行政院國家科學委員會專題研究計畫　成果報告

微陣列晶片基因表現資料之分析—使用以查詢為基礎之遺傳演算法

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-002-086-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立臺灣大學工程科學及海洋工程學系暨研究所

計畫主持人：張瑞益
共同主持人：陳彥良

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 8 月 24 日

行政院國家科學委員會專題研究計畫成果報告

微陣列晶片基因表現資料之分析—使用以查詢爲基礎之學習方式與遺傳演算法
Analysis of microarray gene expression data: using query-based learning methods and genetic algorithms

計畫編號：NSC 93-2213-E-002-086

執行期限：93 年 8 月 1 日至 94 年 7 月 31 日

主　持　人：張瑞益　國立台灣大學工程科學系
共同 主 持人：陳彥良　國立中央大學資訊管理系
計畫參與人員：鞠適存（中央資管）、羅嘉彥（台大工科）、賴良賓（台大工科）

一、中英文摘要

建構一個分析微陣列晶片大量基因表現資料的方法已成為當今最刻不容緩的議題之一。過去五年來，已有許多分群方法被發展用以找出微陣列晶片中的表現基因所代表的功能。這些方法大多是以傳統的自我組織映射圖或是k-means分群方法為基礎建構而成。然而，這些方法仍未臻完美。本研究以查詢為基礎的學習方式為主軸，配合傳統自我組織映射圖之類神經網路演算法，建構一套新的分析方法。實驗結果證明，使用以查詢為基礎之自我組織映射圖方法，將比原始SOM方法具有更快速之學習效率（收斂速度快）與不易受到初始網路拓樸的干擾而影響分群結果（穩定性高）等兩大優點。

關鍵詞：微陣列晶片、以查詢為基礎的學習、類神經網路、自我組織映射圖、分群

Abstract

There is a growing need for a method to analyze massive gene expression data obtained from DNA microarray. Over the last five years, many approaches developed to categorize the gene expression data into functionally meaningful groups. Most of these methods are based on self-organizing maps or k-means clustering method. However, there are some difficulties in these approaches. In this study, we introduced the query-based learning methods into self-organizing maps. The experiment results show that our method is more efficient and more stable than the original SOM method.

Keywords: microarray, query-based learning, neural network, self-organizing maps, clustering

二、計畫緣由與目的

隨著人類基因定序圖譜的解讀完成，科學家們的下一個課題便是了解數萬個基因所代表的意義，和蛋白質組功能間的相互關係。在數量如此龐大的基因群組中，要挑選出我們想研究的標的物相當困難，微陣列(Microarray)基因晶片技術的出現，正是解決此複雜難題的最佳利器。微陣列基因晶片又名生物晶片(Biochip)，其主要特點是精確性高、分析速度快，所需使用的檢體樣品及試劑少，且一次實驗就可獲得整體性實驗數據。微陣列晶片技術挾著其具有一次分析大量基因表現資料的優勢能力，已成為生物科技領域不可或缺的重要技術之一，其應用的範圍涵蓋了基因功能研究、臨床檢驗、新藥開發、菌種篩選等等[1]。

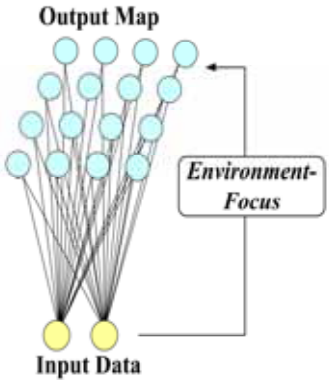
微陣列技術雖然可以產出大量的原始基因表現資料，但若無後端資料分析的方法，也無法得到任何的生物意義。在各種基因表現資料的分析技術

中，分群(clustering)方法是最根本，也是最重要的技術之一。分群在生物學上的假設前提為：相似功能或是具有 coregulated 關係的基因，會有相似的基因表現 pattern[2, 3]。因此，藉由將相似 pattern 的基因歸到同一群，可以省去生物學家在茫茫資料大海中探索基因功能的時間。目前已有相當數量的文獻與研究並證實 clustering 對於探索基因表現資料有相當顯著的價值。

