以類神經網路預測基因的多重功能

江素倩　陳春賢

長庚生物資訊中心、長庚大學資訊管理研究所

E-mail Address

m9244003@stmail.cgu.edu.tw　cchen@mail.cgu.edu.tw

摘要

隨著人類基因組的解碼，對於健康與醫療的觀念逐漸產生重大的改變，傳統以病徵來診治疾病的醫療方式，將隨著對基因組的日益了解，而逐漸被以基因病理、基因診斷與基因治療為主的「基因體醫學」所取代。而後基因體時代的重要任務之一，便是深入了解基因組，因此對於基因功能的了解為後基因體時代主要的課題之一。由於人類基因體較複雜且缺乏完整的相關實驗，而酵母菌為實驗室重要的模式生物，且自1996 年定序完成後，已累積相當規模的實驗資料，因此本研究以酵母菌為對象，以已知的酵母菌基因功能架構與酵母菌相關基因表現資料，進行酵母菌基因功能預測，並針對使用叢集分析方法分析基因表現資料只能預測單一基因功能之弱點，以倒傳遞類神經網路分類器為基礎，發展一套足以預測基因多重功能的方法，以期更貼近生物體實際運作情形，並以公開的酵母菌基因相關資料配合已知功能基因資訊進行訓練與學習。

關鍵字：基因功能預測、多重基因功能預測、分類問題、倒傳遞類神經網路

壹、前言

隨著人類基因組的解碼，對於健康與醫療的觀念逐漸產生重大的改變，傳統靠病徵診治疾病的醫療方式，將可能逐漸被以基因病理、基因診斷與基因治療為主的「基因體醫學」所取代[1]。而從基因的角度出發，了解基因的基本功能、疾病的基因機制與機轉，便成為後基因體時代，實現基因體醫學的重要課題之一[8][12]。

由於基因透過轉錄、轉譯過程，以合成蛋白質、酵素等巨分子進行其功能表現，因此透過對轉譯過程中基因表現量的測定，可以觀察相同實驗條件下基因的表現情形，進而推論基因所隱含的功能。近年來，由於基因微陣列(cDNA Microarray)技術的廣泛使用，對於個別基因表現量的定量分析有非常大的幫助與突破[7][18]。透過基因微陣列可以同時檢測數千甚至上萬個基因的表現量，因此基因微陣列的實驗可產生大量的原始資料，生物學家可透過對這些資料所進行的分析，了解生物體的運作模式與機轉，其中包括基因調控、細胞發展、生物演化、基因與疾病之間的關係等[2]。

目前以基因表現量資料來分析基因功能的常用分析方法，多為以非監督式學習為基礎的叢集分析方法，雖然叢集分析方法是重要的資料探勘技術之一，但叢集分析方法一般最主要的限制為一個基因只會被預測為單一功能[5][13]；然而實際上在生物體中，一個基因可能參與多個調控路徑、具備多種功能，因此採用此種方法容易忽略基因可能擁有的其他功能資訊。此外，由於基因體時代投入大量人力物力進行研究，已累積了相當豐富的生醫研究成果，King 等學者於是提出以監督式方法為基礎進行分析的概念[14]，並以決策樹為基礎的分析方法應用於結核分枝桿菌(*M. tuberculosis*)和大腸桿菌(*E. coli.*)的基因功能預測上[15]。

類神經網路可以解決圖形識別(pattern recognition)、函數逼近(function approximation)、分類(classification)等問題[20]，可透過訓練所得到的類神經網路模型對未知資料進行推論，目前已被廣泛地應用於決策分析、醫療分類、銀行、語音辨識、手寫辨識、生物資訊等領域[11]。

倒傳遞演算法為以多層前饋式類神經網路架構為基礎，所發展出來的網路訓練學習法則，採監督式學習，透過目標向量與網路模擬輸出之間的誤差及靈敏度，進行網路連結權重的調整。由於類神經網路對雜

