# 國立交通大學

## 資訊科學系

## 碩士論文

染色體的自動辨識

Automatic Recognition of Human Chromosome

研究生：劉以涵

指導教授：曾文貴　教授

莊仁輝　教授

中華民國九十四年六月

# 染色體的自動辨識

研究生：劉 以 涵　　指導教授：曾 文 貴　博士

莊 仁 輝　博士

國立交通大學電機資訊學院

資訊科學所碩士班

## 摘　要

此篇論文的目的在於針對染色體影像研發一有效且可靠的分析系統，自動化得到該染色體影像的組型，以利於醫療人員的判讀與診斷工作。一個人類細胞內共有 22 對體染色體及 1 對性染色體，隨著組別遞增，各組染色體的長度相對地有變短的趨勢。染色體辨識的另一個重要特徵為染色體黑白紋理分佈，此紋理分佈同樣也是依其組別而有所不同。以這些特徵為基礎，染色體自動化辨識的研究主要分成以下幾個步驟。在中軸擷取階段，我們以方向性切片方式，先求得切片中點，再由這些中點推算染色體的中軸，而染色體紋理的灰階值曲線圖，即為沿著中軸斷面，以特定統計方式求得各中軸點的灰階值所組成。得到一個細胞內所有染色體的灰階值曲線圖後，經過標準化的動作，便可計算兩兩染色體之間的關連性，最後由關連性的數值來判斷染色體的配對以及辨識結果。實驗結果顯示我們所提出的染色體影像自動化辨識流程具有一定的正確性和效率。

# Automatic Recognition of Human Chromosome

Student: I-Han Liu

Advisor: Prof. Wen-Guey Tzeng
Prof. Jen-Hui Chaung

Department of Computer and Information Science
National Chiao Tung University

## Abstract

In this paper we present our approach for automated recognition of human chromosomes. The method we used integrates several novel algorithms to automatically obtain the karyotype of chromosomes of metaphase cell spreads. The characteristics that according to the classes chromosomes are gradually shortened by lengths and each of the 24 classes possesses unique banding patterns can be used to perform chromosome recognition. In the proposed approach, we first obtain the longitudinal axes from chromosome images, and then extract features from band-pattern profiles along the obtained longitudinal axes. After normalization, the features are correlated to obtain the matching scores for every two aligned profiles. Experiments show that the proposed algorithm is efficient and very robust.

# 目錄

**中文摘要** ..... i
**英文摘要** ..... ii
目錄 ..... iii
圖目錄 ..... iv
第一章　簡介 ..... 1
第二章　相關文獻探討 ..... 3
2.1　染色體特徵擷取 ..... 3
2.2　類神經網路分類法 ..... 4
2.3　貝氏分群法（Bayes piecewise classifier） ..... 5
2.4　動態信心分佈函數 ..... 6
第三章　染色體影像特徵 ..... 8
第四章　染色體影像辨識程序 ..... 11
4.1　染色體幾何特徵分析 ..... 11
4.2　染色體中軸之擷取 ..... 13
4.3　染色體紋理特徵之擷取 ..... 21
4.4　染色體紋理特徵之比對 ..... 23
第五章　實驗結果 ..... 28
5.1　染色體中軸擷取 ..... 28
5.2　紋理特徵比對 ..... 31
第六章　結論與未來展望 ..... 36
參考文獻 ..... 37

# 圖目錄

圖 1 染色體影像 ..... 8
圖 2 由圖 1 所得到的染色體組型(Karyotype) ..... 9
圖 3 四個染色體影像與其輪廓 ..... 10
圖 4 二值化的染色體影像 ..... 11
圖 5 染色體周長之分析 ..... 12
圖 6 染色體面積之分析 ..... 13
圖 7 水平方向的染色體切片影像 ..... 14
圖 8 垂直方向的染色體切片影像 ..... 14
圖 9 四個方向的染色體切片中點 ..... 15
圖 10 切片長度分佈 ..... 15
圖 11 過濾切片長度後四個方向的染色體切片及其中點 ..... 16
圖 12 過濾切片長度後的染色體切片另一例 ..... 16
圖 13 中軸點連線結果 ..... 17
圖 14 中軸點連線結果另一例 ..... 17
圖 15 圖 13 經過線段分離處理的結果 ..... 18
圖 16 圖 14 經過線段分離處理的結果 ..... 18
圖 17 圖 15 經過連線的結果 ..... 19
圖 18 圖 16 經過連線的結果 ..... 19
圖 19 圖 17 經過中軸延伸的結果 ..... 20
圖 20 圖 18 經過中軸延伸的結果 ..... 20
圖 25 一條屬於第六對的染色體之灰階值曲線圖 ..... 21
圖 26 一條屬於第二對的染色體之灰階值曲線圖 ..... 22
圖 28 圖 26 經過化簡的結果 ..... 22
圖 29 某細胞中第三對染色體灰階值曲線圖經過標準化以及量化的結果 ..... 23
圖 30 某細胞中第五對染色體灰階值曲線圖做字串對齊後的結果 ..... 24
圖 31 圖 30 經過改良後的結果 ..... 25
圖 32 某一細胞中第三對染色體灰階值曲線經過改良字串對齊演算法處理後的結果 ..... 25
圖 33 對某組資料做關連性計算後的結果 ..... 26
圖 34 圖 33 經過簡化的結果 ..... 27
圖 35 某細胞中所有染色體經過中軸擷取的結果 ..... 30
圖 36 (a) 無法成功擷取中軸之染色體 (b)(a) 的二值化影像 ..... 31
圖 37 無法成功擷取中軸的另一例 ..... 31
圖 38 中軸擷取錯誤之一例 ..... 31
圖 39 各對體染色體做字串對齊的結果 ..... 34

# 第一章　簡介

在醫療體系裡，孕婦產前檢查項目中通常會包括胎兒染色體是否正常的檢驗，而判定正常與否的標準為染色體數目、外觀、以及紋理特徵分佈。目前許多醫療體系在這方面仍採用人工辨識的方式，此舉不但花費許多人力，也非常耗時，因此自動化的染色體辨識程序之研發為當務之急。所以在這裡染色體分析與辨識的主要目的，便是自動化將顯微鏡下所攝得的染色後之染色體影像做分類，來得到染色體組型（karyotype），以利於醫療人員的判讀與診斷工作。

在一個正常人的細胞核中，染色體的數目應該是 46 條，當中包括了 44 條體染色體和 2 條性染色體，體染色體兩兩成對，所以共有 22 對，每一對中的兩個（homologue）基本上紋理分佈是相同的。性染色體視性別而不同，女性的性染色體為一組 X 染色體，而男性則為一條 X 染色體與一條 Y 染色體。22 對體染色體是按照長短來排列的，第一對染色體最長，第二十二對長度最短。然而這些長短是相對的，非絕對的，視細胞生長環境以及萃取時所使用的生物技術而定。每條染色體在其長度的固定比例處會有一個寬度特別窄的地方，稱為著絲點（centromere），在細胞分裂的過程中，原染色體與複製染色體相連接的地方便是著絲點。著絲點的兩端分別為一短軸以及一長軸，短軸稱為 p 臂（p arm），長軸稱為 q 臂（q arm），通常 p 臂長度與染色體全長的比值可以用來當作一項重要的染色體辨識特徵。