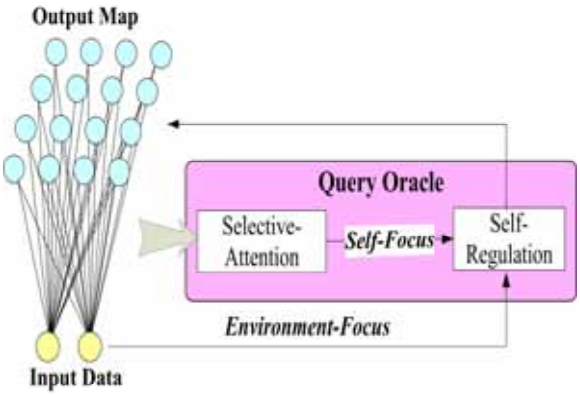
近幾年來，已有許多分群方法被用在基因表現資料的分析上，例如：階層式分群演算法(hierarchical clustering)[4-6]、K-means 演算法[7][8]以及自我組織映射圖(Self-Organizing Maps, SOM)[9-12]等。但可惜的是，這些方法各自均存在一些問題或限制。階層式分群演算法的缺點是缺乏強健性(robustness)、分群結果不具有獨一性(uniqueness)，當資料量大時，也有不易解釋的問題[9][13]。而 K-means 演算法的主要缺點則為對雜訊(noise)敏感[13]。由於基因表現資料分析的特性為資料量大、資料維度高，且極可能含有相當多的雜訊資料或 missing value，使得階層式分群方法與 K-means 方法的應用受到了相當大的限制，因此，許多學者開始應用 SOM 方法求解此一問題。SOM 可以說是目前最廣泛地被應用在分析基因表現資料的方法之一。其主要的特性為，具有強健性、分類精確度佳、抗雜訊能力強，且執行速度也能令使用者滿意[9][14]。這些特性使得 SOM 特別適合用於基因表現資料的叢集分析上。此外，SOM 能將高維度的資料空間映射到較低維度的空間上（通常是二維空間），這使得叢集分析結果易於視覺化表示(visualization)且容易解釋[9][14]。

隨著許多免費或付費工具的提供[15-18]，SOM 儼然已成為生物學家最常使用的叢集分析工具之一。然而，SOM 仍存有一些限制與缺點。例如：其初始化的狀態會嚴重影響叢集分析結果，即是其主要缺點之一[19]。這是由於 SOM 之初始權重與輸入資料之順序均採隨機方法，較好的初始化狀況會使得 SOM 產出較佳的分群結果，反之，分群結果將難以使人滿意。基於上述原因，在實務上，生物學家在使用 SOM 分析基因表現資料時，大都需要進行數次的重複實驗，再從多次結果中選擇最佳的結果，徒費心力與時間。因此，我們導入計畫主持人於 1997 年所提出之以查詢為基礎的自我組織映射圖(Query-Based Self-Organizing Maps, QBSOM)方法[20]，期望能改善上述傳統 SOM 的問題。

QBSOM 的主要概念是在傳統只學習外部資料(external-input samples)的 SOM 之外，再加入內在渴望(internal-desired samples)的學習因子，圖 1 比較了 QBSOM 與傳統 SOM 概念上的差異。傳統的 SOM 只一味的要求學習者完全接收外部資料，如同填鴨式教育；而 QBSOM 則同時考慮了學習者的內在渴望，這種學習的方式，正符合孔子所提倡的：「因材施教」。在本計畫中，我們以 SOM 的網路拓樸(topology)來詮釋內在渴望，亦即 QBSOM 不僅如傳統 SOM 一般學習外部之輸入資料，也會根據 topology 進行學習。傳統 SOM 與 QBSOM 在幾何學上的意涵如圖 2 所示，不像傳統 SOM 只考慮了外部樣本（以 v 表示），QBSOM 同時考慮了 internal-desired sample（以 u 表示），所以 winner neuron 不同，其移動（學習）的方向也不同。因此，在同樣的訓練條件下，QBSOM 比起 SOM 擁有較佳之拓樸特性：不僅具備更良好之收斂特性，更能夠有效降低初始化狀況對於 cluster 結果的影響力。



