2504	103	148	115	163	44	5062	1
-165	1565	-10	14	-145	-3	92	208	30584	29452	-35	4394	52	360	77	120	3611	9050	1
221	1297	709	3309	396	532	439	735	11675	4252	557	2450	742	1311	1027	1323	2870	10540	1

表 8-4：決策規則驗證結果(第四次)

AA777396 <1000	AA985123 <1000	H79533 <5000	T53121 <2000	R95691 <5000	AA913206 <5000	AA701996 <2000	T98611 >5000	A1380522 <2000	A1923787> 2000	AA233079 <5000	W56522< 2000	A1001134 <2000	W01204< 2000	category
4027	2170	17514	3521	10720	3675	6718	648	3495		18880	6100	2246	4163	0
4778	3706	21739	9178	21024	9435	19280	306	11679	264	10742	12956	4649	4717	0
6359	5013	21528	17166	48064	6662	15675	694	11059	959	21758	19626	5317	4044	0
5275	2201	21123	8934	24946	9246	15136	689	7885	479	30122	14549	6647	5317	0
4395	1561	8858	2212	12427	1918	8482	338	4301	323	3963	6348	2281	1597	0
5719	3985	26763	11701	34569	12050	17912	526	15987	258	31683	16493	6708	9523	0
5129	2129	14009	6435	18864	7483	6041	310	9428	224	25789	14229	3744	4540	0
3266	3503	9302	6071	13567	7000	6073	290	8595	94	9692	8950	5106	5346	0
3635	6582	11294	8057	21355	9878	9877	280	5688	199	22425	19294	8149	6511	0
6375	5971	11609	12154	34794	12759	9945	377	4028	325	21827	4844	6719	9474	0
8428	13564	23179	16591	45387	15232	20716	515	7210	558	45318	16203	10294	11482	0
4678	4479	14241	7664	27040	11581	12798	649	7653	243	28371	12416	8884	9300	0
1129	5952	11276	8080	24317	8460	17162	358	10774	276	15460	13379	7049	7784	0
5075	1799	33827	2861	24515	9394	3565	297	10760	264	27529	25192	3157	7787	0
216	201	498	435	1054	193	490	23113	148		1413	174	211	293	1
262	367	1187	800	2620	720	1091	45761	1039	16380	1062	735	462	372	1
401	581	1189	1261	2406	4160	1048	39308	803	15037	1577	736	653	461	1
239	148	761	807	1332	518	428	41833	384	21147	962	562	526	197	1
338	373	483	466	739	367	625	20485	389	5423	621	241	331	187	1
91	228	270	441	1886	404	410	52435	744	22668	749	318	196	778	1
236	92	337	457	1396	573	891	16406	430	5179	1262	316	177	696	1
55	56	131	165	518	208	230	21178	292	7757	265	127	84	197	1
-164	-27	800	177	727	248	103	18508	466	11488	486	241	-34	145	1
-34	79	86	306	1360	304	149	36920	25	7616	245	15	51	178	1
-391	-147	40	325	1943	146	54	40426	302	21041	546	31	-149	175	1
36	-5	174	190	1038	340	202	24192	79	4853	1085	19	88	228	1
-105	256	212	162	876	346	422	21235	166	6685	669	92	158	320	1
470	655	2722	457	1802	1167	269	14051	864	6378	2771	694	785	995	1

表 8-5：決策規則驗證結果(第五次)

