行政院國家科學委員會專題研究計畫　成果報告

蛋白質二維凝膠電泳圖影像處理進階研究：分析與大量自動比對

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-216-016-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：中華大學資訊工程學系

計畫主持人：林道通

計畫參與人員：李訊忠　黃巧妏　廖冠智　黃耀鋒

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94 年 8 月 1 日

行政院國家科學委員會專題研究計畫報告

蛋白質二維凝膠電泳圖影像處理進階研究：分析與大量自動比對(I)

Advanced Research on Protein 2D Gel Electrophoresis Images: Analysis and Large Scale Matching without Landmarks

計畫編號- NSC93-2213-E-216-016
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：林道通　中華大學資訊工程系暨研究所
計畫參與人員：李訊忠 黃巧妏 廖冠智 黃耀鋒　中華大學資訊工程研究所
黃三元　台灣動物科技研究所
林恩仲　台灣大學畜產學系暨研究所

一、中文摘要

二維蛋白質凝膠電泳圖在生物資訊學蛋白質研究領域上扮演著相當重要的角色，我們預計以兩年的時間完成蛋白質二維電泳圖分析系統包含(1)前處理、(2)自動點偵測、(3)大量蛋白質凝膠電泳圖比對、(4)影像資料庫建立。第一年著重在電泳圖的校準與自動點偵測，在本計畫中，我們已提出一項改良式分水嶺演算法對原始影像作影像切割，以適應性的法則來調整分水嶺的閥值，融入標籤法及區域成長法偵測出蛋白質電泳圖上的蛋白質質點。我們也採用方向性分水嶺圖示法及型態學運算來改善過度切割的問題。我們已經收集十五張(大小為 1498 x 1544)豬隻睪丸的二維蛋白質凝膠電泳圖並進行測試，結果相當正面及可行。我們預計建立的資料庫是以豬隻繁殖力功能性相關之蛋白質二維電泳圖影像為主，並結合台灣動物科技研究所，建立一個以豬的繁殖力研究為主軸的功能性基因組及蛋白體資料庫。

關鍵詞：改良式水嶺演算法，點偵測，蛋白質電泳膠片，影像切斷，形態學運算，區域成長
。

Abstract

Two-dimensional gel electrophoresis is one of the most important techniques in proteomics research for subsequent comparison and analysis. In this project, the proposed 2DGE analysis system consists of four major subtasks: (1) pre-processing, (2) automatic spot detection, (3) large scale spot matching, and (4) image data base set up. Spots identification is among the first things of analyzing Two-dimensional gel electrophoresis images. In the first year, we provided an improved solution to the problem of protein spot detection by applying directed graph watershed transform algorithm. We have developed an adaptive mechanism to adjust the level of detail and determine the threshold value of watershed. The over-segmentation drawback is overcome by applying directed graph version of watershed transform algorithm and morphological opening operation. Labeling and region growing techniques are adopted to extracted individual spots features. The proposed spot detection process has been implemented and tested on 15 protein gel profiles (image size: 1498 x 1544) of porcine testis. The detection results are promising.

Keywords : Improved watershed algorithm, Spots detection, Protein 2D gel images, Segmentation, Morphology operation, Region growing..

二、緣由與目的

在蛋白質的研究領域中，二維蛋白質層析電泳圖是處理蛋白質定量、定性分析中一個非常重要的工具，而在生物學家分析的過程中，需要強大且自動化的分析技術去偵測、辨識電泳影像間的差異性，其中最重要就是要有快速、健全的影像分析、比對的軟體工具([1]～[3])。因此我們針對國內的需求及蛋白質工程的基礎建設，配合與台灣動物科技研究所的合作，提出這項計畫設計出全球第一個以豬隻的蛋白質影像分析系統來研究豬的繁殖能力，希望藉此能促進蛋白質工程的發展。