目前在染色體自動辨識相關領域已經有一些成果發表，其中大部分是採用類神經網路來做辨識，或是以不同的分類器（classifier）對染色體做分群。使用類神經網路的好處是可以建構非線性的模型，對於未知的輸入亦可得到正確的輸出，且類神經網路可以接受不同種類的變數作為輸入，適應性強；然而類神經網路以迭代方式更新鍵結值與閥值，計算量大，相當耗費電腦資源，且訓練的過程中無法得知需要多少節點個數，太多或太少的節點均會影響系統的準確性，因此往往需以嘗試錯誤的方式得到適當的節點個數。使用分類器的優點是可以獲得唯

一的解，但其缺點是分類前需要有大量的資料來建立分類器模型，因此不適用於資料量少的個案。

為了改善以上幾種情況，我們建立了一種不同的染色體自動辨識方法，當中包括了染色體中軸的擷取、紋理特徵分佈的萃取、以及配對和辨識的過程。在中軸擷取的部分我們使用二值化的染色體影像來找出中軸所在位置，主要觀念是由四個方向的切片來找出中軸大概的位置，去除交疊的中軸線段並且做連接的動作，最後對切片的斜率做微調以找出較為準確的中軸。得到中軸之後，對應到染色體灰階值影像上，便可沿其方向萃取黑白條紋分佈的變化，產生灰階值曲線圖，最後藉由經過標準化的灰階值曲線圖來計算兩兩染色體之間的關連性，以達到配對以及辨識的目的。

本篇論文共分為六章，第一章為簡介，敘述本篇論文的研究動機以及概述在染色體自動辨識這分面的問題上前人所提出的演算法類別，和一般在解決辨識問題時可能面對的難題。第二章為相關研究與應用，介紹過去幾年在染色體自動辨識方面的研究與成果，以及一些相關的基本概念。第三章的內容為我們所使用的資料庫中的染色體外觀特徵概述。第四章是敘述我們所使用的方法的主要章節，從染色體中軸的擷取，紋理特徵萃取，到染色體配對以及辨識，均有詳盡的介紹。第五章為實驗結果，分析實驗過程中各種參數的調配以及影響，並且由實驗的結果證明在此所提出的演算法具有一定程度的正確性。第六章為結論與未來展望，說明此演算法的特性以及貢獻，並且針對還可以繼續加強之處作說明，期望未來能對染色體影像自動辨識相關問題提供一可靠的解決方案。

# 第二章　相關文獻探討

本章將探討在染色體影像辨識方面的相關文獻，以期瞭解目前在染色體自動辨識領域上所提出的各種方法以及研究的方向。

## 2.1　　染色體特徵擷取

Gunter 和 Gernot 在[8]一文中說到，最精準的染色體自動辨識方式便是沿染色體長軸取出黑白紋理分佈的曲線，因此，獲得可信賴的染色體中軸是一個重要的步驟。Gunter 和 Gernot 所使用的方法為，先找出染色體的輪廓（contour），計算輪廓的曲率，再依據曲率來找出區域性的最大值以及最小值，然而區域性的最大值與最小值不一定是最佳的，因此必須在輪廓曲率中找出最佳的決定點（dominant point）。

在一般的情況下，決定點的個數期望值為二，然而染色體在分裂的不同時期會有不同的形狀，而且由於顯微鏡下的染色體呈現散亂的分佈，有可能發生擠壓以及彎曲的情況，因此決定點的個數不一定會是兩個。若決定點的個數為二，就假設這兩點為染色體的兩端，並且以此找出長軸。

當決定點的個數為三或四時，染色體有可能是處於分裂的後期，因此產生了Y形與X形染色體。對於Y形染色體（決定點個數為三）來說，若決定點分別為 $D_1$、$D_2$、$D_3$，則長軸可能為 $(D_1,(D_2,D_3))$、$(D_2,(D_1,D_3))$、$(D_3,(D_1,D_2))$、$(D_1,D_2)$、$(D_1,D_3)$、或 $(D_2,D_3)$ 這六種連線其中一種。前三種連線方式為一個決定點連到另外兩個決定點的中點，而後三種方式則為點對點的連線。對於X形染色體（決定點個數為四）來說，若決定點分別為$D_1 \sim D_4$，則長軸可能為 $((D_1,D_2),(D_3,D_4))$、$((D_1,D_4),(D_2,D_3))$、或是任兩點的連線，總共八種情形。

得到染色體長軸後，在長軸上每隔固定距離取一個點，並加以計算來得到數據圖（profile）。Gunter 和 Gernot 根據[14]的作法取出三種數據圖，分別為密度數據圖（density profile），梯度數據圖（gradient profile）以及形式數據圖（shape profile）。得到這些特徵之後，以貝氏評量法（Bayesian estimator）來對 Cpr 資料庫中的染色體做分類，可以得到最低 0.6%的錯誤率。

## 2.2 類神經網路分類法

目前已經發表的染色體自動辨識方法中，有許多是使用類神經網路機制來做辨識[1][3][4]。類神經網路是一種基於腦與神經系統研究所啟發的資訊處理技術，它可以利用現成的資料庫，或一組範例（即系統輸入與輸出所組成的資料）來進行系統模型的建置工作。一個類神經網路系統中通常含有三個層別，分別為輸入層、隱藏層、以及輸出層。使用類神經網路機制的好處是不需要建立龐大的資料庫，就算用來做辨識的特徵個數增加了，也不需要增加訓練資料組（training set），因此就算資料庫中的資料很少時也可以使用。此外，類神經網路具有學習的能力，可以用遞迴的方式使得結果越來越好。

Lerner 在[1][4]兩文中將多層類神經網路（multilayer perceptron neural network）應用在特徵維度縮減、影像分割、以及分類上，並且以倒傳遞學習演算法（backpropagation learning algorithm）來訓練此類神經網路。倒傳遞演算法是一種以誤差來驅策參數評估的演算法，主要目的是調節類神經網路中各層之間的權重以及節點的閥值，使類神經網路的輸出能得到最小誤差。計算權重的函式為

$$W^{t+1}(i,j) = W^{t}(i,j) + \mu\delta^{t}(j)X^{t}(i) + \alpha\left(W^{t}(i,j) - W^{t-1}(i,j)\right) \qquad (2.1)$$

其中$t$為遞迴索引，$W^t(i,j)$為節點$i$連接節點$j$的權重，$\delta$為校正函式，$X^t(i)$是節點$i$的輸出，$\mu$為學習速率（learning rate），而$\alpha$為衝量常數（momentum constant）。