(a)



(b)

圖 1　SOM(a)與 QBSOM(b)比較圖

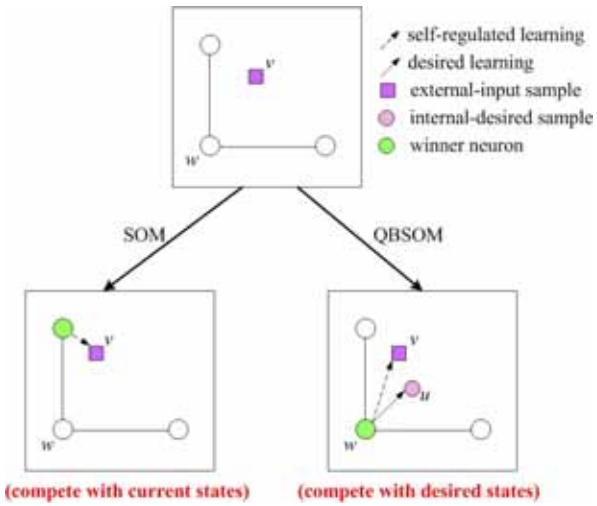


圖 2　SOM 與 QBSOM 幾何學意涵比較

為了瞭解 QBSOM 是否能如同 SOM 一般能夠適用於多種不同之基因表現資料，我們採用了 4 種基因表現的真實資料，分別為：Cho 等人於 1998 年提出之酵母菌(Yeast Cell Cycle)資料[21]、Eisen 等人的酵母菌完整資料[22]、West 等人的乳房腫瘤(Breast tumor)資料[23]以及 Alizadeh 等人的淋巴腺資料(Diffuse large B-cell lymphoma)[24]。針對這四種資料，我們使用 mean square error(MSE)與 Gibbon 等人所提之 Z-score 方法[25]進行分群結果的評估。實驗結果顯示，QBSOM 方法不僅收斂速度較傳統 SOM 快，在重複多次的實驗中，QBSOM 方法皆能穩定地產出優於 SOM 的結果，這顯示了 QBSOM 確實能有效降低初始化狀態的影響，分群能力比傳統 SOM 更加穩定。

三、結果與討論

實驗方式是以 QBSOM 與傳統 SOM 兩種方式進行評估比較。實驗資料則使用前述之 Yeast cell cycle, Yeast-all, Breast tumors 以及 Lymphoma 等四種資料集合。由於有時生物學家會將基因表現資料中差異較不顯著(例如: “flat” data)者，或是缺值過多的資料去除。這不僅在生物學上有其意義與目的（要研究比較 “interesting”的基因），也可以降低不相干資料影響對於分群結果的機會[13][26][27]。因此，我們也將 Yeast cell cycle 分成兩個集合，set1 進行資料過濾，過濾條件為：當 missing value 大於 20%，以及單一 pattern 中，最大值與最小值的差距小於 2 時，將資料去除。Set2 則保持完整的資料。如此一來，我們就可以看出在相同資料(Yeast cell cycle)下，資料過濾與否是否會影響 QBSOM 與 SOM 分群績效的差異狀況。而 Lymphoma 資料的部分，我們也用相同條件篩選出要進行分析的資料。其餘兩種資料，則只使用完整資料進行分析。

經過處理之後，四種大量資料均具有各自的特性：Yeast cell cycle 資料量大、資料維度較小，Yeast-all 資料量大、資料維度也很高，Breast tumors 的資料中不含 missing value，而 Lymphoma 資料量