亂資料的高承受能力，及其對未經訓練資料分類模式的能力，推動了類神經網路在資料探勘分類上的普遍應用。

表 1 功能類別架構

class ID	MIPS class	Function Description
1	[1,0,0,0]	METABOLISM
2	[10,0,0,0]	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
3	[11,0,0,0]	CELL RESCUE, DEFENSE AND VIRULENCE
4	[13,0,0,0]	REGULATION OF / INTERACTION WITH CELLULAR ENVIRONMENT
5	[14,0,0,0]	CELL FATE
6	[2,0,0,0]	ENERGY
7	[29,0,0,0]	TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS
8	[3,0,0,0]	CELL CYCLE AND DNA PROCESSING
9	[30,0,0,0]	CONTROL OF CELLULAR ORGANIZATION
10	[4,0,0,0]	TRANSCRIPTION
11	[40,0,0,0]	SUBCELLULAR LOCALISATION
12	[5,0,0,0]	PROTEIN SYNTHESIS
13	[6,0,0,0]	PROTEIN FATE (folding, modification, destination)
14	[62,0,0,0]	PROTEIN ACTIVITY REGULATION
15	[63,0,0,0]	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
16	[67,0,0,0]	TRANSPORT FACILITATION
17	[8,0,0,0]	CELLULAR TRANSPORT AND TRANSPORT MECHANISMS
18	[98,0,0,0]	CLASSIFICATION NOT YET CLEAR-CUT
19	[99,0,0,0]	UNCLASSIFIED PROTEINS

在本研究中，將類神經網路方法應用在多重基因功能的分類上，透過監督式學習的方式，以倒傳遞演算法作為本研究所使用的類神經網路模型。比例共軛梯度演算法(scaled conjugate gradient)為倒傳遞類神經網路用來進行監督式學習的演算法之一，由於其具有在分類問題與大量參數時較佳的表現[17]，因此我們採用此演算法作為實際訓練類神經網路所使用的演算法。

酵母菌(*Saccharomyces cerevisiae*)為第一個被完整定序的真核生物[10]，且由於其便宜、生長快速、不致病、容易進行基因相關操作等特性，使其成為實驗室中非常重要的模式生物(model organism)，截至目前為止，並累積了相當規模的研究成果[3][4][6][7][9][19]。MIPS[16]更針對酵母菌制定了功能架構(function categories)，對酵母菌基因註解的功能進行有系統的分類。

本研究中將以酵母菌為研究對象，並以 MIPS 的酵母菌基因功能分類架構為依據，探討針對其基因表現資料進行功能分類而設計的類神經網路架構。

貳、資料來源與研究方法

一、資料蒐集與處理

本研究中，採用 Eisen 與 Spellman 所發布使用的酵母菌基因表現實驗[7][19]。Eisen 的資料組中包含 6178 個基因 79 個實驗，觀察基因在包括以 α-factor、elutriation、Cdc15 等三種方式分別進行同步化

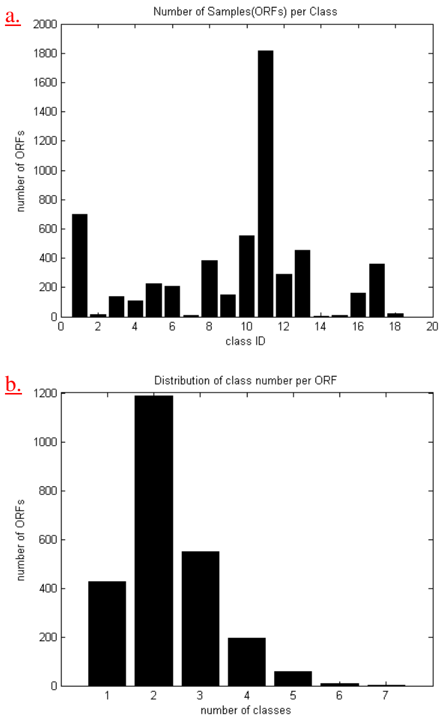

圖 1 原始資料分布圖

(synchronization)後的細胞週期所得到的 47 個基因表現量測值，及觀察酵母菌從雙倍體(diploid)產生單倍體(haploid)中芽孢形成(sporulation)過程的 11 個量測值，另外有包括使用高溫、DDT、低溫等三種不同環境條件下所觀察到的 14 個基因表現量測值，及酵母菌從發

酵生活轉到呼吸生活(無氧轉變到有氧生活)代謝轉變過程的 7 個量測值；Spellman 的資料包含 2467 個基因、77 個實驗，包括以 Cln3、Clb2、α-factor 、elutriation、Cdc15、Cdc28 等六種方式分別進行同步化後所觀察得到的基因表現測量值共 77 個；我們將兩組實驗資料取基因名稱交集、實驗資料聯集，得最後的資料共 2425 筆基因分別於 156 個實驗中的表現資料。針對實驗數據遺失值的處理採用各組實驗之平均值分別填入。

