藉由決策規則的類神經模糊邏輯演算法精鍊人類新穎基因資料庫序列

洪俊銘
崑山科技大學資訊管理系
助理教授
cmhung@mail.ksu.edu.tw

林運輝
崑山科技大學資訊管理所
研究生
hui731229@gmail.com

陳秉祥
崑山科技大學資訊管理所
研究生
bixictn@gmail.com

摘要

人類細胞中基因表達的方式可以分成本質性、暫時性、和誘導性三種不同類別。其中暫時性、誘導性的基因表達是屬於條件式，其轉錄序列很難用EST方法來發現它的存在。然而暫時性表達基因對胚胎發育和疾病的發病來說是非常關鍵的因素，因為他們決定了疾病的結果。本文主要目的是要在暫時性的表達基因中找到更為精準新穎外顯子。因此，從我們之前所建構的新穎基因資料庫中，透過我們所提出的基於決策規則的類神經模糊邏輯演算法，精鍊出符合個別研究目的所需的候選基因序列，這些候選序列具有量少質精的特性，可以供生物學家進行生物實驗之前的輔助參考資訊。

關鍵詞：基因體、新穎基因、轉錄本、基因表達資料庫

Abstract

The expression of genes in mammalian cells can be constitutive, transient, or inducible. Transcripts of transient and inducible genes are difficult to discover using the EST approach. Transiently expressed genes, however, are crucial to embryo development and the pathogenesis of disease because they determine the outcome of disease. Using our new bioinformatics approach, which we believe will facilitate verification of novel transcripts in developing embryos or pathogen-induced cells; we aimed to identify novel exons in transiently expressed genes.

First of all, the proposed method will use a general gene predictor that must be able to produce all possibly optimal or suboptimal candidate exons in human. After applying signal processing, an anchoring procedure in the method will transform and group the candidate sequences into many numeric hashing-signals clusters rapidly. In the meanwhile, an entropy-based theorem in the method will be used to remove the most error matches, repeat matches. Finally, the method will generate the resulting exons identified by alignment with other genomic or EST sequence in cross-species.

Finally, we will be expected to find out thousands of potential novel exons. The a priori results prove the feasibility of the method. Combining the anchoring method embedded an entropy-based filter with an inherently unreliable gene predictor will be used to obtain a small scope of exons that may be potentially novel because the combination may avoid many drawbacks of some traditional gene predictors.

Keywords: genomics, novel gene, transcript, Gene Expression Omnibus (GEO)

1. 前言

在後基因體時代對於新穎基因的預測，可以減少實際實驗成本的支出。其中胚胎時期條件式新穎基因的發現可以用於發展與遺傳有關的基因治療藥物，依美國藥物研究與生產協會(PhRMA)統計，新藥研發所需耗費成本約達5億美元，從先期研發到藥物通過上市許可，所需時間達12-15年之久。為能縮短新藥開發所需時程，加快上市腳步，以達成減少成本並提高研發效能，整合多樣生命科學相關資訊，包括生物資訊學，蛋白質體工程，生物晶片等相關微陣列技術，能為藥物開發流程帶來全新的革命。

由以上可知，有資料探勘技術的幫助，相信對生物資訊技術是一大躍進。我們從先前實驗經驗[41]，發現各個目標分類分布的訓練個案數量可能是影響學習模型的主要原因，而且我們了解到條件式的表達基因應該佔所有的基因一小部分，如果使用傳統的資料探勘方式可能無法找出興趣的樣式，因此根據我們之前成功應用在金融風險資訊上的根據決策規則的類神經模糊邏輯(Neuro-fuzzy Logic with Decision Rules ;NFLDR)演算法，考慮衝突敏感性結構下所挖掘出來的興趣樣式應該也同

樣可以應用於探勘條件式表達的新穎基因上面。另外，由於欲完成本研究必須涉及大量的計算，我們希望能在個人電腦的層次上也能執行此一巨大運算工作，因此必須設計出一個特殊的索引結構來實現叢集雜湊訊號掛錨法(Clustering-Hashing-Signal Anchoring Method; CSAM)演算法。為了以上兩個原因，在本研究中實際發展一個 CSAM+NFLDR 的生物資訊探勘專屬平台，對於找出新穎基因是相當必要的。

此外，預測出來的新穎外顯子是否能在跨物種對應區域中找到對相似性很高的外顯子，則含此未知外顯子的基因很可能是因某種條件下才會表達的新穎基因或是多樣接合型式的基因。此項工作若使用現存一般的序列比較工具來達成，可能必須耗費極大的計算資源。因此，本研究藉由叢集雜湊訊號掛錨法(CSAM)，特定化一般序列比對的問題成為外顯子層次的集合從屬關係的問題。使得兩個大量的序列資料庫能快速的交叉比對，得到一群以外顯子為單位的加值資料庫。

在本研究中，我們以有序雜湊索引為基礎，實現一個以掛錨(Anchoring)為導向的跨物種比對搜尋的新穎基因探勘平台，並透過本實驗室所發展的決策規則下的類神經模糊邏輯學習模型 NFLDR[42]自然計算演算法，根據衝突敏感性結構平衡訓練理論下探勘出來的興趣樣式(interesting patterns)，來建立胚胎時期才會表達的條件式基因訓練集合，藉由這個訓練集合建立一個學習模型並從已經被找出來的候選基因中，預測篩選出真正有可能的新穎基因，本研究所建置的系統(CSAM+NFLDR)主要的目標如下：

(1)**提高整體探勘速度:** 建構大型索引計算(Indexing Computing)SQL 資料庫，可以動態的變化參數進行人類新穎基因的探索以及預測。

(2)**減少人工介入資料探勘程序的程度:** 建構混合式學習模型處理傾斜資料分環境，使得資料清洗步驟能自動化。

(3)**產生量少質精的新穎基因：**產生一組在胚胎時期才會表達新穎基因和它的全長基因序列。

以下各章節內容摘要分述如下。第二節對於研究背景作相關的文獻探討。第三節描述本研究的方法與進行步驟。第四節對本研究之實驗結果。第五節為本研究之結論。

2.研究背景

2.1 生物資訊、資料探勘、與計算分子生物學

自 1987 年 9 月以來，人體基因計劃將解開構成人類生命的藍圖--以美國為主的研究人員開始進行人體 DNA 的排序工作；預估將會有 30 億個鹼基、需要 20 年的時間。但由於電腦軟硬進步相當快速，由英、美、日、德、中國大陸等國出資的人體基因計劃，已提前於 2000 年六月攜手公佈精確率度幾近百分之百的人類基因圖譜。在另一方面，過去廿十多年人工智慧所衍生出來的類神經網路、模糊邏輯及基因演算法等軟式運算也已廣泛地成功運用在各行各業之中，而將大量基因序列資料整理分析加以運用的資料探勘科學也為生命科學領域帶來巨大的影響。

但是如何從大量的序列資料中找出有用的因果關係一直是研究人員努力的目標，也是本研究重要的課題。這方面的研究稱為資料探勘(Data Mining) [34][1][63][52][23][12][56][32][45]，一般採用的方法有 C4.5-like 決策樹[30][31][27][5][60][62]、統計[61][11]、粗糙集(Rough Sets)[47]機器學習(Machine Learning)中的 ID3[20][18]、非參數式回歸(Nonparametric regression)[16]中的 CART(Classification and Regression Trees)[22]等。類神經網路的基本學習方法為逆傳導法(Back-propagation)[59]，更複雜先進的方法則是在統計或非線性迴歸(Nonlinear regression)[57][70]中常用到的 Gauss-Newton Method [72][67][46]或是 Levenberg-Marquardt Method[69][29][37][28][50][51]。但是這些方法都須要用到梯度向量(Gradient vectors)而梯度向量在複雜系統中並不容易計算，因此對於較複雜的大型系統，偏向於使用遺傳演算法(Genetic algorithms)[38]、下坡式 Simplex 搜尋(Downhill simplex search)[55]、雜亂搜尋(Random search) [66][3]、模擬退火法(Simulated annealing) [68][44]等。