大、資料維度中等，且包含了少量(小於 20%)的 missing value。此外，這四種資料本身的 topology 也必然不同。因此，QBSOM 與 SOM 在分析各種類型資料的適應性也能在實驗中看出。的所有資料均正規化(normalized)調整為-1 到 1 之間的值。在訓練參數部分，我們以一模一樣的條件(包括相同的 random initial weight 與相同的 input 順序)分別設定兩個系統，以確保實驗之有效性。實驗資料與相關參數設定如表 1 所示。

表 1：實驗資料集合與相關參數

Data set	# of genes	dimensions	Map size	iterations
Yeast cell cycle (set 1)[21]	972	17	5*5	10,000
Yeast cell cycle (set 1)[21]	6178	17	8*8	25,000
Yeast-all[22]	6221	80	8*8	25,000
Breast tumors[23]	7130	49	10*10	30,000
Lymphoma[24]	8143	40	10*10	30,000

我們將實驗分為兩大部分，第一部份針對 Yeast cell cycle 之 set 1 進行重複實驗(20 次)，這部分主要目的是要驗證兩種方法在 20 次隨機初始化的情形下，分群結果穩定度的表現。除了比較 MSE 以外，我們另比較了 Gibbons 等人所提之 Z-score[25]，Z-score 是用來衡量分群效果的指標，其主要原理是計算 *Saccharomyces* Genome Database (SGD)中的 gene annotations 與分群結果之關連性，得分愈高表示分群效果愈佳。實驗的第二部分則是針對資料量與資料維度較大的資料進行實驗，在這種較為「艱困」的條件下，系統收斂速度不僅影響生物學家實驗的效率，也間接影響了系統的 scalability。我們將綜合比較兩種系統在四個大量資料的 MSE 收斂速度與 MSE 值。為了實驗比較的方便，也為了達成計畫目標，我們以 Java 開發 SOM 與 QBSOM 之基因表現資料叢集分析軟體。

分析結果除了可產生叢集結果之文字檔以外，尚可以圖形顯示結果，如下圖 3 所示，即為以 data set 1 並使用 QBSOM 方法進行分群的結果。藍色線段為 input vector，紅色線為 weight vector（即 cluster 中心）。Cho 將酵母菌的 cell cycle 分成五階段：early G1、 S、G2、late G1 以及 M[21]，這五種階段的基因均有不同的特徵。圖 3 中第一列第四行的 neuron 即代表屬於 Late G1 的 cluster[9][21]。SOM 與 QBSOM 這種易於視覺化表示的特性，確實使得資料更易於解釋。

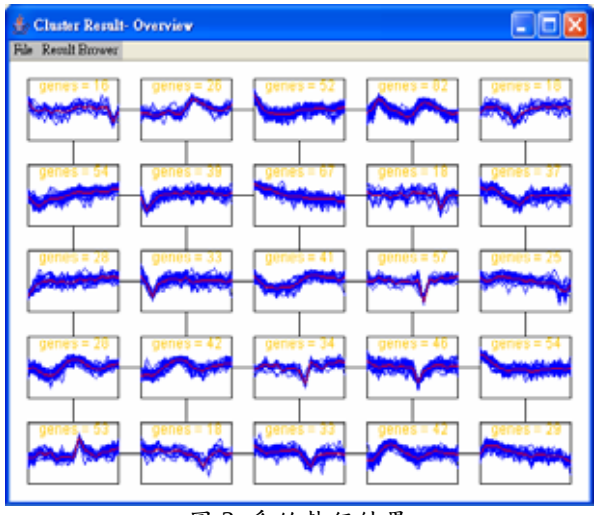


圖 3 系統執行結果

A. Evaluate initialization independency

我們使用 yeast cell cycle 進行重複實驗，通過過濾條件的資料比數為 972 筆，根據[3][19]，在相似的篩選條件與資料筆數下，較合理的資料分群群組數應當約為 20-25 組，因此，我們設定 5*5 的 Map 大小，網路拓樸為 mesh，另外，訓練次數設定為 10,000 次，即約為 10 個 epoch。為了控制變因，實驗方式為隨機產生 20 組不同的 initial weight 與 input 次序，使得在 20 次的實驗中，SOM 與 QBSOM 均能夠使用相同的初始化條件。兩系統之 learning rate 均設定為 0.5，以 linear stepwise 方式遞減。QBSOM 之 query rate 設定為 0.8，以指數方式