AA777396 <1000	AA985123 <1000	H79533 <5000	T53121 <2000	R95691 <5000	AA701996 <2000	T98611> 5000	AI380522 <2000	H52245 >2000	AA233079 <5000	AI375135 <4000	W56522< 2000	AI001134 <2000	AA913206 <5000	H12338 >1000	AI923787> 2000	category
2373	1642	26305	2354	17018	5422	301	3622		20258		5924	3704	4518	147		0
4027	2170	17514	3521	10720	6718	648	3495		18880		6100	2246	3675	315		0
1283	1786	7975	2112	16872	2913	160	2695		2248		4008	2585	3493	76		0
4778	3706	21739	9178	21024	19280	306	11679	408	10742	22548	12956	4649	9435	369	264	0
3757	1941	16237	5184	21823	12344	425	4952	301	19078	5592	6817	3777	7424	180	419	0
6359	5013	21528	17166	48064	15675	694	11059	530	21758	19269	19626	5317	6662	283	959	0
3849	3832	8974	4728	20787	17422	548	9430	313	6935	9191	11553	6416	9685	593	481	0
3266	3503	9302	6071	13567	6073	290	8595	107	9692	12464	8950	5106	7000	44	94	0
4552	4309	18233	16501	22530	11687	282	11842	225	46593	14304	25166	4873	7862	93	154	0
4463	2553	15138	7995	14293	9541	313	5306	111	10139	12417	12414	5770	6793	116	599	0
6646	3270	19470	7150	10679	9590	462	5181	224	10412	13288	21011	5169	6836	365	194	0
4311	8384	29194	17105	37955	22682	965	11827	477	31511	33328	26230	9488	20071	361	412	0
5235	7010	16918	6467	38116	25514	638	11762	411	15114	9161	20313	8371	23356	451	776	0
1129	5952	11276	8080	24317	17162	358	10774	143	15460	12290	13379	7049	8460	198	276	0
103	53	344	191	897	111	13004	462		847		135	208	121	2753		1
216	201	498	435	1054	490	23113	148		1413		174	211	193	1554		1
121	183	460	243	1043	348	12333	82		287		95	218	207	443		1
262	367	1187	800	2620	1091	45761	1039	16633	1062	1285	735	462	720	14227	16380	1
206	196	636	558	686	633	5416	290	4590	839	387	323	406	538	580	8580	1
401	581	1189	1261	2406	1048	39308	803	5391	1577	1177	736	653	4160	1244	15037	1
230	938	508	824	2099	743	30951	1268	3760	932	1011	342	1040	925	3569	6139	1
55	56	131	165	518	230	21178	292	3908	265	237	127	84	208	2015	7757	1
165	72	226	1989	997	252	30226	343	13753	1004	333	274	82	187	3752	13810	1
125	79	318	384	613	446	6305	513	3480	430	361	374	170	574	3186	3050	1
157	131	355	410	385	291	32638	433	5009	627	426	362	248	408	7662	3957	1
43	-50	323	216	791	121	21914	268	12170	3005	703	56	40	435	4874	7864	1
-78	20	221	145	1544	389	40903	299	8402	512	249	83	11	585	10343	22389	1
-105	256	212	162	876	422	21235	166	3125	669	504	92	158	346	2919	6685	1

五、規則解釋與評估

由於醫學研究往往牽涉患者的健康、生命安全，研究模式的解釋能力以及可靠度需以更嚴格的標準衡量，因此本研究選取模式解釋能力 90%以上的 21 個基因為醫療檢測參考因子並建立其個別決策規則，如當基因 AA777396 檢測值小於 1,000 時，則判定為患有乳癌，大於 1,000 時則為正常人；在使用測試集進行驗證分析的時候，採取偽陰性若高於 10%時則該規則予以刪除，偽陽性高於 20%時予以刪除，共刪除四條決策規則（各規則信度效度如上表 7、8 所示）。本研究所建立的決策規則，係純以晶片資料進行分析，由於無法取得相關病歷資料，故無法更深入探討病人基因表現值與相關病歷的關係，如年齡、性別、區域人種等，亦未建立規則間彼此的交互關係。

本研究模式發展生物晶片資料挖礦的架構和方法，可以提供一完整有效的分析架構，從最前端的資料處理開始架構完整的處理步驟、分析模式與結果驗證及彙整決策支援建議。從龐大的資料庫當中，首先必須瞭解資料型態；再進行資料預處理工作，去除雜亂無用的屬性及資料與遺失值的填補並彙整成可用分析的資料集；接著即從降低變數維度的角度，並且採用階段式的步驟進行分析，根據本研究提出的生物晶片資料挖礦模式進行影響特定疾病基因之挖掘，並以史丹佛大學基因晶片資料庫之乳癌實驗晶片 cDNA 資料進行研究驗證，從 41,681 個基因的 64 筆樣本中分階段降低變數維度，到最後歸納出正確率極高之規則，已經證明本方法的效度。

在分析的過程中發現，雖然原始資料為電腦掃描判讀產生，但仍高達近 10%的無效資料無法進行處理及分析而必須捨棄。本研究從原始資料的解構、整合到資料整理，再採用階段性步驟進行研究分析，最後建立出 17 條各別決策規則，可供生醫研究者進行後續分析或決策參考應用。