目前市面上有許多已經發展好的商用軟體提供生物學家分析電泳影像，例如Melanie、Z3…等([4]～[12]、[16])，這些大部分也只能處理一些一般化的資料，然而例如影像偏移、損毀等的狀況，甚至是蛋白質質點的重疊問題，都是目前尚待解決的[12]。因此我們針對這些目的及問題，提出我們的系統架構，並使用模糊推論的機制，加強比對的可信度。在現有處理二維蛋白質電泳圖的技術，大多傾向在電泳圖的前置處理及統計的過程，前者包含修復扭曲、破損的影像，後者多在抽取電泳圖上的點及統計個數，並以直方圖呈現。在整個影像處理階段中，最重要的兩項就是質點的偵測與比對。我們的目的就是要發展一套大量影像分析軟體系統。包含三個基本的步驟：(i)點偵測，(ii) 膠片比對，(iii)資料庫建立及搜尋。在第一步驟，點偵測，決定膠片影像特定的特徵是不是一個蛋白質質點。點偵測的良窳是用自動偵測法則與使用人工去辨識真實的蛋白質質點來作比較，軟體是否能夠偵測到越多真實的蛋白質質點以及最少的誤判將視為判斷好壞的準則。在第二步驟，膠片比對，是使用點比對或者是特徵比對的方式。為了評估膠片比對的能力，我們將使用人為變形的膠片影像跟原本未變形的原始影像來做比較，在處理完 Spot 偵測之後，比對的工作才是在分析電泳圖中作重要的部分，在一般市面軟體中，多半使用 Landmark 的方式來作比對的標準（例如 PDI 公司的 2-D GEL)，而我們現在所提出的方法是以 MRST（Minimal Relation Spanning Tree）自動比對的概念[13]來改進、修正前述的方法，可以達成更有效率及正確的比對。

上述這些方法都只做到前端的處理，並不能完全的表示生物體的資訊，因此需要將生物及醫學方面的知識融入系統中，才能夠發揮出最好的效用。然而為了要產生對樣本間的差異性蛋白質表現有意義的資料，這樣的發現仍然需要依賴大量的 2-D 膠片影像作正確的比對以及比較。最後，我們將蛋白質的電泳圖的資訊完整的儲存在資料庫中，以供資料保存及日後參考，不但可以藉此了解經年累月來的實驗過程以利技術檢討，也可藉此增強在處理二維電泳圖的經驗與知識。因此我們最後會將結果以 WEB 網站的方式呈現，一方面實驗時可以參考多方面的資料，另一方面也能藉由每一次實驗的結果及經驗，作為下次實驗的依據，使實驗的正確性提高。

三、計劃執行結果與討論

第一年計畫執行步驟包含：來源影像前置矯正、電泳圖的蛋白質偵測、統計蛋白質的數量，其執行方法及結果如下：

（一）影像校準

影像校準是整個2D膠片影像分析的重要處理程序，我們採用Multiresolution Image Registration[7]當作前處理校準的方法。這個方法採用了一個從粗糙到平滑的特徵比對。假設我們要比較兩個影像S跟T_n，S是標準的膠片影像、T_n是目標影像乃由某一個轉換函式F構成。如此一來S跟F(T_n)間的相似性可由函數Sim來評估。關係函數定義為兩者的關連：$Sim = corr(S, F(T_n))$。我們選擇了Piecewise Bilinear Mapping(PBM)轉換來表

現兩者變形的關連。一個PBM是由格子構成的（如圖一所示）。為了要建立這樣的格子，目標影像首先要分割成$2^k \times 2^k$的小方塊，k是對應轉換的高度。給定一個高度以及一個格子的索引(i, j)，來源影像S所對應的點分別為$s_{i,j}$，$s_{i+1,j}$，$s_{i,j+1}$，$s_{i+1,j+1}$，目標影像的控制點為$t_{i,j}$，$t_{i+1,j}$，$t_{i,j+1}$，$t_{i+1,j+1}$，將會定義一個對應函數m。點c位於$t_{i,j}$，$t_{i+1,j}$，$t_{i,j+1}$，$t_{i+1,j+1}$ 的方塊當中，目標影像T_n將會根據原始影像依照相對應的控制點對應到m(c)。

$$m(c) = w_{i,j}(a,b)s_{i,j} + w_{i+1,j}(a,b)s_{i+1,j} + w_{i,j+1}(a,b)s_{i,j+1} + w_{i+1,j+1}(a,b)s_{i+1,j+1}$$