類神經網路學習的效果是由學習速率以及衝量常數所控制，然而這些參數並沒有所謂的標準，通常是由使用者憑經驗調配。而類神經網路中的隱藏層（hidden layer）之層數以及隱藏層中的節點數會影響到類神經網路的形狀，因而也會影響到分群的結果以及複雜度。在[1][3][4]中，類神經網路的輸入皆為多維度的特徵，而輸出則視分群的數目而定，在[1]中，Lerner 先用一個類神經網路將所有染色體分到七個群組中，每個群組又個別以一個類神經網路將群組中的項目分至 2 到 8 個不等的類別中；而[4]中則是直接將所有染色體分至 24 個類別中（1～22 對以及 X、Y 染色體）。

## 2.3 貝氏分群法（Bayes piecewise classifier）

當我們有一群已知的物件，且這些物件已經被歸類為幾個群組時，可以依據這些資料將新加入的物件以貝氏分群法歸類至其所屬的群組中。在[4]一文中，作者假設染色體的特徵都是呈現多變量常態分佈（multivariate normally distributed），則一個擁有特徵向量 $x$ 的染色體在貝氏分群法中會依照它的後驗機率（posterior probability, $L(k)$）被分類

$$x \in C_k \quad if\ L(k) > L(j), \quad \forall j \neq k \tag{2.2}$$

其中

$$L(k) = 2\log(P_k) - \log(|\Sigma_k|) - (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) \tag{2.3}$$

上面兩個式子中的$C_k$代表群組$k$，$\mu_k$是特徵的平均向量，而$\Sigma_k$為群組$k$的協方差

（covariance）。在這裡由於只將五種染色體分群，因此某條染色體被分類至群組$k$的機率$P_k$皆為 1/5。

使用貝氏分群法的好處是它可以將錯誤率降至最低，然而要建立一個貝氏分類器之前，必須要先得到每個群組的條件機率密度函式（conditional probability density function），但是在大部分的應用當中，只能對此函式做大概的評估，且評估的過程非常複雜，需要大量的樣本來建立準確的結果。因此在取樣較少的應用當中，其效用會比類神經網路分類法差。

## 2.4 動態信心分佈函數

Ronald J. S.等人在[2]一文中提出了一種基於類神經網路以及動態信心分佈函數的染色體配對方法，其優點是可以偵測出因結構異常而產生的染色體型變，並且改善傳輸演算法（transportation algorithm）的缺點。傳輸演算法只能察覺染色體數目上的異常，對於染色體本身紋理的異常則無法發覺，因此在使用傳輸演算法時，必須假設染色體本身都是正常的。

在[2]中，作者們的目的是在指定一個類別的前提下，可以從染色體群中找出屬於此一類別的染色體。前處理包括了染色體影像擷取、尋找染色體骨幹、特徵擷取…等，接著以類神經網路在染色體群中找出目標類別的候選染色體（candidate），並且把當中分數最高的一個候選者當作主要染色體（primary）。讓每一條候選染色體都與主要染色體做動態編制（dynamic programming），也就是字串對齊的動作，經過動態編制之後，可以得到兩組信心分佈函數，即為動態信心分佈（dynamic confidence distribution, DCD）

$$1-(\textit{candidate score}/\textit{highest score}) \qquad (2.4)$$

以及距離量測信心分佈（distance measure confidence distribution, DMCD）

$$1 - (\textit{candidate distance/maximum distance}) \qquad (2.5)$$

最後計算動態信心分佈與距離量測信心分佈的乘積，擁有最高分的候選染色體即為主要染色體的配對染色體（homologue）。由於單向的評估結果不一定準確，因此必須作雙向的確認，方法是將剛才得到的配對染色體當作主要染色體，並且再一次以類神經網路從染色體群中找出數條候選染色體並重複先前的步驟，求得一條配對染色體，視其是否為一開始選出的主要染色體，若是，則確定此兩條染色體互為配對染色體，且分類成功；若否，則表示分類失敗。

# 第三章　染色體影像特徵

染色體自動辨識的主要目的，是自動化由顯微鏡下所攝得之染色體影像，如圖 1，來得到染色體組型，如圖 2。顯微鏡下的染色體影像中除了散亂排列的染色體外，通常還會有破碎的細胞核以及其他雜質，這些雜質在前端處理中會被過濾掉，但過濾的方法亦會影響到染色體影像之品質。

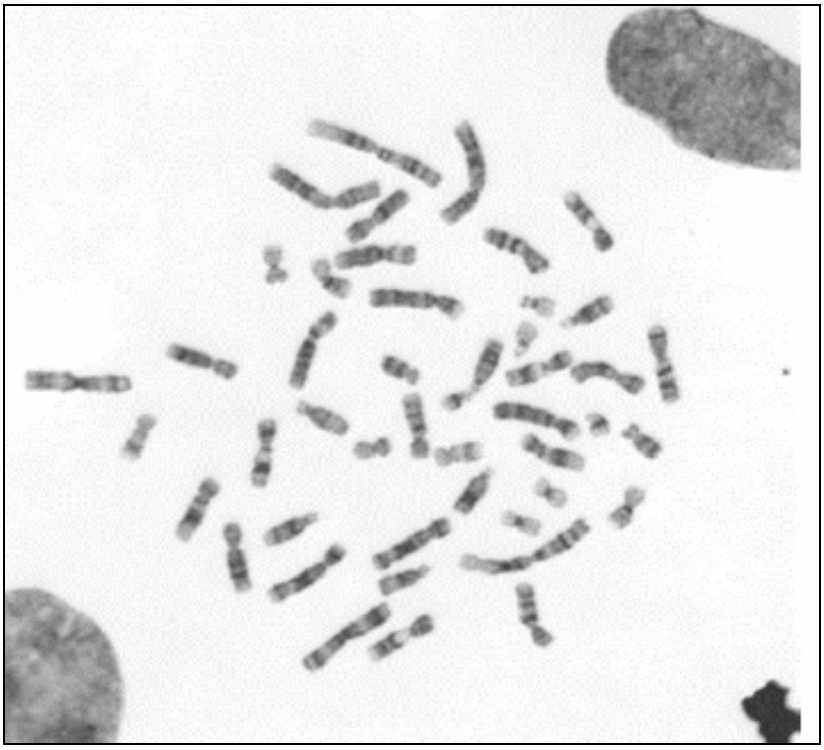


圖 1 染色體影像

由圖 1 中可以觀察到以下特徵：

1. 經過染色技術染色之後的染色體，在灰階顯微鏡下會呈現黑白相間的條紋。
2. 染色體外觀基本上為長條狀，但會有些微扭曲的情況。此外，在某些染色體影像中還可以發現交疊的情況，以及不正常的染色體會發生斷裂、錯置等情形。