MIPS[16]針對基因功能所定的功能分類為階層狀的分類架構，最頂層的分類共 19 類別，每個類別底下又細分子類別，為求訓練樣本資料集合夠大，我們採用最上層的分類架構作為實驗所用的類別標籤，表 1 列出 19 個功能類別標籤。圖 1a 為各類別集合內的樣本筆數分布，橫軸為類別標籤，縱軸為樣本數，樣本數最多的第 11 類為與蛋白質在細胞中的位置有關的基因(SUBCELLULAR LOCALISATION)，共有 1814 個基因具有此類別的功能；第 1 類則是具有新陳代謝(METABOLISM)相關功能的基因，總共有 697 個基因具有此功能。每個基因具有的類別標籤數最多為 7 個，其分布情形如圖 1b 所示，橫軸為基因所擁有的類別標籤個數，縱軸為樣本數。

二、類神經網路設計

針對酵母菌 19 個功能分類，我們針對每個類別個別訓練一個架構為 156-h-1 的類神經網路分類器(h 為隱藏層神經元個數)，每個分類器的輸入神經元均相同(156 個實驗資料)，輸出為 0 或 1，0 代表不具有此分類器所代表的功能類別，1 代表具有此分類器所代表的功能類別。

總共設計了 17 個類神經網路分類器，由於其中第 18 與第 19 個類別分別是功能尚未被明確定義與功能未知的基因，因此不對此兩類別設計分類器。

圖 2 為由 17 個分類器所組成的網路架構圖，輸入層部分共有 156 個神經元，隱藏層神經元依各分類器訓練結果進行調整，各分類器皆有一個輸出層神經元。每個神經元以 sigmoid 函數作為轉換函數，以比例共軛梯度演算法進行網路學習訓練。

三、實驗流程

所有基因被隨機分為三等份，每次選擇其中 1/3 作為測

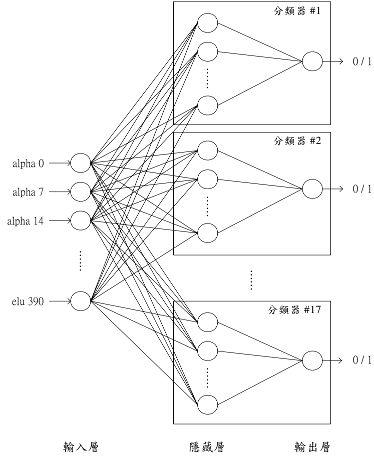

圖 2 類神經網路架構圖

試樣本，另外 2/3 則作為訓練樣本，分別對 17 個分類

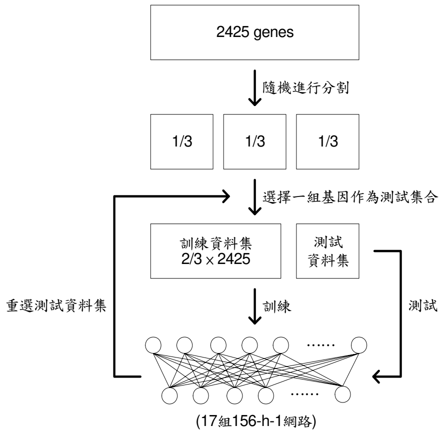

圖 3 實驗流程圖

器不同隱藏層神經元個數的網路架構進行訓練，並計算其準確率，流程如圖 3 所示。個別分類器的準確率計算方式如下式(1)：

$$accuracy = 1 - \frac{misclass}{N} \tag{1}$$

其中，N 為樣本總數，*misclass* 為網路推論輸出與目標輸出不同的樣本數。

參、結果與討論

蒐集得到的 2425 筆基因隨機分為三等份，每次挑選其中一等份作為測試資料集，三組訓練與測試資料的分布情形如圖 4，橫軸為類別標籤，縱軸為具有此類別標籤的基因筆數。