其中 $w_{i,j}(a,b) = (1-a)(1-b)$ 、

$w_{i+1,j}(a,b) = a(1-b)$ 、 $w_{i,j+1}(a,b) = (1-a)b$ 、

$w_{i+1,j+1}(a,b) = ab$ 。a 跟b的值分別是點c的水平以及垂直的比例，如圖一所示。

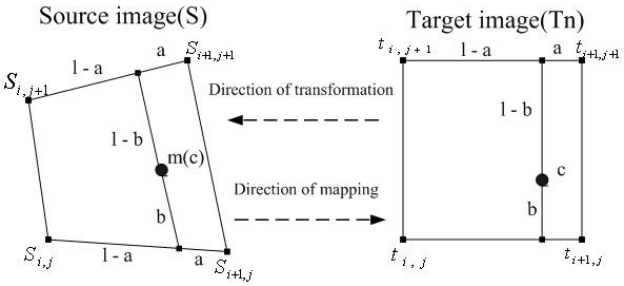
兩個密度以及分佈的交互關係定義成偏微分$\sigma(S)$，$\sigma(T)$ 及$\mathrm{cov}(S,T)$：

$$corr(S,T) = \frac{\mathrm{cov}(S,T)}{\sigma(S)\sigma(T)}$$

$$\mathrm{cov}(S,T) = \frac{1}{|D|_D}\int (S(x)-\bar{S})(T(x)-\bar{T}))dx$$

corr() 是計算兩張膠片影像的相似度函數，cov()是計算兩張膠片影像的covariance。

進行流程一開始是從一個粗糙（低解析度）的層級開始，接著在目標影像中的這些格子中找到一個最大的相似區塊。當關係式corr得到最佳化的結果時，目標影像的level of detail將會增加。經過這個演算法處理數次到最高解析度之後，我們可以得到最後的結果。經過影像校準的處理之後，來源影像將會變形並且影像的尺寸大小將會跟目標影像一樣。接下來我們就可以進行點偵測以及直接比對的動作然後可以去比較兩張膠片影像的蛋白質質點。

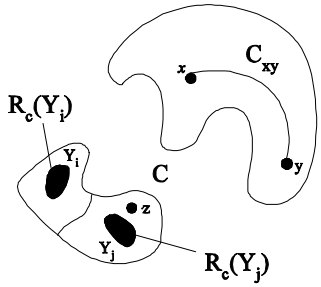


圖一、Piecewise bilinear mapping[28]

（二）影像切割與蛋白質質點偵測

我們現階段的處理重心在影像處理及膠片比對上，我們先將由台灣動物科技研究所提供的二維蛋白質電泳圖輸入至我們的系統中，在依照圖中各項流程處理，最後將蛋白質質點資訊及比對結果輸入至資料庫中儲存。影像取得之後，首要工作就是要將蛋白質質點由影像中找尋出來，我們可以將一張灰階圖形視為一個不平坦的拓樸分佈。這個拓樸分佈的地圖包含了很多的輪廓線並且我們可以在這個拓樸裡面找到最大以及最小的灰階值。一個輪廓 C 如圖二所示，假如 X 跟 Y 是落在 C 這個範圍裡面，可以找到兩個點 X 跟 Y 來定義它們之間的距離。假如兩個點落在不同的輪廓區域當中（如 X 跟 Z 兩點），他們的距離將會被定義成無限大。在兩個點之間，距離可以被表示成

$$d(x,y) = l(C_{xy}), \quad d(x,z) = +\infty$$

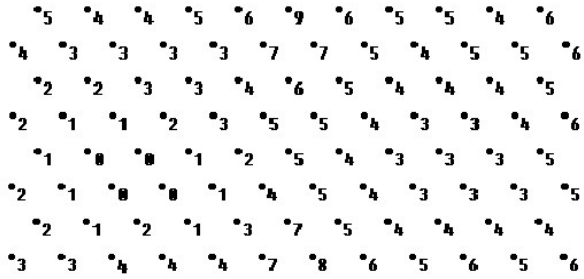


圖二、定義相同區域之兩點距離

因此這個拓樸地圖可以被轉換到一個階層式的灰階影像。傳統的分水嶺轉換法可以分成兩類[14]：模擬洪氾的處理過程，以及著重直接偵測分水嶺的位址。在本研究中我們將結合第二類的演算法來改進分水嶺轉換。首