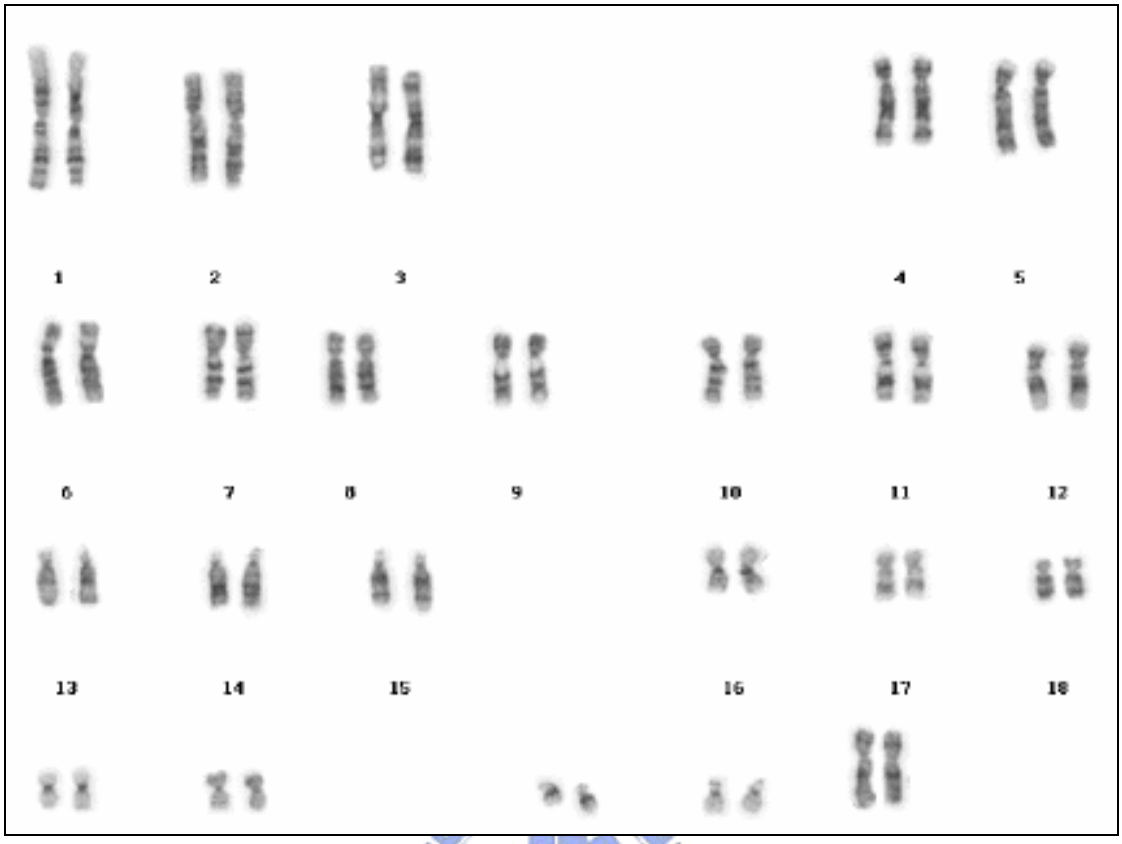



圖 2 由圖 1 所得到的染色體組型(Karyotype)

一個人類細胞內的染色體總共有 22 對體染色體及 1 對性染色體。體染色體每對的紋路相同，因此看起來相似。男性的性染色體為不同的 X、Y 型，女性的性染色體則為相同的 X、X 型。此外，由圖 2 中的染色體組型可觀察到，隨著組別遞增，各組染色體的長度相對地有變短的趨勢。在細胞核分裂時，經過染色體複製後的兩條染色體會以著絲點(centromere)相連，因此著絲點區域的寬度會特別窄，其位置也是依其組別而有所不同。以著絲點位置做區分，染色體的短軸稱為 p 臂(p arm)，長軸稱為 q 臂(q arm)，p 臂長度與染色體全長的比例通常被當作染色體辨識的重要依據。另外一個重要的特徵就是染色體的黑白條紋，此條紋的分布，同樣地也是依其組別而不同。而通常染色體的辨識，是以分裂中期的細胞中的染色體為分析對象。以上這些觀察，在醫療領域有關染色體辨識的研究中，皆被用來作為重要的依據和指標。

染色體影像會因萃取方式、培養環境、人類各器官內細胞以及分裂時期的不

同而有差異性，而這些差異對於辨識所用的方法以及結果都會有很大的影響。我們所得到的染色體影像共有兩種，第一種是如圖 3 中上排的影像，也就是染色體的灰階值影像；第二種是圖 3 中下排的影像，也就是相對於上排灰階值影像的二值化影像。由於我們所使用的資料庫中的影像都是由染色體原圖（如圖 1）經過影像切割而來，因此每張影像中的染色體都已經被調整為直立的狀態。此外，這些影像在外觀上具有以下特徵：沿中軸部分顏色較淺，兩側顏色較深；著絲點的部分不明顯。由於中軸部分顏色較淺，因此在擷取染色體黑白條紋密度剖面圖時，不能只考慮到中軸的部分，兩側的灰階值也需列入計算。而著絲點的不明顯，會影響到 p 臂長度與染色體全長比例的擷取，因此在這篇論文中不使用這項分類指標。評量以上各點之後，最後在實驗中所選用的特徵為染色體周長以及黑白條紋的分佈。