遞減。實驗結果如下圖 4 與圖 5 所示。

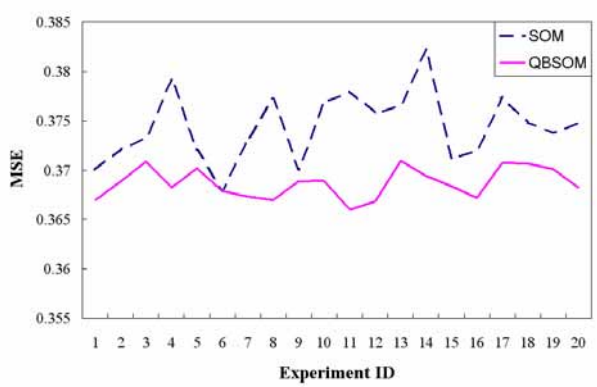


圖 4 實驗 20 次之 MSE 比較

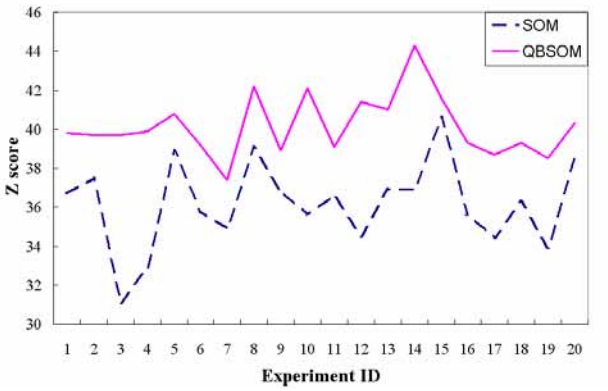


圖 5 實驗 20 次之 Z-score 比較

由圖 4 與圖 5 可知，在 20 次的重複實驗中，QBSOM 在 MSE 與 Z-score 均明顯優於傳統 SOM。即使在實驗 6，SOM 的分類結果最佳時(MSE 最低)，QBSOM 之 MSE 仍相當於 SOM(均為 0.368)。此外，不論是 MSE 或是 Z-score，QBSOM 的表現均較 SOM 穩定。

B. 不同資料之比較

基因表現資料分析的一大挑戰是如何有效率地處理大量且高維度的資料。傳統 SOM 在這方面的表現已是相當優異[9-12][25]。為了測試 QBSOM 是否能承襲 SOM 在這方面之優點，我們使用四種不同的資料進行實驗，實驗資料筆數為 6000-8000 筆，資料維度則從 17-80 個維度不等。我們比較兩種系統在大量資料下 MSE 的收斂狀況與分群完成後的 MSE 值。圖 6-圖 9 為兩系統在不同資料的 MSE 收斂狀況比較。