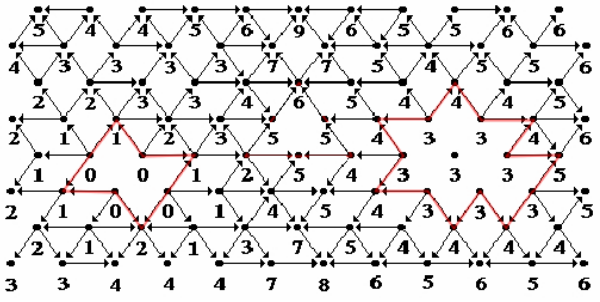
先我們用像素分布來處理膠片影像，如圖三所示。強度較低的像素較小數值；換句話說大的數字代表它的像素是比較黑的。接著我們計算鄰居點的像素強度關係。假如強度比它的鄰居點弱，我們將箭頭指向它的鄰居。否則箭頭將會指向自己。假如它們的強度相同的話將不會有任何箭頭。圖四是圖三經過加上箭頭的結果。根據第二步驟的結果(如圖五所示)，我們可以在影像裡找到兩個獨立的區域。因此我們設定了一個標準當成是臨界值去擴張直到一個可以當作分水嶺的分割線出現。用上述方法做出來的分水嶺區域如圖五中間方塊標記所示。