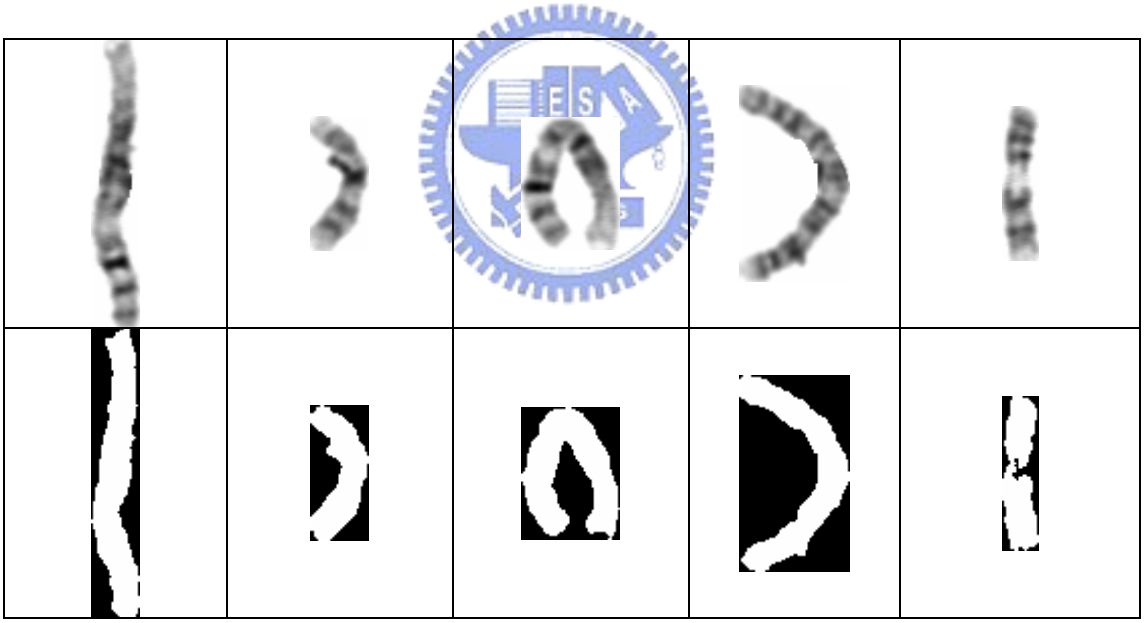


圖 3 四個染色體影像與其輪廓

圖 3 中位於最後一欄的染色體由於切割時所使用的閥值不恰當，因此造成了二值化影像破裂的情形，不利於染色體中軸的擷取，因此在第五章的實驗部分，此種影像不予計算。

# 第四章　染色體影像辨識程序

染色體辨識目前在醫院中主要是作孕婦產前檢查之用，醫療人員可由一個細胞當中的染色體數目、外觀以及紋理正常與否來判斷胎兒的健康情形。然而許多在染色體辨識的部分仍採人工作業流程，耗費許多人力以及時間，因此研發一套自動化的染色體辨識系統是勢在必行的。目前這方面已有些許研究成果發表，使用的方法大多是以類神經網路來對染色體特徵做訓練[1][3][4]，或是以分群的方式來做分類[4][11]。在此介紹我們發展的染色體影像辨識流程，程序包括幾何特徵分析，中軸之擷取，紋理特徵之擷取，以及特徵相互比對。這裡所使用的幾何特徵為染色體周長與面積，並以此做染色體初步的分類，最後再配合紋理分佈等特徵來做比對以及辨識。

## 4.1　染色體幾何特徵分析

在正常的染色體組型中，除了性染色體之外的二十二對染色體是由長至短排列的，因此染色體的長度分析是染色體辨識當中十分重要的一環。由於精準的染色體長度不易取得，在這個部分我們選擇以較容易計算的面積與周長來對染色體做大概的分組。

圖 4 二值化的染色體影像

實作的部分是使用資料庫中的二值影像（如圖 4）。計算周長時，若一個像

素本身為白色，而其上下左右四點其中一點為黑色，則將此點判為染色體周長上的點，最後統計所有周長上點的數目即為一條染色體之周長。面積的計算方式則是統計影像中所有白色的像素總和。

針對染色體周長部分的分析如圖 5，此為針對某細胞中所有的染色體之周長所做的分析。第一行為四十六條染色體的圖檔名稱，開頭數字相同的為一對；第二行為每條染色體的周長；三至六行分別為其長度增減 5%、10%、15%以及 20%時，周長在此範圍內的染色體數目。紅色的字體表示同一對的另一條染色體不在範圍內，黑色的字體則相反。對於人類來說，若拿到的檢體為男性細胞，則其第二十三對染色體的兩條之間長度差異頗大，因此在作長度的分析時不能列入計算。以圖 5 這個例子來說，第二十三對分別為 x、y 染色體，因此為男性檢體。分析圖中數據，長度增減 5%以內可以找到同對染色體的機率為 18/44，長度增減 10%以內的機率為 25/44，長度增減 15%以內的機率為 27/44，而長度增減 20%以內的機率為 33/44。對於初步的分組來說，長度增減 5%能找到同對染色體的機率太低，而增減 20%的範圍內候選者又太多，因此最佳的範圍落在增減 10%或 15%之間。

| | | 5% | 10% | 15% | 20% | | | 5% | 10% | 15% | 20% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 01-1m.bmp | 300 | 0 | 0 | 0 | 3 | 12-2m.bmp | 138 | 0 | 4 | 8 | 10 |
| 01-2m.bmp | 250 | 4 | 4 | 6 | 8 | 13-1m.bmp | 115 | 2 | 3 | 7 | 11 |
| 02-1m.bmp | 260 | 4 | 4 | 4 | 7 | 13-2m.bmp | 146 | 1 | 6 | 8 | 9 |
| 02-2m.bmp | 255 | 4 | 4 | 5 | 8 | 14-1m.bmp | 124 | 3 | 4 | 6 | 8 |
| 03-1m.bmp | 223 | 2 | 4 | 10 | 16 | 14-2m.bmp | 130 | 2 | 4 | 6 | 10 |
| 03-2m.bmp | 250 | 4 | 4 | 6 | 8 | 15-1m.bmp | 128 | 2 | 4 | 7 | 8 |
| 04-1m.bmp | 211 | 4 | 7 | 11 | 15 | 15-2m.bmp | 104 | 1 | 6 | 10 | 12 |
| 04-2m.bmp | 214 | 4 | 5 | 11 | 15 | 16-1m.bmp | 112 | 1 | 4 | 9 | 12 |
| 05-1m.bmp | 201 | 4 | 10 | 12 | 12 | 16-2m.bmp | 94 | 5 | 7 | 9 | 11 |
| 05-2m.bmp | 251 | 4 | 4 | 6 | 9 | 17-1m.bmp | 103 | 2 | 7 | 10 | 11 |
| 06-1m.bmp | 187 | 5 | 8 | 10 | 14 | 17-2m.bmp | 120 | 2 | 5 | 5 | 8 |
| 06-2m.bmp | 194 | 6 | 9 | 12 | 12 | 18-1m.bmp | 99 | 4 | 8 | 9 | 11 |
| 07-1m.bmp | 221 | 3 | 5 | 10 | 16 | 18-2m.bmp | 98 | 4 | 8 | 9 | 11 |
| 07-2m.bmp | 204 | 3 | 11 | 12 | 12 | 19-1m.bmp | 91 | 2 | 7 | 9 | 11 |
| 08-1m.bmp | 191 | 5 | 8 | 10 | 12 | 19-2m.bmp | 96 | 5 | 8 | 9 | 11 |
| 08-2m.bmp | 190 | 5 | 8 | 10 | 13 | 20-1m.bmp | 79 | 1 | 4 | 4 | 7 |
| 09-1m.bmp | 159 | 3 | 5 | 6 | 10 | 20-2m.bmp | 84 | 0 | 2 | 7 | 11 |
| 09-2m.bmp | 190 | 5 | 8 | 10 | 13 | 21-1m.bmp | 76 | 2 | 3 | 4 | 5 |
| 10-1m.bmp | 151 | 4 | 6 | 6 | 9 | 21-2m.bmp | 74 | 2 | 3 | 4 | 4 |
| 10-2m.bmp | 154 | 4 | 5 | 6 | 8 | 22-1m.bmp | 98 | 4 | 8 | 9 | 11 |
| 11-1m.bmp | 154 | 4 | 5 | 6 | 8 | 22-2m.bmp | 93 | 3 | 6 | 9 | 10 |
| 11-2m.bmp | 157 | 4 | 5 | 7 | 8 | xm.bmp | 193 | 6 | 9 | 11 | 12 |
| 12-1m.bmp | 178 | 0 | 6 | 10 | 14 | ym.bmp | 72 | 1 | 3 | 3 | 4 |