隨著資料庫系統應用的日漸增加，近年來資料探勘的重要性也逐漸被重視，各式資料探勘 (data mining) 方法也已經被提出。正如一般所知，資料探勘所處理的資料可以從不同種類的來源獲得，也因此資料的型態可能各不相同。但目前似乎沒有一套資料探勘的方法可以同時適用於所有的應用，因為實際上每種方法都有其適合處理的資料型態。

一般來說，資料探勘服務採取 CRISP-DM process (CRISP：CRoss-Industry Standard Process for Data Mining)，包含以下程序：

(1) 訂定探勘目標及方法。

(2) 資料準備(Data Preparation)。協助資料擷取,資料清理(DataCleaning)，資料轉換(Data Transform)、整合、篩選等準備工作。

(3) 資料探勘(Data Mining)分析執行，依探勘目標選擇最佳的資料探勘核心及探勘方法。如分群分析(Clustering) 、分類分析(Classification)、協銷分析(Association Rule Discovery)、序銷分析 (Sequential Pattern Discovery)等。

(4) 資料探勘(Data Mining)結果產出及解讀。經上述分析後，以報表方式或模型呈現，協助使用單位解讀資料探勘結果,提供企業分析建議。

資料發掘之目的在從大量資料中，尋找其規律性及其中隱藏的知識,期能據以對未來作決策支援之輔助;分類方法即藉著各個資料類別，歸納出各類別的特異性。由於真實資料集本身常由比例不甚平均的分類記錄所組成,隨時間的變化累積而成的交易集合更容易產生資料記錄之間予盾與冗餘;加上常見的已知錯誤和不定量的未知屬性值對訓練集進行干擾,直接探勘這些資料集導出的學習模型其分類預測正確性通常很差。所以過去採用主成分析(Principal Component Analysis)、因子分析（factor analysis）、判別分析（discriminate analysis）、叢集分析（cluster analysis）、典型相關分析（canonical correlation analysis）的多變量分析方法和計算資料分佈的平均數、標準差、標準數(z score)、偏離理論置信區間、分佈的程度、顯著性的差異來計算資料分佈情形後進行訓練集分析、過濾和校正的資料清洗前置作業,然後選用適當的資料探勘工具進行分析和解釋結果，若未發現有用知識，則調整參數後再重新執行這些程序直到發現滿意的答案為止。然而，為數眾多的資料探勘演算法和許多演算法參數必須靠經驗的選擇才能有效地完成目標。雖然，資料探勘技術發展至今對一般性的問題已有良好解決方案，國內外也已有十分優越的軟體供業界使用。但在國內外文獻有關一般分類預測的議題上，多數著重分類預測總體正確率的提昇，未考慮每一類別其分類預測正確率是否均超過 50%,一些不甚明顯的特徵常被邊緣化。尤其，在各類別數目分佈不平均的真實資料集中情況更為嚴重。一般而言,記錄證據之間相關性低時,引用以 Shannon 理論為基礎的決策樹模式及其合併其他演算法的混合模型均可輕易解決此問題;當記錄證據之間相關性非常高時，摻雜 Fuzzy logic 的混合模型則較能適應複雜的環境變化,如神經模糊模型(Neuro-fuzzy models)。然而，記錄證據數目相對較少的類別,為防止被誤以為雜訊而排除，並成為預測模型的一部份，就必須先以資料清洗步驟進一步處理後才能得到較好的結果。如何適當清洗資料才可以保證重要訊息不致遺失，是一個值得研究的課題。

本研究從計算分子生物學的角度去設計一個彈性的計算架構,能夠用來快速瀏覽人類的新穎基因，並且透過資料探勘的方法分析、過濾與預測比較有可能的新穎基因,這些預測出來新穎基因可以提供給生物學家找尋驗證真正新穎基因的一個方向,大幅縮減實驗盲目測試所需要的成本。在這個研究中利用本實驗室提出的 CSAM 演算法[41]來構築一個計算搜尋平台,並在這個平台上發展一個基於模擬自然計算-根據決策規則的類神經模糊邏輯(NFLDR)學習模型進行新穎基因的有效分類預測。我們以胚胎時期暫時表達的基因為目標,設計出的學習訓練機制能對於每一分類項目都能正確及公平提昇其分類預測正確性。我們如何在上述問題、想法及現況中取得一良好的解決方案？我們相信自然計算(Nature computing)可能是一另一種契機。以下將簡略介紹自然計算的內涵：

自然計算是近幾年興起的計算新思維，在 S. Kuffler, J. Nicolls, A.Martin 所著一書(From Neuron to Brain, 2nd edition)闡明這種思考方法的可行性和其挑戰的議題,其精神是結合藉自然生性法則進行運算的意涵與平行演算法精神而成的行為計算科學。在計算方法上乃是結合類神經網路（Neural networks）及模糊邏輯（Fuzzy logic）的優點，輔以不須導式的最佳化（Derivative-free optimization）方法，例如基因程式規劃（Genetic programming）及模擬退火法（Simulated annealing）等對自然界有對應模擬的演算法,且可對特有的資料庫進行分析及微調,以建立一個具有學習能力的模型，並能自我即時（On-line adaptation）調適因時間變化而產生的新環境，

以獲取跟搜尋目標最近的結果。綜觀來說,由以自然生性為基礎的自然計算所衍生出的四個子領域正逐漸受到重視。分述如下:

(1).演化計算(Evolutionary algorithm)使用突變、重組和生物天性的優勝劣敗法則的概念。

(2).類神經網路(Neural networks)以類似

大腦高度連接神經元的自然多層結構來感知環境。

(3).分子計算(molecular computing)是以分子生物學上的範例(paradigm)基礎的 DNA 計算。

(4).量子計算(quantum computing)則是在量子物理上基於量子平行理論運算的方法。

第三跟第四項均為非傳統計算的範疇，它們徹底改變在傳統電腦計算領域上的思維。(參見國際期刊網址 http://www.wkap.nl/journals/naco)然而，它們基本目的都是為解決計算上棘手的問題如 NP problems 及 Ill-posed problems。

2.2 跨物種比較基因組的研究、費時的大規模排比

為理解生物體的生物規律而衍生之比較的基因發現方面的研究已經有了很大的進展。特別是，Batzoglou 首先提出[8] ROSETTA 程式藉由排比人類及小鼠對應基因體的區域(syntenic)的方式以小鼠基因組來註解人類的基因組。其後，類似於此以比較基因組為基礎的識別基因方法陸續地被提出，例如 CEM 程式[7]，TWINSCAN[48], SGP-1[73]，和 SGP-2[35]。這些程式比較兩個基因體的 DNA 序列既不利用蛋白質同源性也非使用經證實過的 EST(Expressed Sequence Tag)證據來預測外顯子序列，他們僅借用基因架構和接合點守舊(conserved)型樣(pattern)兩項資訊便可提高基因預測能力。

由於兩個基因體的 DNA 序列排比必須花費大量的計算時間，因此[13]提出一個較快的全域排比(global alignment)方法，AVID，以選擇好的錨(anchors)來決定序列排比的區域位置。首先將輸入兩條基因體先以重覆序列屏蔽程式遮住重覆序列，然後以字尾樹(suffix tree)方法找出最大的比對。一個字串的字尾可由沿著一個字尾樹的樹根到樹葉的路徑上串接所有經過的字元來表示並且字尾樹的分支代表不同的字尾共用相同的字首(prefix)。因此樹的每

一個內部節點都代表重複的子字串。找出兩條序列最大的比對就是，將兩條序列連接起來中間以邊界字元分隔成為一條新字串並且表示成字尾樹後，找出跨越邊界字元最長的重複子字串。接著會從最大的比對集合中挑選不重疊並且不交錯的比對當一些錨(anchors)。如果有足夠的錨來定位兩序列的比對則反覆分割此序列直到所有比對沒有重複比對， 否則直接引用傳統排比 Needleman-Wunsch 演算法[54]。

