應用VQ影像技術於蛋白質二維凝膠電泳影像快速濾除之研究

陳同孝 何英治 謝明麗 蔡蕙芳 蔡君微*
台中技術學院 中山醫學大學 東海大學 中山醫學大學 台中技術學院
資訊管理系 資訊管理系 生物系 醫事技術學系 資訊管理系
*通訊作者 p3632@thcb.e-building.net.tw

摘要

目前DNA定序工作已完成，接下來要針對DNA做功能性了解、分析，生物資訊從此走進了「功能性基因體時代」。蛋白質二維凝膠電泳影像上所顯示的資料，就是生物體內蛋白質的含量及分佈之情況。目前蛋白質二維凝膠電泳影像比對工作多半是使用或是藉助軟體工具將欲查詢影像和資料庫中蛋白質二維凝膠電泳影像逐一以重疊方式進行比對，這過程尚需以人工方式進行手動做些微移動調整，比對後取出相異的蛋白質進行質譜分析。在蛋白質資料庫中，所面對的是龐大蛋白質影像資料量，使用半人工半自動方式進行比對，是十分沒有效率的。蛋白質二維凝膠電泳影像特有的特徵值運用下，可以利用文章中的技術濾除不相同細胞所產生之蛋白質二維電泳凝膠影像，未來就不再需要進行詳細比對程序。便可省去不必要的比對時間，提高比對時效。運用本研究的方法相信未來必可大幅提升生物學家在資料庫中比對蛋白質二維凝膠電泳影像的效率。

關鍵字：蛋白質二維凝膠電泳影像、質譜分析、濾除。

一、前言及研究背景

人體中所有的結構及活動都是由蛋白質分子所構成，因此蛋白質是組成人體最重要的成份，蛋白質的變異使人產生疾病，生物學家們需了解蛋白質功能，藉以提供基因表現圖譜分析、立疾病基因表達資料庫、找出與疾病相關之基因產物作為診斷及預診之用、確認某蛋白質其生理功能及是否參與某種疾病的致病機轉 提供作為治療藥劑(新藥)以及改良現有藥物研發之用。

生物學家們以蛋白質二維凝膠電泳影像來找出正常人與病人體內蛋白質變異情形，本研究以蛋白質二維凝膠電泳影像做為研究實驗基礎生化技術，蛋白質二維凝膠電泳影像研究重要流程如下：

(1)電泳技術取得蛋白質膠體—藉著蛋白質純化活性、分子大小、環境PH值、使用藥劑含量、使用藥劑濃度……等不同控制變因的組合，以電泳方式利用蛋白質分子帶正、負淨電荷的特性而自然的在二維凝膠上半固體狀介質中分散開來[04]、[05]。分子泳動的程度決定於分子形狀、大小、電荷密度、膠體濃度和所給予電壓[01]、[06]。

(2)膠體染色—電泳完後的凝膠尚需染色處理，才會表現出蛋白質電泳完呈現的色帶[07]。

(3)蛋白質樣本數位化影像取得—在取得電泳膠體後，接著就某一疾病要在正常細胞與異常細胞所產生的膠體之間找出兩者相異的蛋白質樣本點。在找出相異蛋白質樣本點的過程中，生物學家不可能以實際膠體來比較。再者當實驗累積到相當次數時，相對實驗相關資料、膠體數量將非常驚人，為了建立完整實驗記錄資料庫，詳細記錄實驗時間、實驗編號、實驗環境、實驗說明還需記錄實驗的膠體產物。因此，將蛋白質樣本電泳膠體數位化，除了便利記錄實驗的產物也便利比對。

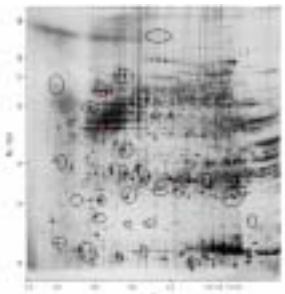
(4)影像特徵值擷取—生物學家們利用蛋白質的分子量、含量、濃度、帶電量來分辨蛋白質樣本點。蛋白質二維凝膠電泳影像中蛋白質樣本點顏色深淺可表達濃度的資訊，電泳實驗利用蛋白質帶電量的不同使蛋白質向四處泳動，因此，由蛋白質樣本點散開程度、分佈位置可以解釋帶電量和分子量，蛋白質樣本點大小可判斷含量。從蛋白質二維凝膠電泳影像中取出蛋白質樣本點大小、蛋白質樣本點數量、蛋白質樣本點位置和蛋白質樣本點顏色深淺四項影像特徵值便可了解是否為相同細胞所產生之影像。

(5)蛋白質樣本比對分析—以蛋白質二維凝膠電泳影像特徵值，經比對分析後將取出正常細胞蛋白質二維凝膠電泳影像與異常細胞蛋白質二維凝膠電泳影像中相異的蛋白質樣本點，再進行質譜分析，進而鑑別蛋白質身份，因此比對的正確性將影響後續的質譜分析。

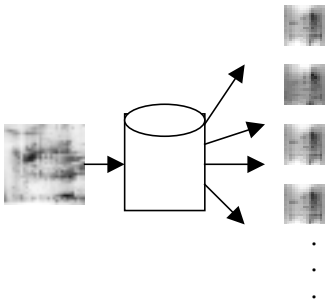
目前蛋白質二維凝膠電泳影像比對工作多半是使用或是藉助國外所開發之Melanie或Z3等工具軟體將欲查詢影像和資料庫中蛋白質二維凝膠電泳影像逐一以重疊方式先找出相同細胞或是相同疾病所產生的蛋白質二維凝膠電泳影像，再從中找尋相異的蛋白質樣本點。經比對軟體自動比對後發現即使是來自相同細胞或是相同疾病的影像，其中蛋白質樣本點有無法完全重疊的情況。因此，生物學家們在經比對軟體自動比對之後，還需以人工方式做些微移動調整。