圖 5 染色體周長之分析

針對染色體面積部分的分析如圖 6（與周長分析的部分使用同一組影像）。第二行為各條染色體的面積。從這個例子可看出，面積增減 15%的範圍內可以找到同對染色體的機率明顯較 5%與 10%為高，因此在初步分組時使用增減 15%的範圍應該是最適當的。此外，在面積增減 20%的範圍內可找到同對染色體的機率為 43/44，較周長增減同範圍的結果（33/44）好許多，因此以面積來做初步的分組或許是較佳的方案。

由以上的結果可以看出，面積與周長的統計的確可以作為染色體初步分組的依據。但資料庫中每組影像計算出來的資料差異頗大，因此範圍的選取是這部分較為重要的議題。

| | | 5% | 10% | 15% | 20% | | | 5% | 10% | 15% | 20% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 01-1m.bmp | 1950 | 0 | 2 | 3 | 5 | 12-2m.bmp | 959 | 4 | 6 | 8 | 9 |
| 01-2m.bmp | 1732 | 2 | 4 | 6 | 7 | 13-1m.bmp | 696 | 3 | 6 | 7 | 8 |
| 02-1m.bmp | 1782 | 2 | 4 | 5 | 6 | 13-2m.bmp | 747 | 3 | 5 | 6 | 8 |
| 02-2m.bmp | 1807 | 2 | 3 | 5 | 6 | 14-1m.bmp | 715 | 4 | 5 | 7 | 8 |
| 03-1m.bmp | 1490 | 0 | 4 | 6 | 10 | 14-2m.bmp | 726 | 5 | 5 | 6 | 8 |
| 03-2m.bmp | 1609 | 1 | 3 | 6 | 8 | 15-1m.bmp | 760 | 2 | 5 | 6 | 9 |
| 04-1m.bmp | 1577 | 1 | 3 | 7 | 9 | 15-2m.bmp | 696 | 3 | 6 | 7 | 8 |
| 04-2m.bmp | 1349 | 2 | 5 | 8 | 11 | 16-1m.bmp | 611 | 1 | 2 | 7 | 9 |
| 05-1m.bmp | 1305 | 3 | 6 | 8 | 10 | 16-2m.bmp | 528 | 2 | 3 | 4 | 8 |
| 05-2m.bmp | 1387 | 1 | 5 | 8 | 10 | 17-1m.bmp | 539 | 3 | 3 | 4 | 8 |
| 06-1m.bmp | 1220 | 4 | 6 | 9 | 10 | 17-2m.bmp | 628 | 1 | 1 | 7 | 10 |
| 06-2m.bmp | 1268 | 3 | 7 | 8 | 11 | 18-1m.bmp | 562 | 2 | 4 | 5 | 6 |
| 07-1m.bmp | 1256 | 3 | 6 | 9 | 11 | 18-2m.bmp | 547 | 3 | 3 | 5 | 7 |
| 07-2m.bmp | 1103 | 1 | 4 | 12 | 14 | 19-1m.bmp | 439 | 3 | 4 | 5 | 6 |
| 08-1m.bmp | 1077 | 2 | 4 | 10 | 13 | 19-2m.bmp | 400 | 1 | 4 | 6 | 7 |
| 08-2m.bmp | 949 | 5 | 6 | 8 | 10 | 20-1m.bmp | 434 | 3 | 4 | 5 | 6 |
| 09-1m.bmp | 839 | 0 | 2 | 10 | 12 | 20-2m.bmp | 450 | 2 | 3 | 5 | 7 |
| 09-2m.bmp | 949 | 5 | 6 | 8 | 10 | 21-1m.bmp | 296 | 0 | 0 | 1 | 2 |
| 10-1m.bmp | 1025 | 0 | 7 | 9 | 12 | 21-2m.bmp | 334 | 0 | 1 | 2 | 4 |
| 10-2m.bmp | 938 | 5 | 6 | 8 | 10 | 22-1m.bmp | 390 | 1 | 3 | 6 | 7 |
| 11-1m.bmp | 947 | 5 | 6 | 8 | 10 | 22-2m.bmp | 354 | 0 | 1 | 3 | 5 |
| 11-2m.bmp | 907 | 4 | 6 | 7 | 11 | xm.bmp | 1169 | 2 | 6 | 8 | 15 |
| 12-1m.bmp | 1182 | 2 | 6 | 9 | 14 | ym.bmp | 422 | 2 | 5 | 5 | 6 |

圖 6 染色體面積之分析

## 4.2 染色體中軸之擷取

由於染色體二值化影像所含的資訊較單純，因此在染色體中軸擷取的部分我

們使用二值化的影像來做處理，再將擷取出來的中軸對應到灰階值影像上以進行特徵擷取的步驟。因為考慮到染色體的旋轉方向不一定，所以在這裡我們擷取染色體中軸的第一個步驟便是對每條染色體的二值化影像取四個方向的切片，此四個方向分別是水平方向、垂直方向、45°斜線方向與 135°斜線方向，以特定間距切割染色體輪廓影像，可得到該切片方向的切片影像，如圖 7 的水平方向切片影像，以及圖 8 的垂直方向切片影像所示。圖中每條切線段的中點以藍色表示，為染色體中軸軌跡點，將各方向所有切片的中點疊加起來，可以得到圖 9。由圖 9 可發現，在這些中點當中，除了近似中軸軌跡的點以外，尚有一些離中軸較遠的雜點。再比較圖 7、圖 8 與圖 9 可以知道，該雜點是由某些未經過染色體中軸的切片所產生，而這些切片有一個共通的特性，即其長度甚大於或甚小於染色體的寬度。因此，只要濾去切片長度不合理者，即可去掉絕大部分不正確的染色體中軸點。