利用 AVID 當前置處理器產生近似排比，[2]提議一種新的基因發現和排比程式，SLAM。SLAM 是一種以比較法為基礎(Comparative-based) 的基因認知(gene recognition)方法，藉由以下原理來運作：

在介於相關生物體之間的守舊區(conserved regions)是比分歧區(divergent regions)更可能成為編碼區。他們描述一種機率的架構來同時地找到基因的結構和人鼠基因體對應區(syntenic genomic regions)的排比。此方法的關鍵特色是藉由發現在人鼠對應(syntenic)序列之間的最佳排比來強化基因預測能力，同時發現生物學上有意義的排比來保護在編碼外顯子序列之間的對應。此機率架構是建構在一般化的成對隱藏式 Markov 模型(GPHMM) [58]上。而 GPHMM 乃混合先前常應用在基因發現上的一般化隱藏式 Markov 模型，以及可應用在序列排比(sequence alignment)上的成對隱藏式 Markov 模型。SLAM 是同時擁有基因發現和排比的功能的程式，它可將兩個相關但尚未註解的(unannotated)DNA 序列對齊並且鑑別出其完整的外顯子/內隱子結構。此法是藉由區別出守舊的非編碼序列(conserved noncoding sequence; CNS) 而獲得正確預測的結果。與以前的方法相比，SLAM 藉由相關生物體之間基因組的排比來明顯降低預測基因結構時產生的偽陽性(false-positive)比率。

PSEP[17]漸進式信號擷取與補綴演算法(Progressive Signal Extracting and Patching; PSEP)完成一個可預測基因與多樣接合分析(Alternative Splicing; AS)的工具。該演算法係基於人與老鼠的 EST 對基因體(genome)，以及基因體對基因體的比對結果，以一連串的漸進式訊號擷取與補綴來預測基因與多樣接合分析。PSEP[17]跟目前知名的跨物種基因預測程式，如 ROSETTA[8]、TWINSCAN、SGP-1/-2、以及 SLAM，相比可得到更高的精準度。

2.3 核甘酸層次排比的缺點

然而，在此研究中發現直接以核甘酸序列進行排比有以下缺點：

1. 未經屏蔽過的核甘酸序列而加以排比會產生非常多的重複排比(repeat match)，由[17]所報告的比對結果過濾高達 88%的可能雜訊

便可得知。

2. 然而經屏蔽過核甘酸序列會破壞所要預測的基因結構。

3. 因為在跨物種基因體之間以核甘酸序列排比的特異性比氨基酸序列排比為低，所以匹配的標準不一。

4. 排比速度過慢需要大型計算機才能處理，不適合全面性反覆地改變排比參數來調整產出結果。

因此讓生物學家探索新穎基因的表現特徵時，不容易以較高的資料庫層次自行設定參數，來全面性掃瞄所有序列而非特定的對應區以符合其研究目的所需的量少質精而且包含多重註解的基因序列。

2.4 氨基酸序列的外顯子層次排比的策略

我們根據以下策略改進以上缺點：

1. 藉由實作雙連殘基理論過濾可能的雜訊，可使用不必經屏蔽的序列當輸入使得所要預測的基因結構不致被破壞。實驗證明此理論方法非常接近最佳值。

2. 使用外顯子對外顯子轉譯後的氨基酸序列排比而非 EST 對基因體(genome)或基因體對基因體的排比，因為跨物種排比的特異性會提高。

3. 直接使用氨基酸序列來進行排比，因為如此可以廣泛應用氨基酸外形資訊來進行全方位的比對，在信號處理上較為接近功能以及折疊方式或結構的比較。

4. 以極快速的群聚雜湊信號索引結構允許藉由調整參數後再重建資料庫來探索新穎基因。

CSAM 演算法以特殊的蛋白質信號共鳴技術快速辨識以及跨物種確認序列的遠距離相關，具備傳統方法已有的功能，另外本計算平台提供一個允許自訂參數的資料庫對資料庫排比的查詢環境。類似於 SLAM，我們的新穎基因預測程式使用 GENSCAN[14]當基礎基因預測程式，它一個是一種 GHMM-based 單個的基因組基因發現程式。它們考量一個包含人類基因組(genomic)序列基本的轉錄(transcriptional)，轉譯(translational)，接合信號(splice signals)以及長度分佈和外顯子，內隱子和基因間(intergenic)區的編排(compositional)特徵，用來描述基因結構的一般的機率的模型。一般來說，GENSCAN 模型用於識別在基因組 DNA 裡的完整的外顯子/內隱子的基因結構。GENSCAN 程式的可以在一條序列內預測多重基因，處理部分以及完全基因，並且可一致預測發生在其中之一或者 DNA 兩股上的基因。

2.5 探索新穎基因的實際作法

對這項研究目的－探索人類新穎基因來說，首要工作是要在所有可能的預測基因中區別出已知基因，它是一項牽涉巨大的資料庫對資料庫排比的大規模工程。為降低一些多餘的計算，我們所提的方法並不使用表達序列標籤(Expressed Sequence Tag; EST)資料庫而是使用在公開領域上的基因導向群的非冗餘集合，UniGene，來完成資料庫對資料庫的排比計算。雖然如此，這樣排比排除的計算仍然是非常費時的工作。本研究方便探索各式各樣主題的新穎基因將系統區分成三個部份，分別為外顯子資料庫的建置、序列排比與排除處理、以及跨物種保守序列比對分析。為此本研究提出一種以排序雜湊索引(OHI)架構為基礎的群聚雜湊信號掛錨(CSAM)方法，在外顯子層次上的序列進行相似度的搜尋。它利用一種蛋白質信號共鳴的方法快速地掛鉤住在外顯子層次的所有近似排比。目前本研究跨物種保守序列比對分析部分只包含與人類相似的小鼠和大鼠基因體的排比。事實上，根據如此快速的雜湊信號索引架構其他物種的排比工作也能非常容易地完成。

2.6 基於熵的過濾器雜湊信號索引掛錨

然而很多被預測出來的外顯子是由重覆序列所組成，以此雜湊信號對齊外顯子序列會造成許多重覆比對(repeat match)。因此，本研究同時也提議一個基於雙連殘基的熵理論的過濾器來區別這些重複比對而不使用任何遮罩器，如 MaskerAid[9]。首先，根據 NCBI(National Center for Biotechnology Information)上公開領域(public domain)的人類以及老鼠基因體 DNA 序列以基因預測工具 GENSCAN [14]預測並分割成較短的外顯子序列存入關聯式資料庫。接著，以相同方式以 NCBI 上較為完整的 UniGene DNA 序列轉譯成蛋白質序列的資料庫。以 UniGene 資料庫取代 EST 資料庫不但可以達到相同效果，而且能大幅降低重覆比對的成本。為了進一步加速比對效率，每一個外顯子上的胺基酸以七個胺基酸寬度的滑動窗為單位，由頭至尾依序雜湊出一個儘量唯一但有疏水性(hydrophobicty)群聚特徵的訊號值。每一個胺基酸對應一個尖峰，連續尖峰形成一個依位置次序表現的特徵曲線。假設有兩條特徵曲線，分別來自基因體 DNA 上

所預測的較短外顯子，以及一條 mRNA 所轉譯出來的較長的外顯子序列的組合。其特徵曲線上的尖峰點之值愈多相等，表示此較短外顯子可能包含於較長的外顯子組合之中。在一個預測基因中愈多的外顯子產生信號錨共鳴，則表示它們包含於任一個已被發現的基因之中，代表愈有可能是已知基因。實作上，此外顯子資料庫為衡量每一外顯子排比不同的相似程度分別列出以殘基、外顯子、和基因為單位三種不同層次的評估指標屬性。