由下圖一所示，圖一中以小圓圈圈起的部份蛋白質樣本點為軟體比對認為兩影像相似蛋白質樣本點，其他仍有非常多無法找到可重疊相對應沒有定義出來的樣本點。然而，這些未被定義的樣本點或許是兩影像實驗時造成的膠體扭曲，只需經過人工稍微調整就可被定義出。未來蛋白質二維凝膠電泳影像資料庫完成後，面對的是龐大資料量，若還使用半人工半自動方式進行比對，是十分沒有效率。如下圖二所示，若有一張二維凝膠影像要進入資料庫做比對，當資料庫裡有多張影像，則必須逐一做多次的半人工半自動比對。生物學家們應該省去比對過程的時間，移至比對後對人類更有意義的蛋白質結構分析、物理功能辨別的工作。為了協助生物學家們省去比對過程所花費的時間，我們在比對工作之前設計一個前置作業，可不影響比對準確性之下幫助生物學家們快速濾除與欲查

詢影像不相似的蛋白質二維凝膠電泳影像，最後只需針對濾除後所得相似蛋白質二維凝膠電泳影像，以Melanie[08]或Z3[09]等工具軟體執行軟體自動比對和人工調整。這樣必能減少無謂的比對時間，本研究提出簡單、快速且不需複雜運算，即能幫助生物學家們達到良好的濾除成效，大幅度提升比對效率是本研究最終目的。



圖一：圖中以小圓圈圈起的部份蛋白質樣本點為軟體比對認為兩影像相似蛋白質樣本點。



圖二：若有一影像要進入資料庫做相似性影像比對查詢，若資料庫內有多張影像則我們必須做多次半人工比對。

二、研究題目

本研究提出的方法可充份利用蛋白質二維凝膠電泳影像獨特的四項特徵—蛋白質樣本點大小、蛋白質樣本點顏色深淺、蛋白質樣本點位置、蛋白質樣本點數量，以及解決蛋白質二維凝膠電泳影像在處理和製作過程中可能因實驗環境不同造成的誤差—解析度不同、放大、縮小、拍攝亮度不一致、蛋白質二維凝膠電泳影像平移，達到不影響準確度下快速濾除蛋白質二維凝膠電泳影像資料庫中與欲查詢影像不相似的蛋白質二維凝膠電泳影像。朝兩大方向來考慮研究方法：

(1)考慮蛋白質二維凝膠電泳影像上蛋白質樣本點數量、樣本點大小、樣本點顏色深淺以樣本點位置四項特徵：

由於蛋白質是由二十種胺基酸分子以不同數目及排序聚合而成，而每種胺基酸所帶電荷不一，在特定酸鹼值時，幾乎沒有任何兩種蛋白質分子具有完全相同的帶電量與分子量。蛋白質二維凝膠電泳影像第一維以各蛋白質帶電量不同來分離，第二維以分子量不同來分離在電泳膠片上不同蛋白質。生物學家們利用蛋白質帶電量、分子量、濃度來鑑別每一個蛋白質的身份[12]。

觀察蛋白質二維凝膠電泳影像，由蛋白質樣本點散開程度和分佈的位置，可以了解蛋白質的分子量、所帶電荷、該蛋白質的等電點。由蛋白質樣本點顏色深淺，可以了解該蛋白質樣本點所含蛋白質濃度多寡。為了判別是否由相同種類蛋白質所產生之電泳影像，可由蛋白質樣本點數量、樣本點大小、樣本點顏色深淺以樣本點位置四項特徵的判讀來得到蛋白質帶電量、分子量、濃度。

經濾除後不相似的影像意味著，與欲查詢影像是由不同種類細胞蛋白質進行電泳實驗的產物。目前比對工作尚處於半自動半人工方式，因此，不再需要與欲查詢影像進行下一階段費時的人工細步重疊比對程序。

(2)考慮影像數位化過程中所發生的錯誤：

任何數位化影像在處理及製作過程中，都可能取得時拍攝或掃描解析度不同、影像大小不同、拍攝亮度不一致和影像平移產生誤差。蛋白質二維電泳膠體在數位化過程中，即使是相同的膠體在不同次的數位化的結果都不會相同。

本研究主要以影像VQ(Vector Quantization)向量量化壓縮技術做為基礎理論，再以本研究提出的計算方法可清楚表達出蛋白質二維凝膠電泳影像的資訊，求得專屬於每張蛋白質二維凝膠電泳影像一系列特徵值，做為影像濾除工作的依據。

三、研究步驟

VQ(Vector Quantization)向量壓縮[02][03]的作法首先將欲進行壓縮影像大小(S x S) 像素的影像以面積(N x N) 像素進行非重疊式分割，將會取得(M x M)個大小為(N x N) 像素的影像區塊。N表示每一個切割影像區塊的邊長，(M x M)表示當每一個影像區塊大小為(N x N) 像素的情形下整張(S x S) 像素的影像共可以切割出(M x M)個影像區塊。以S=128且N=4為例，切分割後將會得到M=(128/4=32)，(32 x 32)個面積為(4 x 4) 像素的影像區塊。按著事先以LBG演算法訓練完成的2^k位元的編碼書，而編碼書中每一個編碼字都有(N x N)個元素，稱之為(N x N)維。依序一次取一個分割好的影像區塊和編碼書中各編碼字一一來進行比較。運用下面距離公式將每一個影像區塊中各點和編碼書中對應的向量值相減，各項差值取絕對值平方之後加總，做為判斷距離小大的依據。VQ壓縮流程如圖三所示。

距離公式：

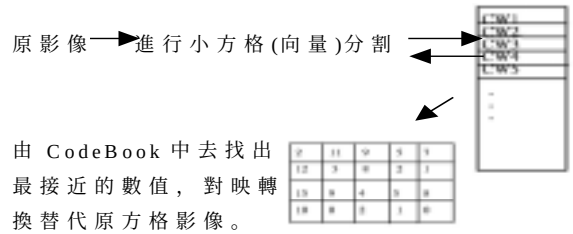
$$\text{Sum}=\sum_{i=1}^{N\times N}|\,Xi-Yi\,|^2 \qquad (1)$$