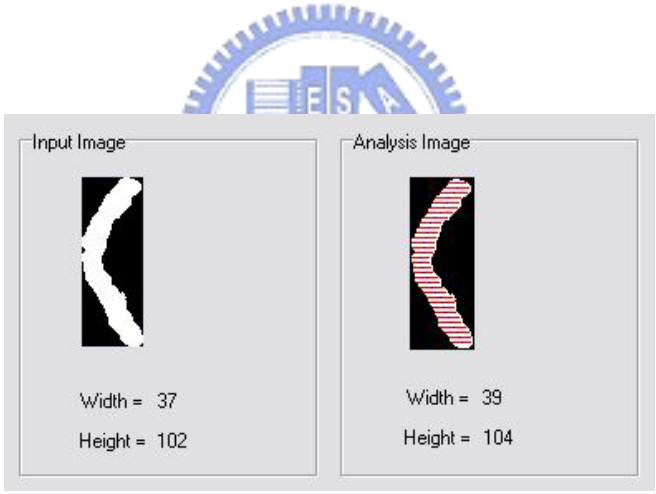



圖 7 水平方向的染色體切片影像

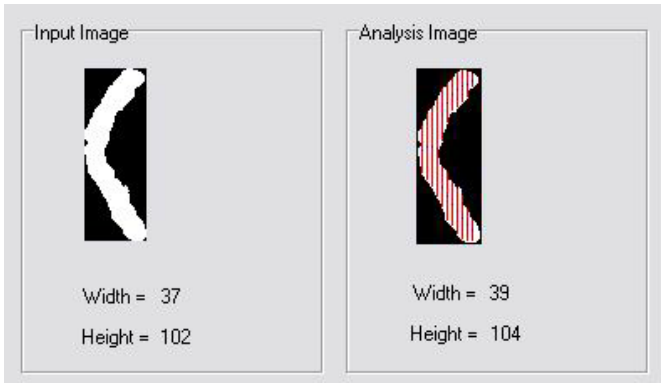



圖 8 垂直方向的染色體切片影像


圖 9 四個方向的染色體切片中點

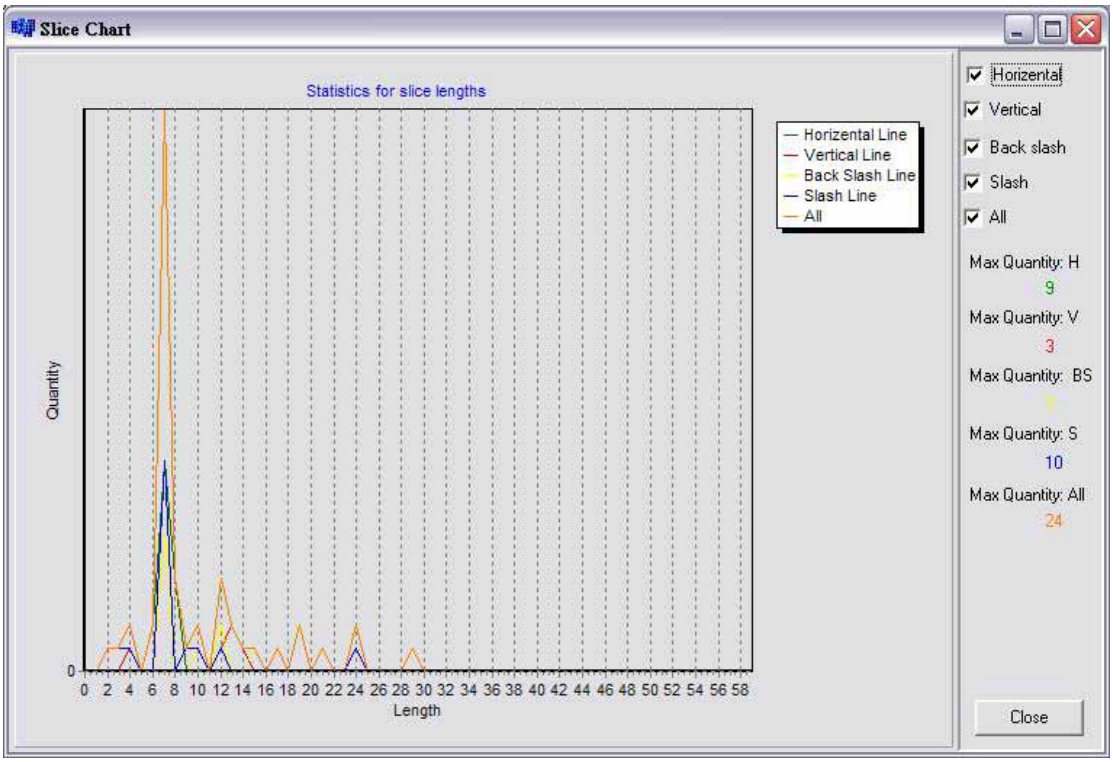



圖 10 切片長度分佈

此外，若將各方向的切片長度以特定區間分組，可發現切片長度有特定分佈，如圖 10。由圖 10 中可以得知，四個切片方向（以不同顏色代表）出現次數最多的切片，皆落在相同的切片長度區間（橫軸），此條染色體的寬度應該就在此長度區間內。此現象說明可以利用切片長度來過濾掉不合理的切片，進而去掉離中軸軌跡太遠的切片中點。而過濾之後的切片中點，即代表染色體中軸所通過的近似軌跡。另外，實際應用中取樣不用像圖 7 那麼密，可放寬取樣間距至染色

體寬度的 1/2 到 1/3 之間，即可得到足夠數量的切片，如圖 11。比較圖 9 與圖 11 可以發現，過濾切片長度之後，中軸點數量已經減少，便於後續處理，且剩下的中軸點確實有足夠的代表性，可以據以得到中軸。圖 12 是此方法對不同染色體影像進行實驗的另外一個例子。

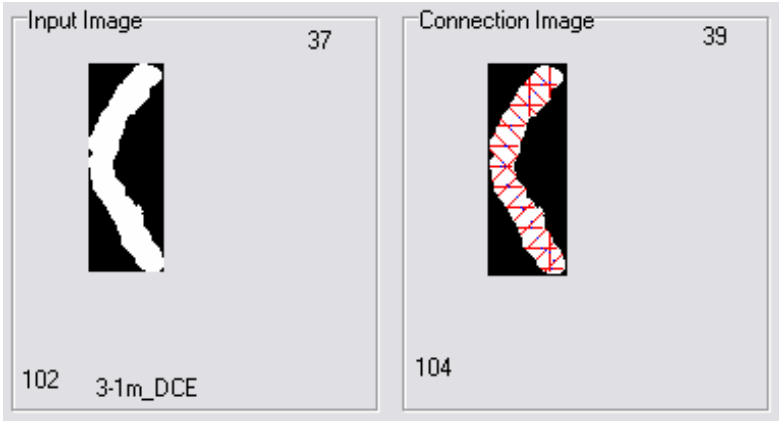



圖 11 過濾切片長度後四個方向的染色體切片及其中點

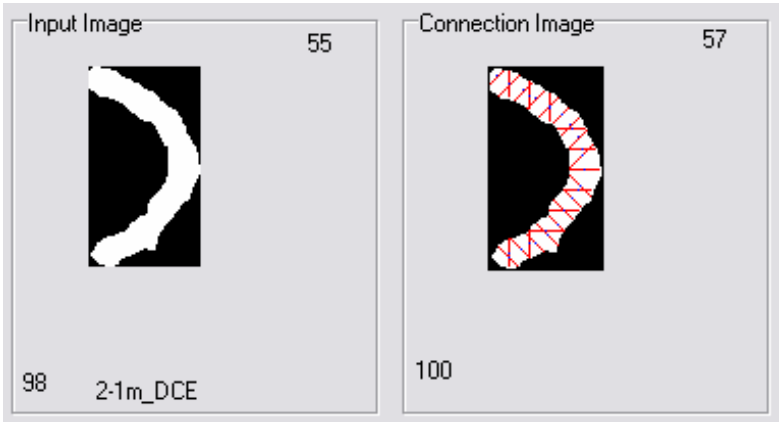



圖 12 過濾切片長度後的染色體切片另一例

接下來便進行中軸點連線的步驟。雖然過濾後的中軸軌跡點對於中軸軌跡已具有相當的代表性，但是仍然有一些問題需要克服：

1. 由圖 11 以及圖 12 中可以注意到，有些相同方向的中軸軌跡點的切片相距甚遠，其連線顯然不能代表染色體中軸。
2. 不同方向的中軸軌跡點的切片，有些會重疊或是非常靠近，造成他們的連線有相交或是重疊的情形發生。