先計算出影像中各像素點的質，如下圖：



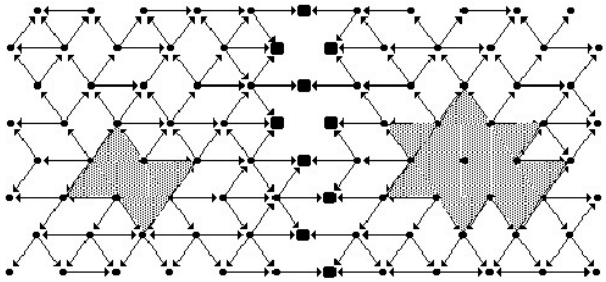
圖三：影像像素分佈圖

再計算每個點對周圍鄰點的關係：若是比周圍點小則指標指向周圍，若周圍點較小，則指標指向自己。利用這種關係，就可以將影像形成像下圖一樣的結果。



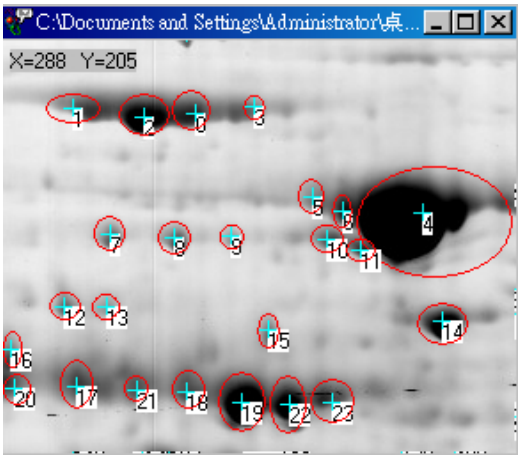
圖四：像素關係圖

利用上圖的結果，我們可以發現兩個區塊是獨立的（無指標朝內），我們就可以藉由此兩塊區域向外作某程度（設臨界值）的擴張，就可以找到下圖中間由較大黑點所形成的分

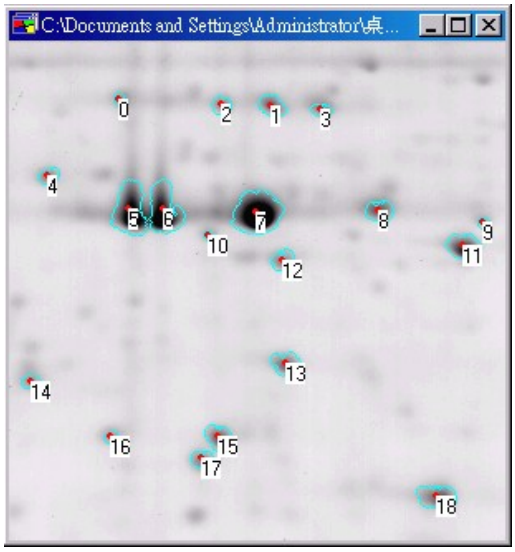


圖五：分水嶺示意圖

隔線，此分隔線就是分水嶺。然後再利用區域成長演算法將每一個被切割出來的質點編號，並標示每個蛋白質質點的區域及中心座標，其切割、偵測的結果如下圖六所示。



(a) 蛋白質質點偵測結果 Case001



(b) 蛋白質質點偵測結果 Case002

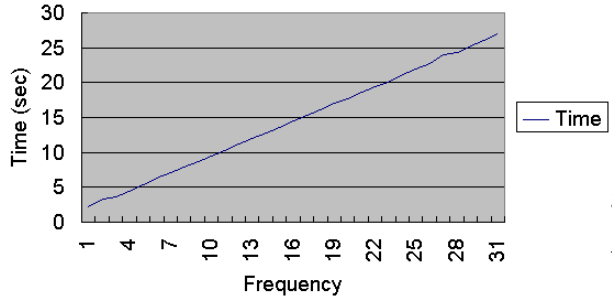
圖六：兩個蛋白質質點偵測結果

利用這種點偵測的方式，可以避免蛋白質質點重疊的問題，而且可以降低條紋雜訊的影響程度，切割出較佳的影像。以適應性的法則來調整分水嶺的閥值，融入標籤法及

遞迴區域成長法偵測出蛋白質電泳圖上的蛋白質質點，其演算法如下：

```
region Growing()
{
    if (there is no pixel meets the same
condition) then break;
    else Label this pixel;
    Region Growing();
}
```

我們發現此程式執行的速度和型態學運算"opening"運算次數成正比，如下圖所示。



圖七：程式執行的速度和型態學算"opening"運算次數之比例。

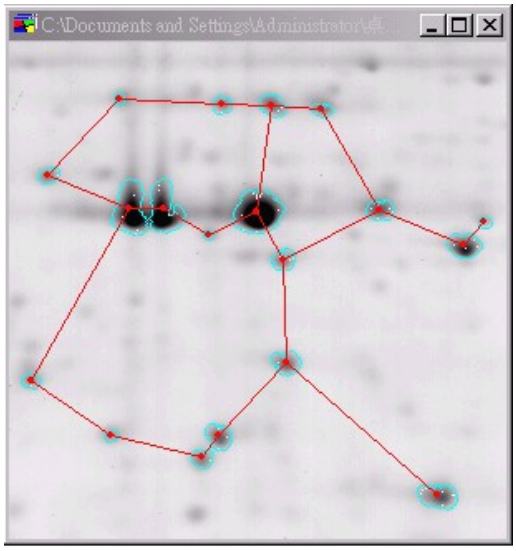
由於蛋白質電泳圖上的蛋白質質點並沒有統一的特徵，因此我們無法藉由建立模組來辨別質點間的差異性。在此我們使用相鄰關係圖形（relative neighborhood graph, RNG）及加百利圖形（gabriel graph, GG）[20]，以幾何平面連通圖形建立比對用特徵的方法。其演算法分別如下：

（1） RNG 圖形：

決定 RNG 圖形 N 個點的集合中若任兩點 a, b 間的距離小於 $a,\ b$ 到其他任何點的距離，則將 a, b 連線，其演算法如下：

$$\forall a,b \in N,\ \ and\ \ a,b\ \ is\ \ connected,$$
$$if\ \ only\ \ if\ \ d(a,b) < \left\langle \max\{d(a,c),\ d(b,c) \middle| \forall c \in N\ \ and\ \ c \neq a,\ \ b\} \right\rangle$$