X_1至$X_{N\,x\,N}$表示壓縮影像任一個影像區中，從(N x N)個元素中取第i個元素。

Y_1至$Y_{N\,x\,N}$表示編碼書中任一個編碼字，從(N x N)個元素中取第i個元素。

由編碼書中找出跟每一個影像區塊最接近 距離最小的編碼字，做為此影像區塊的所代表數值(indexes)。最後，整張(S x S) 像素的影像將成為由(M x M)個數值組成一張索引表，如此便完成影像壓縮，而這張(Mx

M)的索引表即是影像經VQ向量量化壓縮後的產物[03][10][11]。



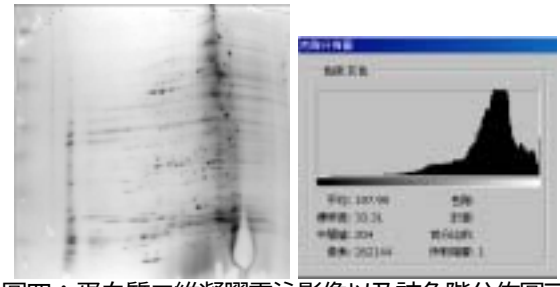
圖三：VQ壓縮流程步驟

以下各步驟將一步步詳細說明，本研究所提出的方法。

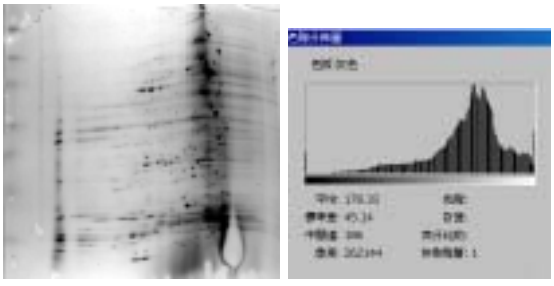
3.1 將蛋白質二維凝膠電泳影像進行亮度平均—解決拍攝亮度不一致

數位影像在拍攝時容易因拍攝亮度不一致導致影像顏色上差異，以繪圖軟體替蛋白質二維凝膠電泳影像平均亮度，將每張蛋白質二維凝膠電泳影像中最亮和最暗的像素定義為白色和黑色，然後再依比例來重新分配中等的像素值。在繪圖軟體中亮度平均化功能會裁掉 0.5% 的白色和黑色像素，也就是說，它在辨識影像中最亮和最暗的像素時，會忽略兩端的前 0.5% 確保白色和黑色值是以代表性的而非極端的像素值為基礎[13]。

將圖四平均亮度前的影像與圖五平均化亮度後影像做比較，可明顯看出圖五中蛋白質二維凝膠電泳影像中黑色的蛋白質樣本點被突顯出來，背景白色部份明顯轉為較白。圖四色階分佈圖中平均值和中間值與圖五色階分佈圖中平均值和中間值做比較，圖五中因平均亮度使得像素重新分配比例，黑色像素部份被突顯，使得像素值平均值與中間值降低。亮度偏亮的蛋白質二維凝膠電泳影像，蛋白質樣本點顏色偏灰，且微量蛋白質不易發現，這將造成比對誤差。以平均亮度，解決亮度不同造成比對誤差。



圖四：蛋白質二維凝膠電泳影像以及該色階分佈圖可以看出亮度上來說是偏亮。



圖五：蛋白質二維凝膠電泳影像以及該色階分佈圖明顯看出亮度上已平均化。

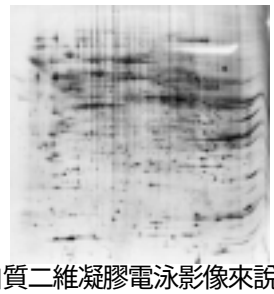
3.2 將蛋白質二維凝膠電泳影像進行VQ影像壓縮

取數張蛋白質二維凝膠電泳影像做為訓練影像，執行LBG演算法所得到2^k個位元的編碼書。蛋白質二維凝膠電泳影像為灰階圖，此2^k個位元的編碼書表示訓練影像上所有像素值經量化，分為2^k個群體後灰階分層階度，以0到(2^k-1)編碼字做為這2^k群像素值的代碼。以表一來說明編碼書，表一是以一個蛋白質二維凝膠電泳影像做為訓練影像執行LBG演算法後所產生的2^3位元編碼書，共有2^3=8個灰階分層階度，分別以0到7的編碼字表示之，每一個編碼字都有(N x N)=16維向量。並將(N x N)維向量值加總，以向量加總值大小進行排序，此排序的功用是為了本研究後續作業執行。排序的目的是將編碼書裡所有的灰階分層階度以黑色程度排列，以了解編碼書中灰階分層階度的相對差異。由表一中可以清楚的看到當編碼字越小，表示該影像區塊所表現的顏色越深。

選擇一張代表性影像來訓練編碼書是很重要的，就蛋白質二維凝膠電泳影像來說應選擇影像中黑白分佈均等的影像來進行訓練，這樣才能使編碼書清楚的分隔出每一個灰階分層階度。若選用的訓練影像黑色部份較少，則整個灰階向量值加總偏高，意味著黑色程度較小，這樣的編碼書無法明確表示出蛋白質二維凝膠電泳影像中較黑的部份。下圖六為本研究採用進行編碼書訓練的訓練影像。

表一：以蛋白質二維凝膠電泳影像做為訓練影像執行LBG演算法後所產生的2^3位元編碼書。

Codeword	向量值(N x N維)	總灰值加總	黑色程度
0	58 57 57 58 54 53 53 53 54 54 53 53 54 58 57 57 59	1369	1
1	88 80 88 80 80 79 77 80 80 80 79 80 81 88 80 88	2846	2
2	97 97 97 98 97 97 97 97 97 97 97 97 97 97 97 97	2417	3
3	ac ac ac ac ad ad ad ad ad ad ad ad ad ac ac ac ad	2761	4
4	bd bd bd bd be bd bd bd be bd bd bd bd bd bd bd	3826	5
5	ce ce ce ce ce cf cf ce ce cf ce ce ce ce ce ce	3269	6
6	dd dd dd dd dd dd dd dd dd dd dd dd dd dd dd dd	3736	7
7	f0 f0 f0 f0 f0 f0 f0 f0 f0 f0 f0 f0 f0 f0 f0 f0	3872	8