3.研究方法與進行步驟

詳細步驟分為兩大部份。(一)、為資料探勘及軟式計算模型建立。(二)、聯邦型資料庫(Federated Database)建立。

(1)、首先，遵循資料探勘步驟，廣泛收集資料並加以編碼。以 MATLAB 6.0 以上版本內含的 Neural Network Toolbox 4.0 實施基本探勘程序來初步發現資料集基本特性概觀，並暫時以手動方式調整環境參數及輸入組成成份，進行多項實驗，接下來以 MATLAB 的 M file 自建近似 Sigmoid Belief Networks 或稱 logistic belief nets(由 1992 Neal 等人發展)的預測機(Stochastic machine)模型。由於 MATLAB 專長為矩陣運算，較不適合樹狀物件結構的實作，會影響接下來建模必要的分析比對資料環境。

(2)、為更實際應用 NFLDR 衝突敏感性資料探勘模型，我們計劃將它應用在生物資訊關於新穎基因的預測研究上。本研究建置的 CSAM 計算平台之架構圖如下：

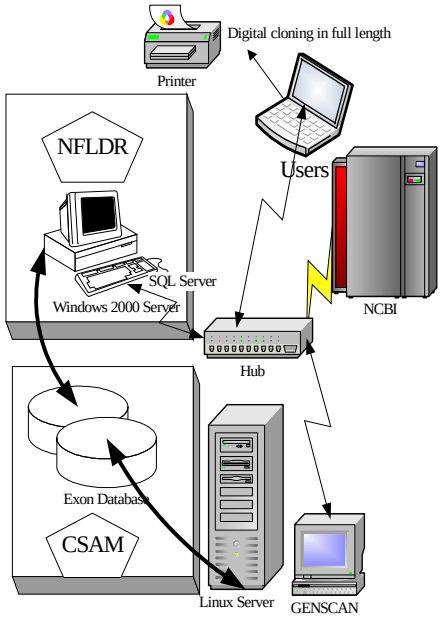

圖 1 CSAM 計算平台之架構圖

3.1 群聚信號(Clustering-Signal)

對於外顯子序列我們以特徵信號共鳴的方式將它們快速聚集成群，克服因長度不一及矩陣衍生的問題、一般群聚演算法的做法及其慢速群聚的缺點。

一般來說，群聚演算法大致區分成階層式和非階層式(non-hierarchical)兩大類[43]。非階層式方法通常依某一衡量標準將 N 個物件反覆地群聚成最佳化的 K 群。而階層式的方法傳回一個巢狀群(nested cluster)的階層，在那裡每群通常由兩或更小的群聯集而成。在階層式方法中，凝聚式(agglomerative) 群聚的方法[34]從單個物件的群開始並且把他們遞迴地合併進更大的群。反之，分裂式(divisive) 群聚[71]的方法則由包含全部物件的群開始並且把它遞迴地分成更小的群。然而，不管用群聚階層式或者非階層式的方法，要它們反覆地找到最佳的群是很費時的。利用群聚信號與氨基酸序列的關係，群聚序列排比， 序列中有部份位置會相似 motif conserved domain 藉以減少多餘的比對。

大多數群聚演算法必須計算在群之間的距離以便決定合併或者分解他們。以最簡單的方式像 UPGMA 那樣的階層式的方法，[65]則透過使用算術平均數計算在群之間的距離。在對大規模基因表達資料分析上，[21]應用一個標準會凝聚的階層式群聚演算法，average-linkage analysis。接著，[10]以分享特徵的方式整合了會凝聚的階層式群聚演算法和 K-mean 演算法[26]。此法形成有限個數的群的方式是根據一個外顯子相對於所有其他基因群中與某一特定基因群有最小的平均距離，親和力(affinity)值，和一個切斷門檻值(cut-off)兩項因素來決定該外顯子加入或移移除自此特定基因群。

Han[36]使用一種內含相互連結節點的超圖(hypergraph)，i.e.群，允許比對節點，i.e.基因，以部分相似的概念去群聚資料。此外，[53]階層地將物件放入一個無循環的有向圖(acyclic directed graph)，其中每群可能來自是許多父群的一部分。然後此演算法以整體結構圖達到最簡潔(compact)的衡量標準來最佳化群的數目。另一方面，依據最大的熵原理算出群內距離平方和(sum of squared within-cluster distances)的成本，Alon[4]利用階層方法分裂較大的結腸癌基因表達的資料歸類出數個特徵群。之後，Califano[15]藉由群聚高度相關的微列陣樣品群來鑑定最重要的細胞顯型

(phenotype)分類的基因群。相同的技術也能用來找到在他們的表達外形(express profile)的子集上方高度共同表達基因(coexpressed gene)的群。以上這些群聚方法皆允許基因跨越分佈在不同群之間。這使得他們特別適合於真核基因表達。

根據生物學上常見結合多個調控功能序列片段(regulatory motifs)或者具多重功能類別的基因。事實上，因為多個轉錄因子和加強子的複雜控制作用，所以群聚基因表達的方法應該允許基因跨越分佈在不同群之間。同樣地，CSAM 方法設計出先天性具有任一外顯子序列分佈在不同基因群甚至不同物種基因群內的能力。而且，與這些要反覆計算距離的方法不同，CSAM 方法在一回合內即可決定出所有具有最小距離的群。因此，CSAM 方法僅需要一回合便能自然消除在每群中最壞適合(ill-fitting)的匹配基因。

3.2 定義群聚雜湊信號及目標定量化

一個群聚雜湊信號分成兩個部分，一個以累加鄰接的氨基酸外形，像是疏水性指數，來表示它的特性。另一個部分使用特定位置的雜湊技術來識別序列中氨基酸共有序列的型樣(pattern)。CSAM 指定以此群聚信號可能性分佈為參數的雜湊函數當作歸屬函數(membership function)給每一群。允許每一基因中各個外顯子可以個自分類到許多群內，而不是將每個外顯子分類成為特定的群。每種氨基酸的許多外形(profile)特性，像氨基酸疏水性，可用來建構一個信號用來描繪一個有物理化學上意義的特徵外形。此信號，稱為群聚信號，是根據不同氨基酸位在不同的排列情況下，所展現的外形頻率的一個合成值。

由序列上氨基酸[疏水性指數]累加之後形成(雜湊群聚信號) 再以一個的固定大小的視窗範圍內的毗鄰氨基酸以重疊方式往下移動一個單位在系統建置之初，所有的氨基酸序列會根據此信號合成值的規則，累計每一氨基酸以及固定數目的左右數個毗連的氨基酸疏水性指數，來產生許多特有的外顯子或蛋白質外形資料庫。另外，避免群聚目數太少造成外顯子序列層次的排比對象範圍過大，額外的特定位置的雜湊值同時被加入此群聚信號之中。然後再根據信號頻率相同時引發的共鳴(sympathy)特性，將基因所屬的外顯子在共鳴處以大規模掛錨方式緊緊相互固定住。

CSAM 以雜湊群聚信號中的位置排列資訊來分裂大的群,而以外形資訊來凝聚 agglomerative 小的群因此透過群聚雜湊信號，以索引聯合比對方法迅速地形成許多重疊外顯子的基因群。因為雜湊群聚信號的位置雜湊值遠小該氨基酸外形頻率合成值，如此極小的雜湊值不會影響到群聚帶有相似外形合成值的外顯子序列集合。因此雜湊值模糊地隱含位置資訊可以避免對一條序列在氨基酸層次上直接使用傳統方法做序列排比，因而明顯地降低序列比較的時間複雜性。以雜湊群聚信號中的位置排列資訊來分裂大的群,而以外形資訊來凝聚小的群,不但免除了反覆群聚的步驟解決費時的問題而且保留序列外形資訊使得此排比更能讓具生物意義。因為 CSAM 不必依賴相似矩陣而可擺脫程式學習上的偏好，所以比其他方法來得更容易發現具新穎外顯子的基因。我們的實驗結果以及之前的不平衡資料集的研究皆顯示此種學習上的偏好會隱蓋許多有趣的型樣(pattern)。生物序列的多樣性因此在不依賴排比矩陣的偏好下可透過外形信號的共鳴而表達出來。