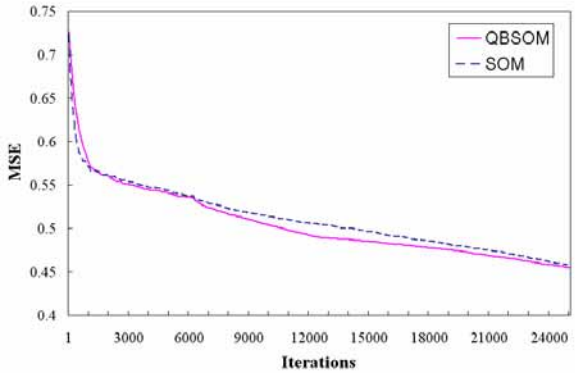


圖 6 使用完整 yeast cell cycle 資料

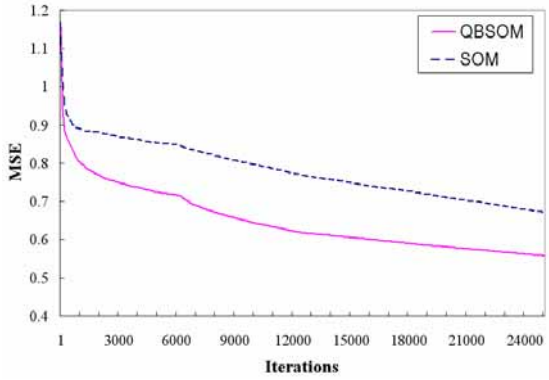


圖 7 使用 yeast all 資料

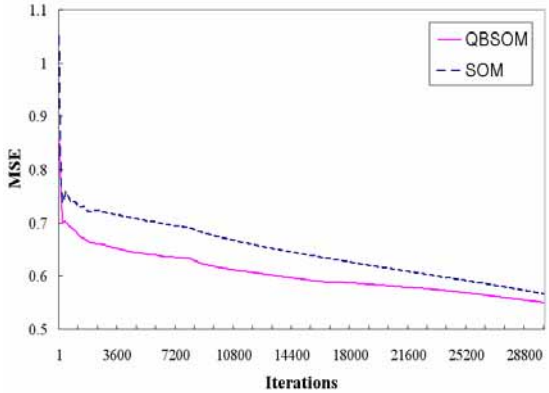


圖 8 使用 Lymphoma 資料

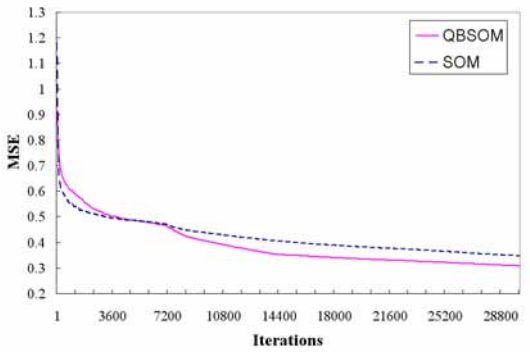


圖 9 使用 breast tumor 資料

由圖 6-圖 9 可知，QBSOM 在四種不同資料集合下，收斂速度都較 SOM 快，此外，MSE 之最終值亦均較 SOM 更低。值得注意的是，QBSOM 在四個圖中，均具有了兩個轉折點，使得其學習曲線看起來是有三個階段：第一階段 Query-Based learning 的作用最強(query rate=0.8)，而歷經了 1 個 epoch 之後，query rate 為 0.32，此時，Query-Based learning 與傳統 SOM 的 competitive learning 的影響力相去不遠，在第 3 個 epoch 之後，傳統 SOM 的學習機制便發揮了絕對的影響力，Query-Based learning 的影響在此後會趨近於 0。我們將兩個系統在四種不同資料的 MSE 終值彙整於圖 10。更進一步地分析，我們發現，當資料維度愈高時，QBSOM 的表現就會愈佳，例如在使用 yeast 資料(80 維)時，QBSOM 的表現遠優於 SOM。

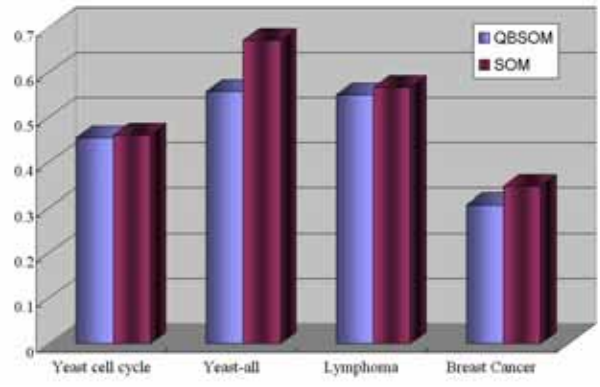


圖 10 兩系統在四種不同資料之 MSE 比較

四、計畫成果自評

本實驗結果證明所提出之 QBSOM 方法相較於先前之傳統 SOM 方法，能夠有效地降低網路初始化狀態對於分群結果的影響，亦具有收斂速度較快的優勢。此外，當資料量愈高、資料維度愈高時，這種優勢就更加明顯。我們已針對此一方法設計出一套分群分析的工具，將可提供生物學家多一個分析工具的選擇。此外由實驗結果可知，QBSOM 的確是相當具有發展價值的方法，未來我們將持續探索相關問題，例如：如何決定最佳的分群組數等；同時，我們亦將持續改良此一方法，期望能做出更強力的基因表現資料分析工具軟體。