圖六：就蛋白質二維凝膠電泳影像來說應選擇黑白分佈均等的影像來進行訓練，這樣才能使編碼書清楚的分隔出每一個色階。

透過上述灰階分層階度的原理，我們就可以表現出蛋白質二維凝膠電泳影像重要特徵之一—蛋白質樣本點顏色深淺的問題。以 K=2、N=4、S=256 則 M=(S/N)=64 為例來說明，第一步訓練出 2^2 位元、2^2=4 個編碼字、(N x N)=16 維的編碼書，並將編碼書內編碼字之間順序以各編碼字之 16 個元素值總和進行排序。編碼字分別以 0,1,2,3 索引值表示之，產生之編碼

書如下表二，往後各步驟說明實驗皆使用表二所示編碼書進行。

將大小為(S x S)=(256 x 256) 像素的蛋白質二維凝膠電泳影像進行影像區塊切割，得到(M x M)=(64 x 64)個像素為(N x N)=(4 x 4)的影像區塊。依序將每一個影像區塊在表二編碼書中找出相對映的編碼字，並將相對映的編碼字取代該影像區塊，做為該影像區塊的索引值。最後，將得到一張(M x M)=(64 x 64)個編碼字所組成的索引表。

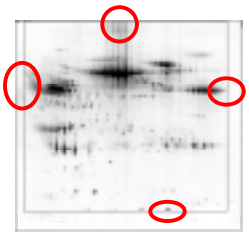
表二：以k=2、N=4、S=256則M=(S/N)=64為例來，訓練出2^2位元、(N x N)=16維的編碼書。

Codeword	像素值(N x N維)	像素值Avg	黑色程度
0	57 54 54 56 52 50 48 51 52 50 48 52 56 55 55 57	1329	1
1	[illegible]	2287	2
2	bb bc bc bb bc bd bd bc bd bd bd bd bc bc bd bc bc	3012	3
3	de de de de de de de de de de de de de de de de	3552	4

3.3 取出索引表中重要特徵區塊—除去影像上影像濾除準確性的雜訊

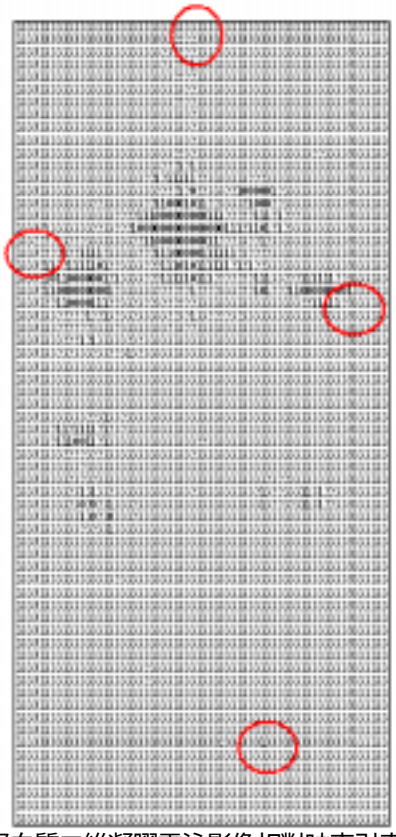
影像平移誤差可能發生在拍攝時影像左上角原點位置不一致，如圖七左邊所示。或是影像四邊經過些許裁切，稱之為影像平移，如圖七右邊所示。若是影像被平移或是影像大小不一，則在重疊比對時影像上蛋白質位置相關資訊將不正確。此步驟我們要來取出各蛋白質二維膠電泳影像上可包括該影像上所有蛋白質樣本點的重要特徵區塊，而去除影像上不重要的雜訊。以下圖八所示，找出蛋白質二維凝膠電泳影像中可以包涵所有蛋白質樣本點的區域。不論影像如何被平移或是大小不一，依舊可以取得蛋白質二維凝膠電泳影像中重要大部份的重要樣本點，藉此去除影像四週不必要的雜訊，如此便可不影響本研究濾除的正確性，我們針對影像上重要的資訊進行濾除工作。

圖七：影像平移誤差可能發生在拍攝時影像左上角原點位置不一致，如左圖所示。或是影像四邊經過些許裁切，稱之為影像平移，如右圖所示。



圖八：找出蛋白質二維凝膠電泳影像中可以包涵所有蛋白質樣本點的區域，去除影像四週不必要的雜訊。

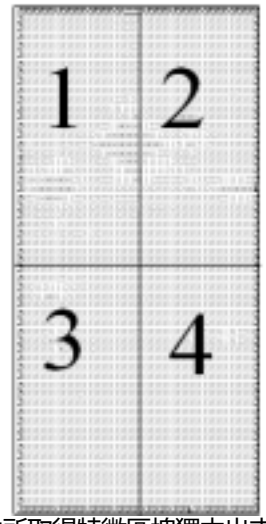
透過研究研究方法3.2產生之表二編碼書，將圖八中蛋白質二維凝膠電泳影像以S=256、N=4經過VQ影像壓縮，取得其相對映編碼字。可得下圖九(M x M)=(64 x 64)的索引表，將圖八和圖九做相對映位置對照可以發現，圖八影像小方框之外白色的部份在圖九相對位置其編碼字是3，表示此區黑色程度最小也就是色最白的部份，此部份沒有任何蛋白質樣本點，我們將此區域視為雜訊地帶。而圖八中蛋白質點的部份可由圖九相對位置發現是由一群編碼字2到0所組成。要找出包涵所有蛋白質樣本點的區域，只需找出圖九中四邊最外緣索引值不是3(不是白色)的四個點，圖八及圖九當中的四個圓圈的位置。之後所有計算步驟都針對蛋白質二維凝膠電泳影像索引表中取出的可包涵所有蛋白質樣本點的區域。