如此便可得到 RNG 圖形，換言之其中圖形點與點間的連線不會交錯，是一個平面連通圖。如下圖所示，RNG 圖形的連線狀態。



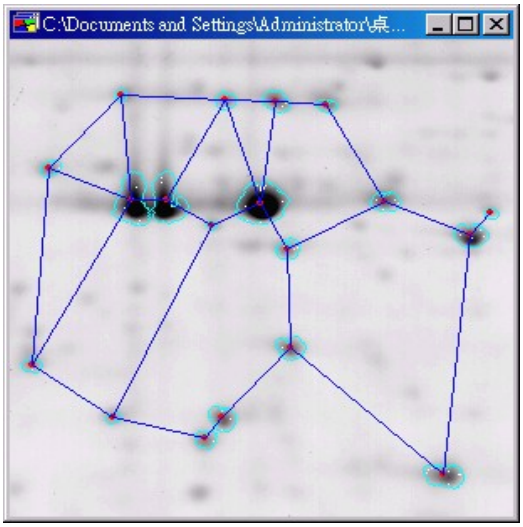
圖八：RNG 圖形示意圖。

（2） GG 圖形：

決定 GG 圖形中任一兩點 a, b 間的連線的演算法如下：

$$\forall a,b \in N,\ \ and\ \ a,b\ \ is\ \ connected,$$
$$if\ \ only\ \ if\ \ d(a,b) < \min\{\sqrt{d^2(a,c)+d^2(b,c)}, \forall c \in N\ \ and\ \ c \neq a,\ \ b\}$$

便可得到 GG 圖形，也是一個平面連通圖。如下圖所示，GG 圖形的連線狀態。



圖九：GG 圖形示意圖

由於這種圖形在同樣的影像中，並不會因為影像的縮放、偏移和旋轉的改變而有所影響，因此在特徵選擇上具有唯一及不變性，適合我們用來比對膠片間的差異。而我們所以選擇 RNG 及 GG 兩種圖形，除了可以

滿足上面兩點之外，他們彼此之間具有集合的關係，由上面圖形所示，可以清楚發現 RNG 圖形是 GG 圖形的子圖所以我們以 GG 圖形為主，RNG 圖形為輔作比較，因 GG 在分支度的呈現上也比 RNG 來的多樣性，所以我們在比對的時候使用 GG 圖形當作比對路徑的基礎模型。

四、計畫成果自評與未來工作方向

一、第一年計畫已完成項目：

（一）資料收集：已收集 15 張（1498*1544）的蛋白質二維電泳圖影像；

（二）完成來源影像前置矯正；

（三）完成電泳圖蛋白質點偵測成式，目前測試正確率達 92%以；

（四）進行統計蛋白質的數量；

（五）持續測試與驗證。

此計畫目前執行了一年，現在的結果還不錯，其中幾項不錯的結果可歸納餘下：

- 可切割複雜區域；
- 蛋白質質點標示清楚；
- 特徵值維持單一不變性；
- 推論系統擁有區域可調性；

就目前的影像處理階段而言，結果大致良好，預期建立完整的資料庫系統可以增進在蛋白質工程領域上的研究進度，一方面可將生物學家的研究成果完整的保存；另一方面藉由交互參考不同的資料庫，可以互補有無，豐富各研究機構間的研究資源，促進生物資訊工程發展。

五、參考文獻

[1] D. Bollag, M. Rozycki, S. Edelstein, Protein Methods, Second Edition, Wiley, New York, NY, 1996.

[2] W. Patton, “Review: Bilolgist’s perspective on analytical imageing systems as applied to protein gel electrophoresis ,” ELSEVIER, 1995. Journal of Chromatography A. 698, pp. 55-87.

[3] P. Nugues, “Two-Dimensional Electrophoresis Image Interpretation,” IEEE Trans. On Biomedical Engineering, 1993. vol. 40, pp. 760-770.