伍、結論

資料挖礦是一種不斷循環的資料分析與決策支援過程，從大量資料中探索和分析，以發現出有意義的樣型或規則，並將其整理成有價值的知識。資料挖礦方法已經成功的應用在許多領域，而應用在生命科學和醫學領域尤其是基因研究卻是一個嶄新的方向。隨著生命科學的知識及技術的快速發展，生物資訊發現所累積的大量資料已不是傳統統計技術可以解決的了，從生物晶片探索基因的影響為例來說，生物晶片一次就能夠紀錄成千上萬個基因表現的樣型，是一個具有極大變數量卻僅有很少的樣本數的資料處理問題。在傳統統計假設上，即因自由度的關係而無法進行實驗設計，亦難以處理複雜交互作用的情形下的分析。本研究所提出之生物晶片資料挖礦架構提供一個有效的方法，由乳癌實驗晶片 cDNA 資料分析結果也可以驗證其效度。

然而，本研究尚未涵蓋人類基因功能網路架構(regulatory network)，因此並未進一步建立規則間的交互關係，且所採用的決策樹分支閥值的方法是等分區段的方式，未來研究可以整合跨領域的專家以進行更深入的研究，包括閥值的敏感度分析以建立更精密的決策規則等，亦可以結合晶片製造領域及生醫領域專業知識，更深入的考量因設備或晶片或實驗所造成的影響，並取得本土生醫資料進行實驗和實證研究，以及對於挖掘出的規則或樣型進一步整合領域專家進行解釋、分析與醫檢決策甚而發展視覺化生物晶片資料挖礦之決策支援系統。

陸、致謝

本研究承蒙國科會專題研究計畫(NSC93-2213-E-007-033; NSC94-2213-E-007-049)與國際合作研究計畫之經費補助，以及史丹佛大學晶片資料庫中乳癌晶片資料之協助，特此致謝；作者並感謝兩位匿名審查委員及 Professor Francois Sainfort、吳欣怡博士、許嘉裕等先進的寶貴建議與指教。

柒、參考文獻

1. 王鴻儒、簡禎富、李培瑞、徐紹鐘，2002『決策樹資料挖礦架構及其於半導體製程之實證研究』，科技管理學刊，第七卷．第一期：137~160 頁。
2. 何國傑、葉開溫、鄭石通、靳宗洛，2001，基因工程與生物技術概論-基因選殖及 DNA 分析，台北：藝軒圖書出版社。
3. 江晃榮，2003，經濟巨人 Bio-生物科技的千億商機，台北：世茂出版社。
4. 耿直、鄔宏潘、謝邦昌、趙雅婷、蘇志雄，2003，生物醫學統計學-理論與資料分析應用，台北：鼎茂圖書出版有限公司。

5. 簡禎富、吳文婷，1997『醫療決策分析：以唐氏症之診斷為例』，醫療資訊雜誌，第六期：39~53 頁。
6. 簡禎富、林鼎浩、徐紹鐘、彭誠湧，2001『建構半導體晶圓允收測試資料挖礦架構及其實證研究』，工業工程學刊，第十八卷．第四期：37~48 頁。
7. 簡禎富、李培瑞、彭誠湧，2003『半導體製程資料特徵萃取與資料挖礦之研究』，資訊管理學報，第十卷．第一期：63~84 頁。
8. 簡禎富、林鼎浩、劉巧雯、彭誠湧、徐紹鐘、黃佳琪，2001『建構晶圓圖分類之資料挖礦方法及其實證研究』，工業工程學刊， 第十九卷．第二期：23~38 頁。
9. 簡禎富、蕭禮明、王興仁，2004『建構半導體製造管理目標層級架構與製造資料之資料挖礦』，工業工程學刊，第廿一卷．第四期：313~327 頁。
10. 簡禎富、王興仁、陳麗妃，2005『利用資料挖礦提升半導體廠製造技術員人力資源管理品質』，品質學報，第十二卷．第一期：9~28 頁。
11. Baldi, P., and Brunak, S. *Bioinformatics: The Machine Learning Approach*, The MIT Press, London, 2004
12. Bergeron, B. *Bioinformatics Computing,* Prentice Hall, New Jersey, 2002
13. Breiman, L., Friedman, J. H., Olshen, R. J., and Stone, C. J. *Classification and Regression Tree*, Wadsworth, Belmont, CA, 1984
14. Chen, X., "Gene Selection for Cancer Classification Using Bootstrapped Genetic Algorithms and Support Vector Machines," *Proceedings of the 2003 IEEE Bioinformatics Conference* 2003, pp: 504-505
15. Chien, C.F., Chen, S., and Lin, Y., "Using Bayesian Network for Fault Location on Distribution Feeder of Electrical Power Delivery Systems," *IEEE Transactions on Power Delivery*, (17:13) 2002, pp: 785-793
16. Chien, C.F., Wang, W.C., and Cheng, J.C., "Data mining for yield enhancement in semiconductor manufacturing and an empirical study," *Expert Systems with Applications*, (33:1) 2007, pp: 1-7
17. Chuang, H., Liu, H., Brown, S., McMunn-Coffran, C., Kao, C., and Hsu, F. "Identifying Significant Genes from Microarray Data," *Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering* 2004, pp: 358-365
18. Fayyad, U., "Data mining and knowledge discovery in database: implication for scientific databases," *Scientific and Statistical Database Management* 1997, pp: 2-11
19. Gregory, P., and Tamayo, P., "Microarray Data Mining: Facing the Challenges," *Knowledge Discovery and Data Mining*, (5:2) 2003, pp: 1-5
20. Han, J. and Kamber, M. *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 2000
21. Hartigan, J. A. *Clustering Algorithms*, John Wiley & Sons, New York, 1975
22. Khabzaoui, M., Dhaenens, C., and Talbi, E., "A Multicriteria Genetic Algorithm to analyze DNA microarray data," *Proceedings of the Fourth IEEE Symposium on*