圖九：在蛋白質二維凝膠電泳影像相對映索引表找出中可以涵蓋所有蛋白質樣本點的區域。

3.4 特徵區塊切割定位—取得蛋白質樣本點位置

蛋白質樣本點的位置是蛋白質二維凝膠電泳影像四項重要特徵之一，有時蛋白質二維凝膠電泳影像發生扭曲或是影像平移現象，我們將蛋白質樣本點分區進行定位，這樣一來即使蛋白質二維凝膠電泳影像發生扭曲或是影像平移，每個扭曲或是被平移蛋白質點依舊是落在其所屬定位區內。此步驟目的在於替每一個蛋白質樣本點做定位。圖十顯示將圖九中所取出的特徵區塊獨立出來，定義一個g做為分割定位區的指數，以g=2分割2^g=4個定位區為例加以說明。將每個定位區給予編號，不論區分為多少個，定位區編號方向皆由上而下，由左而右，所有的蛋白質二維凝膠電泳影像採用相同的方向統一定位編號。當依據定位編號依序進行比對，便可考量到蛋白質樣本點位置資訊，也可避免因蛋白質二維凝膠電泳影像扭曲，造成比對誤差。



圖十：將圖十中所取得特徵區塊獨立出來，以分割2^2個定位區為例加以說明。將每個定位區給予編號。

3.5 編碼字數量統計—取得蛋白質樣本點大小

在研究方法3.4時，將可包涵大部份重要蛋白質樣本點的特徵區塊進行2^2個定位區分割，此步驟開始要來計算比對所需的一系列特徵值。圖十一中四個定位區內所有索引值以分區分群的方式進行統計，所謂分區指的是依各定位區，分群指的是依各索引值，分別統計出第一個定位區中索引值是0的個數、第一個定位區中索引值是1的個數、第一個定位區中索引值是2的個數…、第二個定位區中索引值是0的個數、第二個定位區中索引值是1的個數…、第三個定位區中索引值是0的個數…、第四個定位區中索引值是3的個數。以本例切割為四個定位區且採用2^2=4位元的編碼書，則每一定位區內有4個特徵統計值，而整張影像有16個特徵統計值，可視為每一張蛋白質二維凝膠電泳影像特有的特徵值，做為濾除時的重要依據。以X_{ij}來表示每一個特徵統計值，i定義為定位區編號，j定義為索引值編號，例如：X_{22}表示第二個定位區內索引值為2的個數，X_{40}表示第四個定位區內索引值為0的個數。而16維特徵統計值以([X_{00}, X_{11}, X_{12}, X_{13}], [X_{20}, X_{21}, X_{22}, X_{23}], [X_{30}, X_{31}, X_{32}, X_{33}], [X_{40}, X_{41}, X_{42}, X_{43}])表示之，每一個[]符號內都有四個數值，表示這四個數值都來自同一個定位區。透過此步驟所計算出的統計值，可以了解在每一個定位區裡任何黑白色分層階度蛋白質樣本點的大小。

VQ壓縮品質決定於進行壓縮時所使用的Codebook大小，本研究使用256、512、1024、2048以及4096五種Codebook來進行實驗，找出本方法最適當、實驗濾除效果最好的Codebook。

3.6 調整蛋白質二維凝膠電泳影像特徵值算法—考慮影像大小不一致、解析度不同造成的誤差

影像處理上的放大、縮小都會造成影像解析度變異，因此將解析度 不同和影像大小不一致視為同類型問題，在此步驟可一併解決。蛋白質二維凝膠電泳影像資料庫中所存在的資料影像，尺寸不會都相同，尺寸不相同的兩張蛋白質二維凝膠電泳影像是無法用傳統重疊的方式進行比對，但若以放大、縮小來解決尺寸不一致的情況，將會造成區塊效應，無法表達蛋白質本點資訊。需將研究方法3.5中由16個統計值所組成的一系列特徵值([X_{00}, X_{11}, X_{12}, X_{13}], [X_{20}, X_{21}, X_{22}, X_{23}], [X_{30}, X_{31}, X_{32}, X_{33}], [X_{40}, X_{41}, X_{42}, X_{43}])做算法上的調整，即可克服影像處理、製作時造成的誤差問題。

對於研究方法3.5中每一個定位區除了要分群統計出各索引值每一群數量外，尚需算出每一個定位區內共有多少個索引值。以T_1,T_2,T_3,T_4分別表示第一個定位區內索引值個數總數…第四個定位區內索引值個數總數。最後，以第一個定位區內索引值為0的數量在第一個定位區內所占的比率、以第一個定位區內索引值為1的數量在第一個定位區內所占的比率、…以第四個定位區內索引值為3的數量在第四個定位區內所占的比率，算該影像一系列新的特徵值：

```
for i=0 to 3
  for j=0 to 3
      Xij=Xij/Ti+1
```

3.7 自動化蛋白質二維凝膠電泳影像濾除

假若現有一張欲查詢影像要進入蛋白質二維凝膠電泳影像資料庫內與資料影像做濾除時，不論是欲查詢影像或資料庫裡的資料影像，都以相同的K、S、M、N以及g五種變因執行以上的研究方法3.1到研究方法3.6，依實驗所設定K、S、M、N以及g變因的不同，所取得一系列特徵值的長度、含意亦不同。以先前實驗S=256、M=4、N=4、g=2，定位區數量為2^2為例，當研究方法3.6執行後蛋白質二維凝膠電泳影像資料庫裡的資料影像與欲查詢影像都會取得一組由16個比率值所組成的特徵值([X_{10}, X_{11}, X_{12}, X_{13}],[X_{20}, X_{21}, X_{22}, X_{23}], [X_{30}, X_{31}, X_{32}, X_{33}], [X_{40}, X_{41}, X_{42}, X_{43}])。為了避免混淆，將欲查詢影像取得的一組由16個比率值所組成的特徵值另稱為([Q_{10}, Q_{11}, Q_{12}, Q_{13}], [Q_{20}, Q_{21}, Q_{22}, Q_{23}], [Q_{30}, Q_{31}, Q_{32}, Q_{33}], [Q_{40}, Q_{41}, Q_{42}, Q_{43}])。將欲查詢影像的16個比率值所組成的特徵值運用距離公式，公式一(1)，與蛋白質二維凝膠電泳影像資料庫中其他資料影像的特徵值做最小距離的P張影像篩選。