3.3 有序雜湊索引（Ordered Hashing Indices）

大多數的序列查詢都僅可能只存取到，不是在大資料庫中很少的一部分序列，就是一條完全序列中的一個片段。如果這些比對查詢是在兩個大規模的資料庫中交互進行，那麼大多數記錄存取則會是多餘的。理想上，快速序列搜索演算法的設計都應該考慮直接存取所需要的資料。為了獲得查詢的更直接的存取，CSAM 方法提供與群聚信號相關的一個有序雜湊索引(OHI)架構。使用這個架構能讓查詢外顯子資料庫的序列更有效率。此外，顯子資料庫建造目的是為了要在外顯子層次中，透過一個序列資料庫的輸入利用一個快速的比對方法，將另一資料庫序列上部份相似的外顯子序列篩選出來。有序索引是一種經排序過的值的索引，可供階層或循序方式存取資料值。而雜湊索引是一種經雜湊過的值的索引，可將大量的值儲存在均勻分佈的儲存桶中，使得查詢該值時可直接以定址方式快速存取資料值。OHI 是一個有效率的資料結構，歸因於結合有序索引以及雜湊索引的優勢。此儲存桶(群)，由一個雜湊函計算出一個值並指定給它。因為 OHI 用來大幅降低時間和空間的複雜性，因此它特別適合對兩大的資料庫進行交叉查詢。

3.4 排比(Alignments)

接下來我們透過了解傳統蛋白質序列排比可以比較出 CSAM 在排比的意義上有何不

同。傳統蛋白質序列比對中常使用殘基相似性計分矩陣來估計整個氨基酸序列之間的相似性。氨基酸評分矩陣常見的有 Fitch 矩陣[24]、BLOSUM 矩陣[39](BLOcks SUbstitution Matrix)、以及 PAM 矩陣[19](Point Accepted Mutation)。Fitch 矩陣則只由氨基酸遺傳編碼的相似度來建構。模塊替換矩陣 BLOSUM 是一組 20X20 的胺基酸相似度矩陣，其中的數值將用來計算在不同最小相同殘基數百分比下的兩段 10 個胺基酸序列間的相似度，用於解決序列的遠距離相關[39]。高於或等於 80%相同的序列組成的字串可用於產生 BLOSUM80 矩陣。同樣的，有 62%或以上相同的字串用於產生 BLOSUM62 矩陣。至於 PAM 矩陣是假設任一次的突變與此位置過去的歷史無關，則胺基酸被置換為其它胺基酸的機率，可用突變一次之機率的矩陣連續自乘表示。為避免 100 個胺基酸中發生多次突變，必須找相似性在 85%以上的蛋白質來估計不同形式胺基酸置換的機率來估計 PAM-1 矩陣的數值。經過自乘 250 個 PAM 單位後的突變機率矩陣，PAM250[25][64]能靈敏地測到相距甚遠的親緣關係。但 PAM250 是進化距離較遠的矩陣是從初始模型中推算出來而不是直接計算得到的[19]，因此其準確率受到一定限制。關於序列搜尋，PSI-BLAST 疊代搜尋(Position-Specific Iterated BLAST; PSI-BLAST)[6]，是另一種將雙序列比對和多序列比對結合在一起的蛋白質資料庫搜尋方法。利用第一次搜尋結果構建位置特異性計分矩陣，透過多次疊代找出最佳結果。但如果在比對前不把膠原蛋白、同源多聚體等低複雜度的重覆序列屏蔽掉，自動疊代搜尋過程會因為這些重覆序列的干擾而失敗[40]。基於此序列模式的資料庫搜尋靈敏度較高、特異性較好，所以可以發現一些距離較遠但卻具有生物學意義的相似序列。然而 PSI-BLAST 除了需要大量的計算資源外，對於搜尋結果的分析解釋常常相當困難。另一種序列搜尋，FastA 演算法[49]可識別出相匹配的被查詢序列中很短的序列片段並將位於同一對角線上相互接近的短片段連接起來。我們的計劃是不使用計分矩陣、不透過多次疊代以及在比對序列前不必將低複雜度的重覆序列屏蔽掉的方式來排除和匹配外顯子轉譯序列。此提議方法結合類似多序列比對如 PSI-BLAST，並將位於同物種上相互接近的外顯子片段以及跨物種上生物學意義的相似外顯子片段連接起來如 FastA。因為此計劃牽涉到資料庫對資料庫的交互搜尋，因此需能快速地完成大規模的序列比對使得實現全方位探索新穎基因存在成為可能。而 CSAM 方法參照傳統序列比對方法的優點並改進其提出新的做法貢獻來避免使用簡化的計分矩形模型的缺點，達到特定目的。

3.5 重複序列屏蔽問題(Repeat Sequence Problem for Masking)

在 1999 年 Smith 在他發展的序列屏蔽程式 RepeatMasker(http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl)的文件之中提到有關輸入一條經屏蔽過序列來預測基因氏應注意的幾個問題並提出一些使用 RepeatMasker 的建議。因此，與標準排比程式不同，我們的排比程序直接使用未經屏蔽過的序列。例如，即使只將散落性的(interspersed)重複序列遮罩起來，都可能使得基因預測程式無法正確識別實際的基因的位置。因為出現在外顯子編碼區的末尾(a tail end of coding region)上的轉位子(transposon)有時會被誤判為重複序列。因此，基因預測程式不應該在輸入前將低複雜區以及三核甘酸重複(trinucleotide repeat)的序列遮罩住，避免造成程式預測錯誤。而且，如果預測程式使用經屏蔽過的 DNA 當輸入序列，則此程式會將位於 3'端未轉譯區(UTR)延伸部份的散落性(interspersed)重複序列被高估(overestimate)成許多 polyA 的信號。最後，由於重複序列內可能包含一些本身能貢獻出接受者接合點（acceptor splice sites）的連續嘧啶 (polypyrimidine) 區，所以一些正確的接合點可能會被妥協（be compromised）掉而消失。

3.6 重複匹配的消除(Repeat Match Elimination)

兩條未經屏蔽過序列的排比可能會產生許多錯誤的重複序列匹配，因此我們提出雙連殘基(biresidue)的概念。一個氨基酸在胜月太鏈(peptide)裡面亦稱為一個殘基，而雙連殘基只含兩個殘基是一條最短的胜月太鏈(peptide)。本研究定義一個基於熵的雙連殘基(biresidue)的評估子，利用它來識別兩條外顯子序列的排比是否由重複匹配所構成。將這些重複匹配消除後可以得到更正確的比對結果。輸入序列可能被 RepeatMasker，(http://ftp.genome.washington.edu/RM/ RepeatMasker.html；Smit and Green)，程式處理。但是，比對被分成兩組：那些重疊的重複序列我們叫重複匹配，並且那些不重疊重複序列我們叫真實匹配。兩序

列進行匹配的結果是希望找出最大匹配，而它不必定是唯一的。我們的程式在處理真實和重複匹配上使用不同於其他程式的方法，詳述如下。

3.7 重複匹配過濾器(Repeat Match Filter)