五、參考文獻

1. http://www.biochipmaster.com/
2. S. Granjeaud, F. Bertucci, and B. R. Jordan, “Expression profiling: DNA arrays in many guises,” *BioEssays*, vol.21, 1999, pp.781-790.
3. A. V. Lukashin and R. Fuchs, “Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal numbers of clusters,” *Bioinformatics*, vol. 17, no.5, 2001, pp.405-414.
4. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, 1998, pp. 14863-14868.
5. P. T. Spellman, G. Sherlock, M. Q. Zhang, D. Botstein, and B. Futcher, “Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization,” *Mol. Biol. Cell*, vol. 9, pp. 3273-3297.
6. M. Takahashi, D. R. Rhodes, K. A. Furge, H. Kanayama, S. Kagawa, and B. B. Haab, “Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic

classification," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, 2001, pp. 9754-9759.

7. R. Herwig, A. J. Poustka, C. Müller, C. Bull, H. Lehrach, and J. O'Brien, "Large-scale clustering of cDNA-fingerprinting data," *Genome Research*, vol. 9, 1999, pp. 1093-1105.
8. S. Tavazoie, J. Hughes, M. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat. Genet.*, vol. 22, 1999, pp. 281-285.
9. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, 1999, pp. 2907-2912.
10. P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," *FEBS Letters*, vol. 451, 1999, pp. 142-146.
11. J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong, "Analysis and visualization of gene expression data using Self-Organizing Maps," *Neural Networks*, vol. 15(2002 special issue), 2002, pp. 953-966.
12. K. Torkkola, R. M. Gardner, K. K. Tamma, and C. Ma, "Self-organizing maps in mining gene expression data," *Information Sciences*, vol. 139, 2001, pp. 79-96.
13. D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," *IEEE Trans. on Knowledge and Data Engineering*, vol.16, no. 11, 2004, pp.1370-1386.
14. J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns," *Bioinformatics*, vol. 17, no. 2, 2001, pp. 126-136.
15. GEPAS: http://gepas.bioinfo.cnio.es/tools.html, [online available].
16. Cluster: http://rana.lbl.gov/EisenSoftware.htm, [online available]
17. GeneCluster2: http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html, [online available]
18. ArrayMiner: http://www.optimaldesign.com.
19. S. Wu, A.W.-C. Liew, H. Yan, and M. Yang, "Cluster analysis of gene expression data based on self-splitting and merging competitive learning," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8 , issue: 1 , 2004, pp. 5 – 15.
20. Ray-I Chang and Pei-Yung Hsiao, "Unsupervised query-based learning of neural networks using selective-attention and self-regulation," *IEEE Trans. Neural Networks*, vol.8, no.2, March 1997, pp:205-217.
21. R. J. Cho and *et al*. "A Genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, 1998, pp. 65-73.
22. Eisen's Lab: http://rana.lbl.gov/EisenData.htm
23. M. West and *et al*, "Predicting the clinical status of human breast cancer by using gene expression profiles," *PNAS*, vol. 98, no. 20, 2001, pp. 11462-11467.
24. A. A. Alizadeh and *et al*. "Dinstinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, 2000, pp. 503-511.
25. F.D. Gibbons and F. P. Roth, "Judging the Quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, issue 10, 2002, pp. 1909-1916.
26. S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, 2003, pp. 2088-2096.
27. Y. Moreau, F. De Smet, G. Thijs, K. Marchal,

B. De Moor, “Functional bioinformatics of microarray data: from expression to regulation,” *Proceedings of the IEEE*, vol. 90, issue 11, 2002, pp. 1722-1743.