3.8 找出P組中實驗編號出現頻率最高者

檢查以最小距離篩選出P張蛋白質二維凝膠電泳影像它們的實驗編號，同一組實驗所產生之蛋白質二維凝膠電泳影像擁有相同的實驗編號，而同一組實驗中，除了實驗編號相同外，實驗環境、組合變因、細胞種類等變因都是相同的，為了客觀取得某細胞或是某一疾病的蛋白質二維凝膠電泳影像，所以我們進行連續多次實驗，此數塊電泳膠體所取得之數位化影像則稱為同組。統計出在最小距離的P張蛋白質二維凝膠電泳影像中哪一組的實驗編號出現頻率最高。最後，生物學家們只需針對P組中出現頻率最高的實驗編號所代表的該組電泳影像進行比對即可。

四、研究結果

本研究所使用的研究平台為P4 1.3MHZ、128MB RAM、作業系統為Microsoft Windows XP程式撰寫工具為Edit Plus，Java程式語言。我們所提出的方法不但簡單、快速且不需複雜運算，即能達到良好的濾除成效。以下為研究結果分析。

目前資料庫內原存有103張蛋白質二維凝膠電泳影像分別以2D01至2D103來表示(含一張欲查詢影像)，在研究研究方法3.6我們取出特徵區塊以及調整特徵值為比例算法已經解決了蛋白質二維凝膠電泳影像尺寸不一致的問題，因此我們所提出的方法可以接受所有尺寸的電泳影像。預設P為10，濾除後將由資料庫中取出10組與欲查詢影像在蛋白質樣本點大小、蛋白質樣本點個數、蛋白質樣本點顏色深淺以及蛋白質的位置四項特徵最相似的蛋白質二維凝膠電泳影像。包含欲查詢影像在內的103張蛋白質二維凝膠電泳影像中2D40、2D52、2D21、2D33、2D14、2D17以及欲查詢影像2D76皆是由中山醫學大學實驗室所提供相同細胞在不同實驗境下所產生的相同實驗組7張影像，其餘96張為分別從網路上取得不相關種類細胞所產生之蛋白質二維凝膠電泳影像。經過本研究提出的濾除方法之後，對這資料庫而言2D40、2D52、2D21、2D33、2D14、2D17這6張影像是與欲查詢影像2D76為同一組實驗產物，最為相似，是經過初步濾除後與欲查詢影像最小距離的結果影像。

在本研究所提出的方法中有幾個關鍵性的變數：

- S-蛋白質二維凝膠電泳影像原始大小。
- N及M一併討論-(N x N)表示蛋白質二凝膠電泳影像中影像區塊的大小，以及每一個編碼字的向量維度。而(M x M)表示蛋白質二維凝膠電泳影像中影像區塊的個數。
- K-決定編碼書大小的指數。
- g-切割定位區的數量的指數。

以上五種重要變因不同組合都會造成不同的實驗結果，我們嘗試以不同變因的組合，進行多次實驗，找出在本研究提出方法下最適合蛋白質二維凝膠電泳影像、實驗效果最好的變因組合。

利用研究所列距離公式一(1)逐一計算欲查詢影像與資料庫中資料影像兩者之間的距離，取出距離最小10張蛋白質二維凝膠電泳影像作為結果影像。檢查在這10張的結果影像的實驗編號，找出10張中實驗編號出現頻率最高者，則欲查詢影像則和此實驗編號出現頻率最高者為同一組實驗所產生的蛋白質二維凝膠電泳影像。

4.1實驗參數之選擇

(1)由每一組實驗變因組合顯示，欲查詢影像亦為資料庫內供查詢資料影像之一時，發現在與欲查詢影像距離最小的之影像是欲查詢影像本身，其距離為0，查詢者與被查詢者是一模一樣的。

(2)針對編碼書大小不同來說，編碼書所含編碼字越多VQ壓縮效果越好。因編碼字個數越多越能詳細將蛋白質二維凝膠電泳影像內灰階深淺做量化分群。以S=512、g=8、N=4、M=128而K分別為2、3、4的實驗變因所得到濾除結果加以說明，表三為當編碼書大小為2^2=4時經濾除後得到的前10張結果影像，表四為當k=3，編碼書大小為2^3=8時經濾除所得到的前10張結果影像，表五為當編碼書大小為2^4=16時經濾除所得到的前10張結果影像。

表三：當編碼書大小為4時經濾除後得到的前10張結果影像

實驗編號	影像名稱	與欲查詢影像的距離
D0003	2D76	0
D0003	2D21	19068
D0011	2D63	20472
D0003	2D52	20568
D0065	2D62	20899
D0031	2D103	22559
D0002	2D30	22796
D0001	2D13	22829
D0004	2D27	22836
D0030	2D22	23081

表四：當編碼書大小為8時經濾除後得到的前10張結果影像

實驗編號	影像名稱	與欲查詢影像的距離
D0003	2D76	0
D0003	2D21	24490
D0003	2D52	26471
D0070	2D61	26974
D0031	2D103	28790
D0065	2D62	29140
D0011	2D63	29364
D0001	2D59	29924
D0045	2D43	30068
D0065	2D73	30418

表五：當編碼書大小為16時經濾除後得到的前10張結果影像

實驗編號	影像名稱	與欲查詢影像的距離
D0003	2D76	0
D0003	2D21	28781
D0003	2D52	29777
D0003	2D40	32584
D0011	2D63	33865
D0070	2D61	34962
D0012	2D47	35030
D0022	2D74	35058
D0023	2D29	35271
D0010	2D70	35540

表三至表五中以表格底色為灰者，其與欲查詢影像同樣來自中山醫學大學相關細胞所產生的蛋白質二維凝膠電泳影像，上例三表，我們可以看出編碼書小大為16時效果最好，所以可以說明編碼書所含編碼字越多實驗效果越好。