本研究使用一個基於熵的雙連殘基(biresidue)的一種評估方法，來辨識並消除在序列排比上無義的重複序列所造成的重複匹配。這項研究透過資料庫之間的排比，以外顯子為單位將預測出來的人類外顯子扣除已知的基因所含蓋的外顯子，並比對小鼠及大鼠已知的基因序列。最後，預期產生數千個目標外顯子序列，而這些外顯子序列可能是新穎基因探索對象的一部份，它們可進一步做為實驗分析的參考序列。假設只根據雙連殘基(biresidue)評估子便能快速在兩大資料庫排比的結果中，辨識哪些外顯子匹配是重複匹配。事實上，此假設要成立必須滿足下列三項衡量標準：密實性(massiness)，偶發性(contingence)，和無所不在性(ubiquity) 。一個在外顯子序列出現的雙連殘基要被認定為已發生重複匹配，必須此三項衡量標準同時跨越門檻值。雖然這些雙連殘基也許可能產生一個的真實匹配。如果能將此衡量標準改善其應用在過濾方面的靈敏度和特異性定義好，則它將不會影響消除重複匹配的能力。

3.8 識別重複匹配之密實性、偶發性、無所不在性三項衡量標準

為有效過濾序列排比中的重複匹配，本研究利用先前實驗室所提出的一個基於熵的計算公式[41]整合以下三項衡量標準共同決定過濾的標準：

密實性：可識別的屬於重覆匹配的雙連殘基應該集中在一些序列比對內，而不是均勻分佈在每個序列比對之中。

偶發性：在每個序列比對中的每一條序列任意地同時出現毗連的兩個氨基酸均可視為一雙連殘基。並非所有雙連殘基都是重複序列，但重覆匹配的雙連殘基偶發性會較高。

無所不在性：真實匹配的雙連殘基頻繁地出現在某些序列之中。

綜合上述，我們的方法將利用關聯式資料庫之間的索引結構以一個外顯子為單位排除和匹配與群聚雜湊信號相關聯的外顯子，以簡易而有效的方法節省了大量的計算資源。包含大鼠，小鼠的跨物種比較基因組也已被用來進一步確認此收集到的基因的確存在在其他近似物種之中。最後藉用基於雙連殘基理論的過濾器讓真實匹配的新穎外顯子的基因能被快速地探索出來。

4. 結果

4.1 找到新穎基因外顯子(Exons)抄錄本總共 1130 條

本次研究把先前實驗室所找到 3,223 條新穎基因和研究執行中所收集的基因，過濾其潛在目標外顯子再利用資料探勘的技術，進行序列過濾，得到總共 1130 條新穎基因外顯子抄錄本，依其資料來源分別命名如下：

A-PMExons，來自GenBank所提供的老鼠基因組(alternate assemblies)，利用CSAM演算法與人類基因組進行交叉比對所預測出來的結果。
R-PMExons，來自GenBank所提供的老鼠基因組(reference assemblies)，利用CSAM演算法與人類基因組進行交叉比對所預測出來的結果。
oldExons，先前利用CSAM演算法與人類基因組進行交叉比對所預測出來的結果。

4.2 簡易新穎基因搜尋平台

將其上述所得到的新穎基因外顯子抄錄本，建置一個線上簡易新穎基因搜尋平台，提供生物學家與研究人員實驗參考所需，圖 2 為新穎基因搜尋平台登入畫面和圖 3 主畫面。

圖 2 系統登入畫面

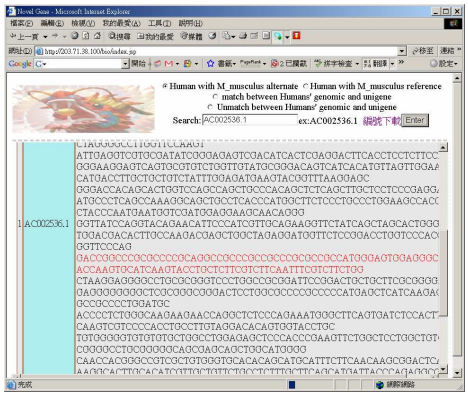

圖 3 系統搜尋主畫面

4.3 序列精鍊

本研究利用人工智慧將先前研究所得到的新穎基因，再一次將序例精鍊，更符合序列量少質精的特性。

4.4 訓鍊資料來源：

本研究從 GEO (Gene Expression Omnibus) Profiles 裡的基因序列採隨機抽樣的方式做為預測訓練集的資料。例如：從 GDS2761 當中的 47293 條序列隨機抽樣 9945 條序列，將其實驗結果的表達特徵編碼為以下二種格式：

有表達　[1 0]
無表達　[0 1]

4.5 特徵值萃取分析：

本研究從 9945 條序例以 MATLAB 的 Bioinformatics Toolbox 進行特徵值萃取分析得到十一個特徵值，如下所述，其特徵值分析示意圖，如圖 4：

(1)**最大分解度：**在所有可能將這個蛋白質分解的酵素中，挑選最大分解片段數比上它的序列長度。

(2)**迴文比率：**在DNA上不論是由哪個方向所取得的核酸序列都相同，這些序列擁有雙重的旋轉對稱，所以從3’端到5’端或是從5’到3’所得到的結果都相同，計算其出現的比率。

(3)**平均胺基酸分佈量之型態分群：**將所有出現在蛋白質中所有胺基酸占多數者編碼為1，其餘編碼為0，形成一個二進位編碼的字串，把這個字串當作分類標籤。

(4)**RS出現數目：**SR 蛋白質分子為調節另類訊息核酸先驅分子剪接之重要因子,它有一個會與轉錄因子有交互作用的domain即為(arginine/serine)多次重複的RS domain可發現在一些未知功能的新型蛋白分子上。

(5)**最大數量的mers之出現數目：**最多雙胺基酸出現的次數。

(6)**分子量等級：**分子量大小的資訊。

(7)**等電點(Isoelectric)量**：顯示分子的電性。

(8) **Trypsin cleave裂解數目：**胰蛋白酶(Trypsin)是哺乳動物體內重要的酵素之一，可將蛋白質切割成小分子的胜肽與胺基酸。

(9) **Chymotrypsin裂解數目：**Chymotrypsin 是專門分解脂肪及蛋白質。

(10)**Glutamine 裂解數目：**麩醯胺（Glutamine）是體內最豐富的胺基酸，是負責周邊組織及內臟器官的氮元素運輸者，是小腸、淋巴球及巨噬細胞主要的能量來源。

(11)**Lysine 裂解數目：**賴胺酸（Lysine)是人體無法自行合成但卻是不可或缺的必須胺基酸。

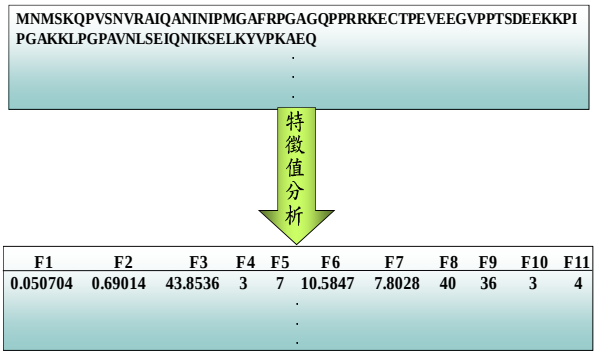

圖 4 特徵值分析示意圖

4.6 資料格式：

將 9945 條序列所獲得的十一個特徵值與表達特徵編碼資料合併，利用 weka 進行人工智慧序列精鍊探勘分析，做為之後 NFLDR 演算法模型預測的訓練資料集。資料格式如圖 5 所示：

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	class
0.050704	0.69014	43.8536	3	7	10.5847	7.8028	40	36	3	4	2
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

有表達[1 0] class = 1
無表達[0 1] class = 2

圖 5 資料格式

4.7 NFLDR 的機器學習模型預測結果：

本研究使用 NFLDR 演算法，將先前所利用 CSAM 演算法與人類基因組進行交叉比對所預測出來的新穎基因，A-PMExons、R-PMExons、oldExons，在考慮衝突敏感性(sensitive)結構下以決策規則探勘條件式表達的新穎基因(因為空間限制原因，實驗結果不予列出)。

4.8 NFLDR 和其他決策模式進行新穎基因交叉比對

將利用 NFLDR 與其他二種決策模式(C4.5、RandomForest)所精鍊出來的新穎基因，進行交叉比對，如圖 6，強化所預測出來的新穎基因，大大減低生物學家進行生物實驗的成本。