[4] Q. Wang, K. Mattila, H. Frey and Y. Neuvo, “Image Processing Applied to the Analysis of Electrophoresis Gels,” China 1991 International Conference on Circuits and Systems, 1991. Shenzhen, China, pp. 789-792.

[5] A. Thompson and T. Brotherton, “Information Extraction from 2D Electrophoresis Images ,” Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society , 1998. Vol. 20, No 2, pp. 1060-1063.

[6] H. Noordmans and A. Smeulders “Detection and Characterization of Isolated and Overlap Spots,” Computer Vision and Image Understanding, April, 1998. vOL. 70, No. 1, pp. 23-35.

[7] S. Veeser, M. Doun, G.-Z. Yang, “Multiresolution image registration two-dimensional gel electrophoresis,” Proteomics, 1, 2001. pp. 856-870.

[8] Y. Xiu, D. Feng, H. Hong, “Novel Elastic Registration For 2-D Medical And Gel Protein Images.”, First Asia Paci£c Bioinformatics Conference (APBC2003) , Adelaide, Australia Conferences in Research and Practice in Information Technology, 19. Chen, Y.-P. P., Ed. ACS. 223-226.

[9] Z. Smilansky, “Automatic Registration for Images of Two-dimensional Protein Gels.”, , Electrophoresis , 2001. 22, 1616-1626.

[10] Y. Watanabe, K. Takahasi, M. Nakazawa, “Automated Detection and Matching of Spots in Autoradiogram Images of Two- Dimensional Electrophoresis for High-speed Genome Scanning,” IEEE , 1997. 496-499.

[11] P. Zhou, D. Pycock, “Robust statistical models for cell image interpretation,” ELSEVIER: Image and Computing, vol 15, pp. 307-316, 1997.

[12] H. Noordmans, “Detection and Characterization of Lsolated and Overlapping Spots ,” Computer Vision and Image Understanding, 4, 1998. Vol 70, No. 1, pp. 23-35.

[13] L. Vincent and P. Soille, “Watersheds in Digital Spaces: An Ef£cient Algorithm Based on Immersion Simulations ,” IEEE Trans. On Pattern Analysis and Machine Intelligence, 6, 1991. Vol. 13, No. 6 ,pp. 583-598.

[14]S. Beucher, “The Watershed Transformation Applied to Image Segmentation,” Signal and Image Processing in Microscopy and Microanalysis, 6, 1992. pp. 299-314.

[15] A. Moga and M. Gabbouj, “Parallel Image Component Labeling With Watershed Transformation,” IEEE Trans. On Pattern Analysis and Machine Intelligence, 5, 1997, Vol. 19, No. 5 ,pp. 441-450.

[16] E. Bettens, P. Scheunders, J. Sijbers, D. Van Dyck, L. Moens, “Automatic Segmentation and Modelling of Two-Dimensional Electrophoresis Gels,” IEEE , 665-668, 1996.

[17] J. Jaromczyk and G. Toussaint, “Relative Neighborhood Graphs and Their Relatives.”, Proceedings of the IEEE, 80, 9, 1992, pp. 1502-1517.

[19] ,J. Garrels, ”The Quest system for quantitative analysis of two-dimensional gels”, Journal of Biology Chemistry, vol. 264, pp. 5269-5282,1989.

第一年進度甘梯圖

月 次 工作項目	第1月	第2月	第3月	第4月	第5月	第6月	第7月	第8月	第9月	第10月	第11月	第12月	備註
適用多種規格電泳圖	V	V											完成
圖形定位	V	V											完成
移除條紋及背景校正		V	V										完成
區域性的處理		V	V										完成
批次處理			V	V									完成
Spot 數量			V	V									完成
Spot 偵測				V	V	V							完成
Spot 切割				V	V	V							完成
過濾 Spot				V	V	V	V						完成
編輯標準組影像				V	V	V							完成
參考組影像平均						V	V	V	V				完成
Spot 統計						V	V	V	V				完成
與其他實驗結果比較								V	V	V	V	V	
期末報告											V	V	
進度累計百分比(%)	5	15	25	40	50	65	70	78	85	88	93	100	