針對影像區塊個數多寡以及影像區塊大小來說。影像區塊個數較多時，影像區塊大小較小，則影像索引表中所含的索引值個數較多，索引表較細緻，能將蛋白質二維凝膠電泳影像表達較詳細。雖然影像區塊大小越小越好，但是我們並不採用(2 x 2) 像素的影像區塊大小，其影像區塊大小過小卻無法正確選擇出編碼書中所對映的編碼字。以S=512、g=6、K=4、N分別以4和8的實驗變因所得到濾除結果加以說明，表六為影像區塊大小為(4 x 4)且影像區塊個數為(128 x 128)時經濾除後得到的前10張結果影像，表七為影像區塊大小為(8 x 8)且影像區塊個數為(64 x 64)時經濾除後所得到的前10張結果影像。

表六：影像區塊大小為(4 x 4)且影像區塊個數為(128 x 128)時經濾除後得到的前10張結果影像

實驗編號	名稱	與欲查詢影像的距離
D0003	2D76	0
D0003	2D21	3994
D0011	2D63	4464
D0003	2D52	4518
D0001	2D13	4829
D0065	2D62	4973
D0004	2D27	5028
D0003	2D40	5144
D0010	2D70	5187
D0002	2D30	5196

表七：影像區塊大小為(8 x 8)且影像區塊個數為(64 x 64)時經濾除後所得到的前10張結果影像

實驗編號	影像名稱	與欲查詢影像的距離
D0003	2D76	0
D0011	2D63	4665
D0089	2D42	4881
D0023	2D45	4946
D0003	2D52	4981
D0003	2D21	5015
D0003	2D40	5059
D0036	2D15	5065
D0031	2D103	5189
D0012	2D97	5196

表六與表七中以表格底色為灰者，其與欲查詢影像同樣來自中山醫學大學相關細胞所產生的蛋白質二維凝膠電泳影像。雖然兩表中查詢到數量相同，但是，可依被查詢到順序來判斷，表九中影像區塊較小，索引表較細膩，可在順序較前時就被發現。可說明影像區塊個數多時也就是影像區塊大小較小，實驗效果越好。

針對切割定位區的數量來說，切割定位區目的在於替蛋白質樣本點做定位，取得位置資訊。原則上來說切割定位區數量越多，實驗效果越好。若切割定位區數量較少，平面上每一個定位區內所分佈的蛋白質數量較多，這些相同定位區內的蛋白質樣本點位置資訊是相同的，位置的資訊將會變得籠統，無法詳細表達出蛋白質樣本點位置資訊。以S=512、N=4、M=128、K=4、g分別以2和6的實驗變因執行加以說明，表八為切割定位區個數是2^2=4時經濾除所得到的前10張結果影像，表九為切割定位區個數是2^6=64時經濾除後得到的前10張結果影像。

表八：當切割定位區數量為4時經濾除後得到的前10張結果影像

實驗編號	影像名稱	與欲查詢影像的距離
D0003	2D76	0
D0001	2D13	107
D0010	2D70	109
D0002	2D30	121
D0051	2D12	131
D0020	2D4	137
D0045	2D27	143
D0011	2D63	144
D0069	2D16	150
D0055	2D19	150

表九：當切割定位區數量為64時經濾除後得到的前10張結果影像

實驗編號	影像名稱	與欲查詢影像的距離
D0003	2D76	0
D0003	2D21	109650
D0003	2D52	117900
D0003	2D40	118700
D0031	2D103	122900
D0065	2D62	124412.5
D0011	2D63	129045.478999999
D0001	2D59	129312.5
D0011	2D24	130395.680999999

表八與表九中以表格底色為灰者是與欲查詢影像同樣來自中山醫學大學相關細胞所產生的蛋白質二維凝膠電泳影像。依兩個表中表格底色為灰者數量來判斷，可說明切割定位區數量個數多時實驗效果越好。

4.2最佳參數組合

從表十最佳變因組合，一一來分析：除了切割定位區數量外其餘四項變因都是符合上述三點所判斷—影像區塊大小以(4 x 4)最佳、影像區塊的個數以(128 x 128)最佳、編碼書大小以16位元為最佳。但是，在切割定位區數量方面，卻會因切割定位區數量過多使得平面上每一個定位區內所包含蛋白質樣本點數量過少，無法確實表達出每一個定位區內所要傳達的資訊。或是面積較大之蛋白質樣本點遭到一分為二，造成無法明確表達出每一個蛋白質樣本點的位置資訊。

表十：實驗效果最好的組合

切割定位區的數量 (2^g)	影像區塊大小 (N x N)	向量維度 (N x N)	影像區塊的個數 (M x M)	編碼書的大小 (2^k)
64	4 x 4	16	128 x 128	16

4.3最佳實驗變因組合後之實驗結果

最後，我們以表十一中取出前十筆記錄如表十三所示，結果影像這10張影像中以實驗編號來看，出現頻率最高為D0003，D0003此組就是與欲查詢影像同樣來自中山醫學大學相同實驗環境、相同實驗變因所產生同一組的蛋白質二維凝膠電泳影像，如表十三所示。最後欲查詢影像只需針對濾除後所得到的實驗編號為D0003這一組蛋白質二維凝膠電泳影像，以Melanie或Z3等工具軟體執行軟體自動比對和人工手動比對，這樣必能減少無謂的比對時間。然而，以表十二中所示除了實驗編號D0003之外的6張影像，在就蛋白質樣本點大小、蛋白質樣本點位置、蛋白質樣本點顏色深淺以及蛋白質樣本點數量四項特徵資訊是十分相似的。

表十一：以本研究實驗證明最好的實驗組合進行全資料庫特徵值計算

實驗編號	排序編號	影像名稱	與欲查詢影像的距離	實驗編號	排序編號	影像名稱	與欲查詢影像的距離	實驗編號	排序編號	影像名稱	與欲查詢影像的距離
D0003	0	2D76	0	D0034	35	2D58	8787	D0016	70	2D31	10022
D0003	1	2D21	5943	D0022	36	2D26	8787	D0024	71	2D98	10022
D0003	2	2D52	6144	D0011	37	2D66	8807	D0033	72	2D28	10034
D0003	3	2D40	6566	D0036	38	2D30	8826	D0067	73	2D48	10071
D0012	4	2D47	7331	D0087	39	2D37	8874	D0054	74	2D54	10148
D0023	5	2D29	7363	D0055	40	2D19	8902	D0035	75	2D53	10148
D0011	6	2D63	7369	D0069	41	2D16	8906	D0022	76	2D74	10150