三組訓練、測試樣本分別對 17 個分類器進行訓練後，總共得到 51 個訓練好的類神經網路結構。測試資料與訓練資料分別計算其分類準確率，並取三組資料集的

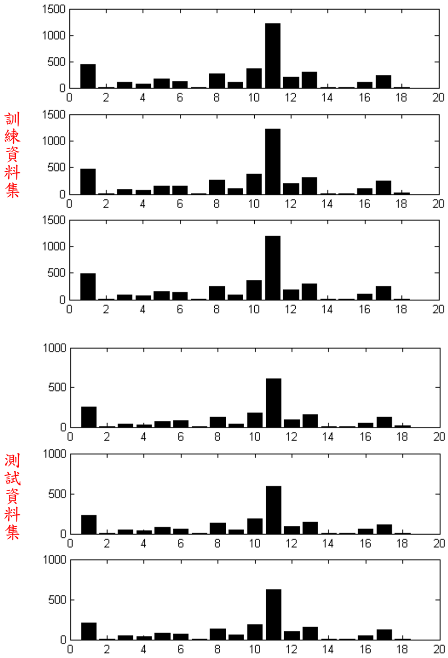

圖 4 訓練、測試資料集分布情形

平均準確率作為個別分類器的準確率。

表 2 隱藏層神經元個數與準確率

classifier	sample size	h	training accuracy (%)	testing accuracy (%)
1	697	19	99.24	66.10
2	14	3	100	98.10
3	135	9	99.38	89.77
4	108	5	99.44	90.85
5	228	15	99.36	84.25
6	205	6	99.44	86.23
7	11	2	99.98	98.95
8	385	18	99.32	75.59
9	147	5	99.48	85.90
10	550	16	99.34	70.14
11	1814	20	99.46	63.05
12	290	14	99.53	89.36
13	452	20	99.38	71.71
14	4	1	100	99.75
15	6	5	100	99.46
16	158	12	99.36	90.19
17	361	19	99.51	75.22

倒傳遞類神經網路架構的隱藏層個數目前並無確切的計算方式，因此本研究中針對隱藏層神經元個數以試誤法的方式，針對神經元個數 h 為 1 到 20 的網路架構進行準確率的計算。圖 5 列出 17 個分類器分別的隱藏層神經元個數及其相對應的準確率，橫軸為隱藏層神經元個數，縱軸為準確率，圖 5a 為訓練資料集計算得到的平均準確率，圖 5b 為測試資料集平均準確率。表 2 列出各分類器訓練準確率最高的隱藏層神經元個數，及在此網路架構下所得到的測試資料準確率，其中 h 為隱藏層神經元個數。

其中，所有 17 個分類器的個別訓練資料準確率都在 99%以上，但第一、八、十、十一、十三、十七個分類器的測試準確率則偏低，根據圖 1a 可看出，這幾個類別的樣本數較大，本研究所使用的類神經網路架構可能不足以學習所有包括測試資料的案例。

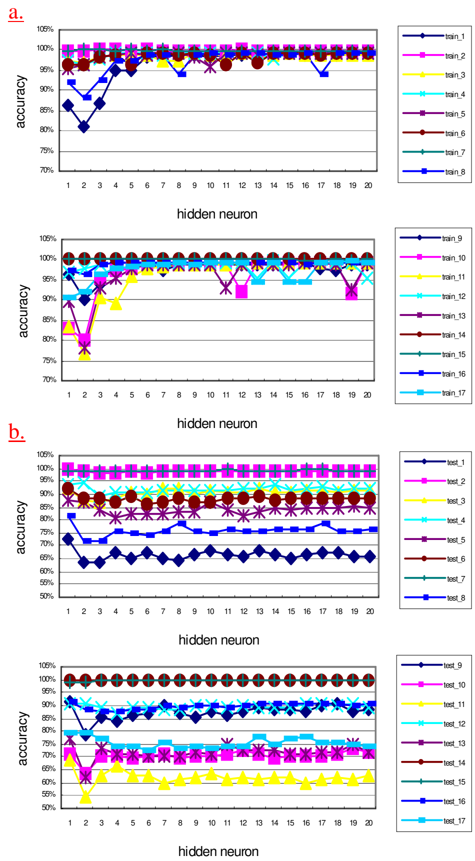

圖 5 網路架構與準確率關係圖

這訓練好的 17 個分類器可被視為一組類神經網路分類器，當功能未知的基因被輸入時，可以同時考量多類別的分類。

肆、結論與未來展望

本研究提出了一個以類神經網路分類器為基礎，針對基因表現資料，將基因多重功能問題拆解成多個二元分類問題的架構進行分析，透過此種分類方式可以有效解決多重類別的問題，而類神經網路本身所具有的容錯特性，對目前充滿雜訊的生醫資訊亦相當適合。