Bioinformatics and Bioengineering 2004, pp: 1874-1881

23. Kass, G. V., "An exploratory technique for investigation large quantities of categorical data," *Applied Statistics*, (29) 1980, pp: 119-127
24. Kristina, A. K., and Salter, L. A., "A comparison of methods for estimating the transition: transversion ration from DNA sequences," *Molecular Phylogenetics and Evolution*, (32) 2004, pp: 495-503
25. Lesk, A. M. *Introduction to Bioinformatics*, Oxford University Press, New York, 2002
26. Miyamoto, T., Uchimura, S., Yoshihiko, H., Iizuka, N., Oka, M., and Yamada-Okabe, Y., "Comparative study of feature selection methods on microarray data," *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering* 2003, pp: 82-83
27. Mukherjee, S. N., "Gene ranking using bootstrapped P-value," *Knowledge Discovery and Data Mining*, (5:2) 2003, pp: 16-22
28. Nadimpally, V., and Zaki, M., "A Novel Approach to Determine Normal Variation in Gene Expression Data," *Knowledge Discovery and Data Mining*, (5:2) 2003, pp: 6-11
29. Peng, C., and Chien, C., "Data Value Development to Enhance Yield and Maintain Competitive Advantage for Semiconductor Manufacturing," *International Journal of Service Technology and Management*, (4:4-6) 2003, pp: 365-383
30. Peng, J., Chien, C., and Tseng, B., "Rough set theory for data mining for fault diagnosis on distribution feeder," *IEE Proceedings-Generation, Transmission, and Distributions*, (151:6) 2004, pp: 689-697
31. Shoemaker, J., Painter, I., and Weir, B., "Bayesian statistics in genetics," *Trends in Genetics*, (15:9) 1999, pp: 354-358
32. Simon, R., "Supervised analysis when the number of candidate feature (p) greatly exceeds the number of cases (n)," *Knowledge Discovery and Data Mining*, (5:2) 2003, pp: 31-36
33. SØrlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., LØnning, P., and BØrresen-Dale, A. , "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Science of the United States of America*, (98:19) 2001, pp: 10869–10874
34. Troyanskaya, O., Dolinski, K., Owen, A., Altman, R., and Botstein, D., "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Science of the United States of America*, (100:14) 2003, pp: 8348–8353
35. Tusher, V., Tibshirani, R., and Chu, G., "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Science*

of the United States of America, (98:9) 2001, pp: 5116–5121

36. Wu, Y., and Zhang, A., "Feature Selection for Classifying High-Dimensional Numerical Data," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2004, pp: 251-258
37. Yun, H., Ha, D., Hwang, B., and Ryu, K., "Mining association rules on significant rare data using relative support," *The Journal of System and Software*, (67) 2003, pp: 181-191