D0045	7	2D43	7710	D0060	42	2D68	8977	D0011	77	2D5	10183
D0065	8	2D73	7711	D0020	43	2D41	8998	D0046	78	2D46	10184
D0089	9	2D42	7834	D0039	44	2D99	9003	D0070	79	2D61	10217
D0033	10	2D13	7839	D0066	45	2D38	9077	D0061	80	2D90	10297
D0023	11	2D45	7849	D0039	46	2D36	9077	D0084	81	2D102	10326
D0065	12	2D72	7923	D0012	47	2D97	9086	D0033	82	2D44	10329
D0044	13	2D67	7938	D0010	48	2D60	9088	D0022	83	2D69	10435
D0011	14	2D95	7969	D0017	49	2D78	9097	D0035	84	2D3	10543
D0023	15	2D23	8106	D0048	50	2D25	9111	D0076	85	2D32	10603
D0012	16	2D7	8145	D0036	51	2D15	9174	D0003	86	2D33	10656
D0045	17	2D27	8189	D0025	52	2D35	9250	D0028	87	2D34	10847
D0053	18	2D49	8193	D0008	53	2D11	9275	D0031	88	2D65	10986
D0022	19	2D50	8237	D0051	54	2D12	9292	D0065	89	2D94	11194
D0066	20	2D51	8237	D0020	55	2D4	9352	D0050	90	2D92	11250
D0025	21	2D10	8269	D0054	56	2D18	9395	D0067	91	2D82	11703
D0030	22	2D22	8277	D0067	57	2D55	9406	D0019	92	2D77	12652
D0067	23	2D9	8293	D0058	58	2D56	9413	D0022	93	2D81	12660
D0045	24	2D101	8314	D0007	59	2D57	9432	D0034	94	2D80	12676
D0031	25	2D103	8342	D0003	60	2D14	9482	D0056	95	2D89	12691
D0058	26	2D2	8435	D0043	61	2D96	9511	D0034	96	2D87	12698
D0067	27	2D71	8448	D0047	62	2D83	9539	D0002	97	2D86	12709
D0009	28	2D75	8477	D0050	63	2D93	9555	D0021	98	2D88	12712
D0003	29	2D17	8495	D0069	64	2D64	9658	D0034	99	2D79	12713
D0033	30	2D59	8529	D0046	65	2D6	9744	D0046	100	2D85	12723
D0065	31	2D62	8534	D0076	66	2D39	9748	D0022	101	2D84	12756
D0011	32	2D24	8642	D0058	67	2D100	9761	D0023	102	2D91	12756
D0010	33	2D70	8657	D0013	68	2D20	9766				
D0070	34	2D8	8675	D0040	69	2D1	9776				

表十二：本研究實驗證明最好的實驗結果

實驗編號	編號	名稱	與欲查詢影像的距離
D0003	0	2D76	0
D0003	1	2D21	5943
D0003	2	2D52	6144
D0003	3	2D40	6566
D0012	4	2D47	7331
D0023	5	2D29	7363
D0011	6	2D63	7369
D0045	7	2D43	7710
D0065	8	2D73	7711
D0089	9	2D42	7834

表十三：實驗編號為D0003就是與欲查詢影像同樣來自中山醫學大學相同實驗環境、相同實驗變因所產生同一組的蛋白質二維凝膠電泳影像

實驗編號	名稱	與欲查詢影像的距離
D0003	2D76	0
D0003	2D21	5943
D0003	2D52	6144
D0003	2D40	6566
D0003	2D17	8495
D0003	2D14	9482
D0003	2D33	10656

五、結論

本研究將VQ這個基本壓縮技術應用在蛋白質二維凝膠電泳影像的研究在不影響比對準確度下來達到加強影像比對效率，由研究結果分析可很清楚看出本研究提出之應用方法不只是可行，更可以達到百分之一百正確性。本研究提出之方法若能使用在蛋白質二維凝膠電泳影像的比對上，相信可以大幅提升生物學家以及科學家們在比對效率，將比對步驟節省的時間移至後續對人類生命更有意義的結構以及功能分析階段工作之上。

六、參考文獻

[01] John M. Walker r," Proteins", Clifton, NJ:Humana Press,c1984

[02] Khalid Sayood , " Introduction to data compression" , San Francisco : Morgan Kaufmann Publishers, 2000

[03] Allen Gersho, Robert M. Gray ,"Vector quantization and signal compression", Boston : Kluwer Academic Publishers ; Taipei : Maw Chang, c1992

[04] J. Leggett Bailey, " Techniques in protein chemistry", [S.l.]:[s.n.],c1967

[05] Tim Hunt, Steve Prentis, John Tooze," DNA makes RNA makes protein", [s. l. : s. n.],c1983

[06]Bonnie S. Dunbar, " Two-dimensional electrophoresis, and immunological techniques", New York:Plenum Press,c1987

[07] D. Rickwood and B.D. Hames, " Gel electrophoresis of nucleic acids:a practical approach", OxfordNew York:IRL Press at Oxford University Press,c1990

[08]http://tw.expasy.org(ExPASy Molecular Biology Server)

[09] www.2Dgels.com (Z3 Website)

[10] Anthony T. Andrews, " Electrophoresis:theory, technique and biochemical and clinical applications", Oxford:Clarendon,c1986.

[11] ANIL k.JAIN, "Fundamentals of digital image processing", Prentice-Hall International Editions, 1989.

[12] John M. Walker r," Proteins", Clifton, NJ:Humana Press,c1984

[13]Adobe Photoshop7.0 User manual.

[14] Stringner Sue Yang and Huber R. Warner," The Underlying molecular, cellular, and immunological factors in cancer and aging", New York:Plenum Press,c1993

[15] Alan D. B. Malcolm," Molecular medicine", Oxford:IRL,c1984