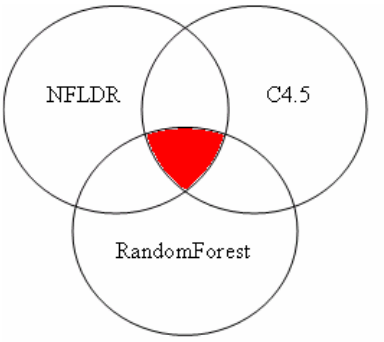

圖 6 交叉比對示意圖

4.9 新穎基因交叉比對結果：

因為 NFLDR 演算法是在衝突敏感性(sensitive)結構下以決策規則進行探勘，因此 NFLDR 會有二組新穎基因序列結果，分別為 sensitive 與 insensitive，所得到的新穎基因總筆數分別是 218 與 145 條。如表 22 為三種模式所交叉比對的結果。

5. 結論

在未來，我們將這些收集到的新穎外顯子將會陸續以二或三級結構分析結合機器學習方法，進一步縮小範圍並事先準確預測蛋白質多重功能讓條件式表達基因的實驗成本大幅降低。本實驗室將有足夠能力協助生化學家快速找到新穎基因及其轉譯之蛋白質功能，對協助國家目前所大力支持的生物科技政策而言具有貢獻。

誌謝

感謝國科會提供本研究相關經費支援(計畫編號：NSC 96-2221-E-168-030)。

參考文獻

[1] Agrawal R., and Srikant R., "Privacy-Preserving Data Mining," ***Proc. of the ACM SIGMOD Conference on Management of Data, Dallas,*** 2000.

[2] Alexandersson M., Cawley, S. and Pachter, L. “SLAM: Cross-species gene Tnding and alignment with a generalized pair hidden markov model,” ***Genome*** 13(3): 496-502,2003.

[3] Ali M., and Storey, C., “Modified Controlled Random Search Algorithms,” ***International Journal of Computer Mathematics,*** 53, pp.229-235,1994.

[4] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. “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, 1999.

[5] Alsabti, K., Ranka, S. and Singh, V. “CLOUDS: A Decision Tree Classifer for Large Datasets,” ***Conference on Knowledge Discovery and Data Mining***, 1998.

[6] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” ***Nucleic Acids*** 25: 3389-3402,1997.

[7] Bafna, V. and Huson, D. H. “The conserved exon method for gene Tnding,” ***Proc. Int. Conf. Intell. Syst. Mol. Biol.*** 8:3-12, 2000.

[8] Batzoglou, S., Pachter, L., Mesirov, J., Berger, B. and Lander, E. “Comparative analysis of mouse and human DNA and applications to exon prediction,” ***Genome*** 10 (7): 950-958,2000.

[9] Bedell, J. A., Korf, I. and Gish, W. “Maskeraid: a performance enhancement to repeatmasker,” ***Bioinformatics*** 16: 1040-1041, 2000.

[10] Ben-Dor, A., Shamir, R. and Yakhini, Z. “Clustering gene expression patterns,” ***J. Comput. Biol.*** 6(3-4): 281-297, 1999.

[11] Bentkus V., Gotze F. and van Zwet W. “An Edgeworth expansion for symmetric statistics,” ***Universit at Bielefeld,*** SFB 343, Preprint 94-020,1994.

[12] Bradley, P.S., Fayyad, U.M. and Reina, C. “Scaling Clustering Algorithms to Large Databases”, ***Fourth International Conference on Knowledge Discovery & Data Mining KDD-98,*** pp. 9-15. AAAI Press, Menlo Park, CA, 1998.

[13] Bray, N., Dubchak, I. and Pachter, L. “AVID: A global alignment program,” ***Genome*** 13:97-102,2003.

[14] Burge, C. and Karlin, S. “Prediction of complete gene structures in human genomic DNA,” ***J. Mol. Biol.*** 268: 78-94,1997.

[15] Califano, A., Stolovitzky, G. and Tu,

Y. “Analysis of gene expression micr oarrays for phenotype classiTcation,” ***8 th International Conference on Intelli gent System for Molecular Biology,*** 2 000.

[16] Carroll, R. J., Maca, J. D., Ruppert, D. “Nonparametric regression in the presence of measurement error,” ***Biom etrika,,*** 86, 3, 541-554, 1999.

[17] Chuang, T., Chen, F. C. and Chou, M. Y. “A comparative method for ide ntiTcation of gene structures and alter natively spliced variants,” ***Bioinformati cs*** 20(17): 3064-3079, 2004.

[18] Davidsson. P. “ID3-SD: An algorith m for learning characteristic decision t rees by controlling the degree of gene ralization,” ***Technical Report LU-CS-T R:*** 95-145, Dept. of Computer Science, Lund University, Lund, Sweden. 199 5.

[19] Dayhoff, M., Schwartz, R. M. and Orcutt, B. C. “Atlas of protein sequen ce and structure,” Vol. 5, ***National Bi omedical Research Foundation,*** Silver Spring, Maryland. 1978.

[20] Dietterich, T. G., Hild, H., & Bakiri, G. “A comparison of ID3 and backp ropagation for English text-to-speech mapping,” ***Machine Learning,*** 18, 51-80, 1995.

[21] Eisen, M. B., Spellman, P. T., Brow n, P. O. and Botstein, D. “Cluster ana lysis and display of genome-wide expr ession patterns,” ***Proc. Natl. Acad. Sci*** USA 95(25): 14863-14868, 1998.

[22] Esposito, F., Malerba, D., and Seme raro, G. “A comparative analysis of m ethods for pruning decision trees,” ***IE EE Transactions on Pattern Analysis and Machine Intelligence,*** 19(5):476-491,1997.

[23] Ester M., Kriegel H.-P., Sander J., Xu X. “A Density-Based Algorithm fo r Discovering Clusters in Large Spatia l Databases with Noise,” ***Proc. 2nd I nt. Conf. on Knowledge Discovery an d Data Mining (KDD-96),*** Portland, O R, pp. 226-231, 1996

[24] Fitch, W. “An improved method of testing for evolutionaxy homology,” ***J. Mol. Biol.*** 16: 9-16,1966.

[25] Fitch, W. “A non-sequential method for constructing trees and hierarchical classiTcations,” ***J. Mol. Evol.*** 18(1): 30-36,1981.

[26] Forgy, E.W. “Cluster analysis of mu ltivariate data: EfTciency versus interp retability,” ***Biometric,*** 21:768-769,1965.

[27] Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T. “Constructing Efficient Decision Trees by Using Opt imized Numeric Association Rules,” 1 46-155,1996.

[28] Fun, M.H. and Hagan, M.T. “Leven berg-Marquardt Training for Modular Networks,” ***The 1996 IEEE Internatio nal Conference on Neural Networks,*** Vol. 1, 468-473,1996.

[29] G. A. Watson “The Levenberg-Mar quardt algorithm: implementation and t heory. In G. A. Watson, editor,” ***Num erical Analysis. Lecture notes in Mat hematics*** 630, 105-116,1977. Springer-Verlag, Berlin, New York.

[30] Gehrke, J., Ramakrishnan, R. and G anti, V. R. “A Framework for Fast D ecision Tree Construction of Large Da tasets,” ***VLDB*** 1998: 416-427,1998.

[31] Gehrke, J.E., Ganti, V., Ramakrishnan, R. and Loh, W.-Y. “BOAT - Optimistic Decision Tree Construction.” ***In Proceedings of the 1999 SIGMOD Conference,*** Philadelphia, Pennsylvania, 1999.

[32] Geist, I. and Sattler, K. “Towards D ata Mining Operators in Database Syst ems: Algebra and Implementation,” ***In Proc. of 2nd Int. Workshop on Dat abases, Documents, and Information Fusion,*** Karlsruhe.2002.