實驗結果顯示，所有 17 個分類器的個別訓練資料準確率都在 99%以上，但其中第一、八、十、十一、十三、十七個分類器的測試資料準確率則較低，根據圖 1a 原始樣本類別分布圖可看出，這幾個類別的樣本數較大，樣本資料所呈現的複雜度因此較大，而本研究所使用的類神經網路架構也因此可能不足以透過學習而歸納出樣本空間的完全特性。

然而在資料蒐集上，由於輸入特徵值維度相當大，在樣本集合不夠大的情況下，若能結合特徵值篩選方法針對各分類器挑出相關性大的輸入特徵，應能有效提升分類準確率；而基因表現資料本身能否正確蒐集到所有與基因功能表現相關的實驗也是潛在的問題之一，若能結合其他與基因相關的特徵輸入，如：序列資料、同源性分析資料、及其他相關特性資料等，再進行特徵挑選，應能提升分類準確率。

伍、誌謝

感謝長庚醫院研究計畫 CMRPD 1008-II 與國科會計畫 NSC93-2213-E-182-010 的支持，提供我們此機會研究此基因多重功能預測方法。

陸、參考文獻

1. 鄭博仁、謝燦堂(2000)，健康密碼：探索人類基因奧秘，初版，台北：聯合文學，頁 11-13。
2. P. Baldi, G. Hatfied (2002), DNA microarrays and gene expression, Cambridge: Cambridge University Press, pp1-17.
3. H. Boucherie, G. Dujardin, M. Kermorgant, C. Monribot, P. Slonimski, M. Perrot (1995), "Two dimensional protein map of Saccharomyces cerevisiae: construction of a gene-protein index," Yeast, 11(7), pp601-613.
4. R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, R. Davis (1998), "A genome-wide transcription analysis of the mitotic cell cycle, " Molecular Cell, 2(1), pp.65-73.
5. A. Clare, R. King (2003), "Predicting gene function in Saccharomyces cerevisiae," Bioinformatics, 19, Suppl 2:II42-II49.
6. J. DeRisi, V. Iyer, P. Brown (1997), "Exploring the metabolic and genetic control of gene expression on genomic scale, " Science, 278(5338), pp680-686.
7. M. Eisen, P. Spellman, P. Brown, D. Botstein (1998), "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci. USA, 95(25), pp14863-14868.
8. D. Eisenberg, M. Marcotte, I. Xenarios, O. Yeates (2000), "Protein function in the post-genomic ear," Nature, 405, pp823-826.
9. A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, P. Brown (2000), "Genomic expression program in the response of yeast cells to environmental changes," Mol. Bio. Cell, 11, pp4241-4257.
10. A. Goffeau, B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq, M. Johnston, E. Louis, H. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S. Oliver (1996), "Life with 6000 genes", Science, 274(5287), pp563-567.
11. M. Hagan, H. Demuth, M. Beale (1996), Neural Network Design, Boston MA: PWS Publishing.
12. P. Hieter, M. Boguski (1997), "Functional Genomics: It's All How You Read It," Science, 278(5338), pp601-602.
13. J. Ihmels, G. Friedlander, S. Gergmann, O. Sarig, Y. Ziv, N. Barkai (2002), "Revealing modular organization in the yeast transcriptional network," Nat. Genet., 31(4), pp370-377.
14. D. Kell, R. King (2000), "On the optimization of classes for the assignment of unidentified reaings frames in functional genomics programmes: the need for machine learning," Trends Biotechnol., 18(3), pp93-98.
15. R. King, A. Karwath, A. Clare, L. Dehaspe (2001), "The utility of different representations of protein sequence for predicting functional class," Bioinformatics, 17(5), pp445-454.
16. H. Mewes, K. Albermann, K. Heumann, S. Liebl, F. Pfeiffer (1997), "MIPS: a database for protein sequences, homology data and yeast genome information," Nucleic Acids. Res., 25(1), pp28-30.
17. M. Moller (1993), "A scaled conjugate gradient algorithm for fast supervised learning," Neural Network, 6, pp525-533.
18. M. Schena, D. Shalon, R. Davis, P. Brown (1995), "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," Science, 270(5235), pp467-470.
19. P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, B. Futcher (1998), "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Mol. Biol. Cell, 9(12), pp3273-3297.
20. G. Zhang (2000), "Neural networks for classification: a survey," IEEE Trans. Syst., Man and Cybern., Part C: Application and Reviews, 4(30), pp451-462.