[33] Gouda, K. and Zaki, M J. “Efficient ly Mining Maximal Frequent Itemsets” ***in 1st IEEE International Conferenc e on Data Mining ,*** San Jose.2001.

[34] Gowda, K. C. and Krishna, G. “Ag glomerative clustering using the conce

pt of mutual nearest neighborhood," ***PatternRecognition*** 10: 105-112,1977.

[35] Guig'o, R., Agarwal, P., Abril, J. F., Burset, M. and Fickett, J. "An assessment of gene prediction accuracy in large DNA sequences," ***Genet. Res.*** 10: 1631-1642, 2000.

[36] Han, E. H., Karypis, G., Kumar, V. and Mobasher, B. "Clustering based on association rule hypergraphs," ***In Proccedings of SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery,***1997.

[37] Hanke, M. "A regularization Levenberg-Marquardt scheme," ***with applications to inverse groundwater filtration problems,*** Inverse Problems 13, p.79-95, 1997.

[38] Harik, G., Lobo, F. and Goldberg, D.E., "The Compact Genetic Algorithm." ***Proceedings of the 1998 IEEE Conference on Evolutionary Computation,*** pp. 523-528, 1998.

[39] Henikoff, S. and Henikoff, J. G. "Clustering in a high-dimensional space using hypergraph models." ***Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci.*** USA 89: 10915-10919, 1992.

[40] Holm, L. "UniTcation of protein families," ***Curr. Opin. Struct. Biol.*** 8(3): 372-379,1998.

[41] Hung C. M., Huang, Y.M.、and Chang M. S., "CSAM: Using clustering-hashing-signal anchoring method to explore human novel genes"、***Journal of Computational Biology***、Vol. 13(10)、pp. 115-128, 2006.

[42] Hung C. M., Huang, Y.M. "Conflict-Sensitivity Contexture Learning Algorithm for Mining Interesting Patterns using Neuro-fuzzy Network with Decision Rules," ***Expert Systems With Applications***, Vol. 34(1), pp. 159-172, 2008.

[43] Jain, A. K., Murty, M. N. and Flynn, P. J. "Data clustering: a review," ***ACM Computing Surveys*** 31(3): 264-323, 1999.

[44] Kirkpatrick, S., Gelatt, C. and Vecchi, M. "Optimisation by Simulated Annealing," ***Science,*** No.220, pp.671-680, 1983.

[45] Knorr, E.M. and Ng, R.T. "A Unified Notion of Outliers: Properties and Computation", ***Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining,*** Newport Beach, CA, August 14-17, 1997.

[46] Knoth, O. "A globalization scheme for the generalized Gauss-Newton method," ***Numer. Math.,*** 56, pp. 591-607, 1989.

[47] Komorowski, J., Polkowski, L., and Skowron, A. "Rough sets: a tutorial". ***In S.K. Pal and A. Skowron, editors, Rough-Fuzzy Hybridization: A New Method for Decision Making,*** Springer-Verlag, Singapore. 1998.

[48] Korf, I., Flicek, P., Duan, D. and Brent, M. R. "Integrating genomic homology into gene structure prediction," ***Bioinformatics*** 17: 140-148, 2001.

[49] Lipman, D. J. and Pearson, W. R. "Rapid and sensitive protein similarity search," ***Science,*** 227(4693): 1435-1441, 1985.

[50] Mansour, A. and Barros, A. K. M. Kawamoto, and N. Ohnishi. "A fast algorithm for blind separation of sources based on the cross-cumulant and levenberg-marquardt method," ***In Fourth International Conference on Signal Processing (ICSP'98),*** 323-326, Beijing, China, 12-16 October 1998.

[51] Mansour, A. and Ohnishi, N. "Multichannel blind separation of sources algorithm based on cross-cumulant and the levenberg-marquardt method.," ***IEEE Trans. on Signal Processing,*** vol. 47, no. 11, pp. 3172-3175, 1999.

[52] Mehta, M., Agrawal, R. and Rissanen, J. "SLIQ: A Fast Scalable Classifier for Data Mining", ***Proc. of the Fifth Int'l Conference on Extending Database Technology,*** Avignon, France, 1996.

[53] Mjolsness, E., Castano, R. and Gray, A. “Multi-parent clustering algorithms for large-scale gene expression analysis.” ***Jet Propulsion Laboratory Technical Report JPL-ICTR,*** 1999.

[54] Needleman, S. B. andWunsch, C. D. “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” ***J. Mol. Biol.*** 48: 443-453, 1970.

[55] Nelder, J.A. and Mead, R. "The downhill simplex method". ***Computer Journal,*** Vol. 7, pp.391-398, 1965.

[56] Ng, R. and Han, J. "Efficient and Effective Clustering Method for Spatial Data Mining", ***Proc. Of 1994 Int'l Conf. on Very Large Data Bases (VLDB'94),*** Santiago, Chile. 1994.

[57] Nix, D.A. and Weigend, A.S. “Learning local error bars for nonlinear regression,” ***Advances in Neural Information Processing Systems (NIPS),*** MIT Press. 1994.

[58] Pachter, L., Alexandersson, M. and Cawley, S. “Applications of generalized pair hidden markov models to alignment and gene Tnding problems,” ***J. Comput. Biol.*** 9: 389-399, 2002.

[59] Pearson, J.K. and Bisset, D.L. Back Propagation in a Clifford Algebra. ***Artificial Neural Networks, 2, I. Aleksander and J. Taylor (Ed.),*** 413:416. 1992.

[60] Rastogi, R. and Shim, K. “PUBLIC: A decision tree classifier that integrates building and pruning.” ***In Proceedings of the Very Large Database Conference (VLDB),*** 1998.

[61] Roberts, S.J. “Novelty Detection using Extreme Value Statistics.” ***IEE Proceedings on Vision, Image & Signal Processing,*** 146(3):124-129, 1999.

[62] Sattler, K. and Dunemann, O. "SQL Database Primitives for Decision Tree Classifiers", ***In Proc. Of the 10th ACM CIKM Int. Conf. on Information and Knowledge Management,*** November 5-10, 2001.

[63] Shafer, J.C., Agrawal, R., Mehta, M. "SPRINT: A Scalable Parallel Classifier for Data Mining", ***Proc. of the 22th Int'l Conference on Very Large Databases,*** Mumbai (Bombay), India, Sept. 1996.

[64] Smith, T. F., Waterman, M. S. and Fitch, W. “Comparative biosequence metrics,” ***J. Mol. Evol.***18(1): 38-46. 1981.

[65] Sneath, P. H. A. and Sokal, R. R. Numerical Taxonomy, Freeman, San Francisco. 1973.

[66] Solis, F. J. and Wets, R. J-B. “Minimization by Random Search Techniques.” ***Mathematics of Operations Research,*** 6:19-30, 1981.

[67] Subramanian, P. K. “Gauss-Newton methods for complementarity problems, J. Optim.” ***Theory Appl.,*** 77 pp. 467-482. 1993.

[68] Szu, H. and Hartley, R. “Fast simulated annealing.” ***Physics Letters A,*** 122(3/4):157-162. 1987.

[69] VanTulder MW, Assendelft WJJ, Koes BW, Bouter LM. “Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for spinal disorders.” ***Spine,*** 22:2323-2330. 1997.

[70] Vidakovic, B. "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," ***J. Amer. Statist. Assoc.,*** vol. 93, pp. 173-179. 1998.

[71] Wan, S. J., Wong, S. K. M. and P., P. “An algorithm for multidimensional data clustering,” ***j-TOMS*** 14(2): 153-162. 1988.

[72] Wedderburn, R.W.M. “Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method.” ***Biometrika,*** 61 439-447. 1974.

[73] Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T. and Guigo, R. “SGP-1: prediction and validation of homologous genes based on sequence alignments,” ***Genome Res.*** 11: 1574-1583. 2001