針對第一點，我們以相鄰切片及相同方向作為中軸連線的條件，所得到的線

段再進行後續的連接，以得到完整的中軸。針對第二點，可以加入動態調整切片長度過濾條件的機制，不但可以避免中軸軌跡點交疊的情形，對於不同方向的中軸線段之間的連線也有助益。圖 13 為中軸點連線的結果，其中不同顏色的線段，分別代表不同方向的切片做連線的結果。圖 14 為另一個例子。可以發現所得到的線段，都有足夠的代表性，提供後續的各線段連接以得到完整中軸的步驟不錯的基礎。

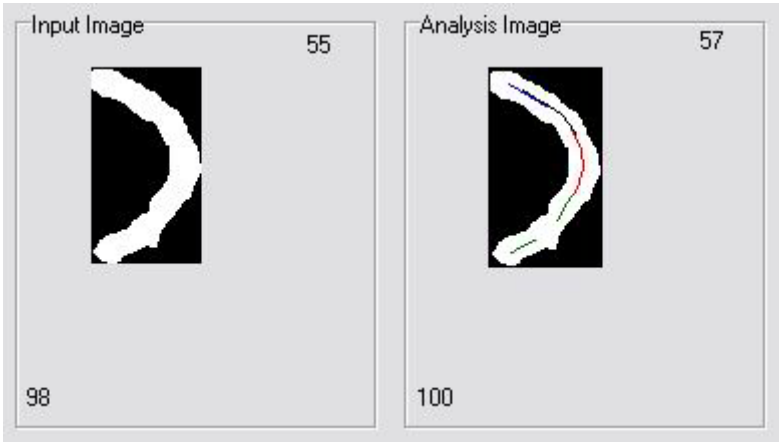



圖 13 中軸點連線結果

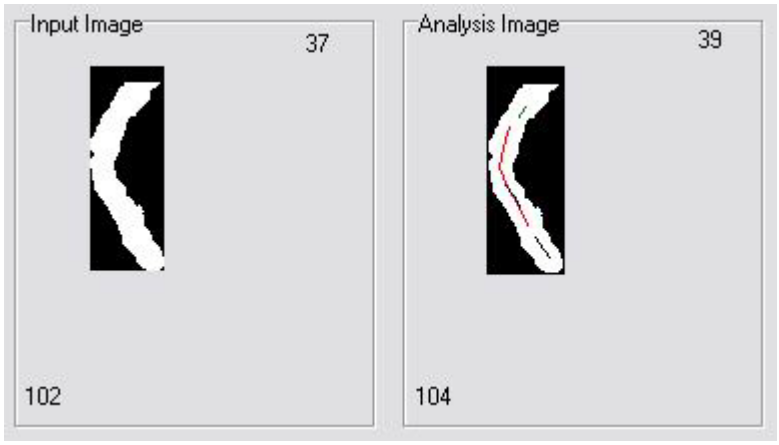



圖 14 中軸點連線結果另一例

在圖 13 以及圖 14 中皆可發現兩種不同顏色線段交錯或覆蓋的情況，例如圖 14 中，較上方的黑色線段覆蓋在紅色線段上。這種線段交錯的現象對於中軸的擷取是無益的，因此必須將之排除。在這裡所使用的方法是比較兩條線段端點 $y$ 軸的大小。圖 15 以及圖 16 分別為圖 13 和圖 14 經過此方式分離交錯線段的結果。

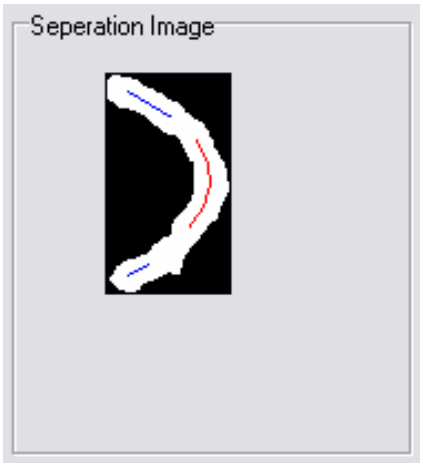



圖 15 圖 13 經過線段分離處理的結果

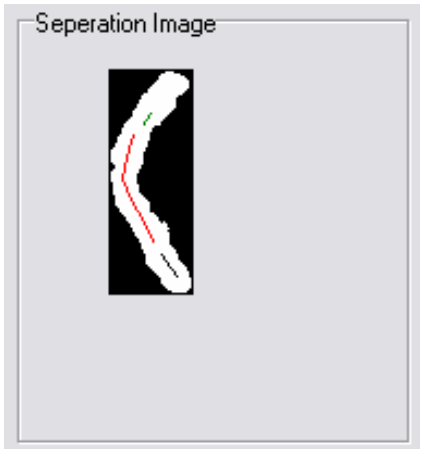



圖 16 圖 14 經過線段分離處理的結果

處理完線段交錯的情況之後，接下來的步驟是對所有相互分離的線段做連線。由於每條線段都會有頭尾兩端，因此每兩條線段之間要做連線的話會有四種不同的排列組合，將所有線段兩兩做連線之後對所有連線的長度做排序，由最短的連線開始判斷它所連接的兩個端點是否已經和別的端點有連線，若是，則刪除這條連線，若否，則保留這條連線。以此類推，直到所有的端點中只剩下兩個端點未與其他端點相連接為止，而這兩個端點就是整條染色體中軸連線的頭尾兩端。連線的結果如圖 17 以及圖 18（圖中紫色的線段即為兩兩的連線）。

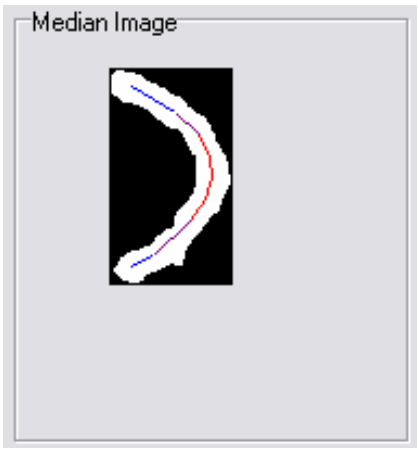



圖 17 圖 15 經過連線的結果

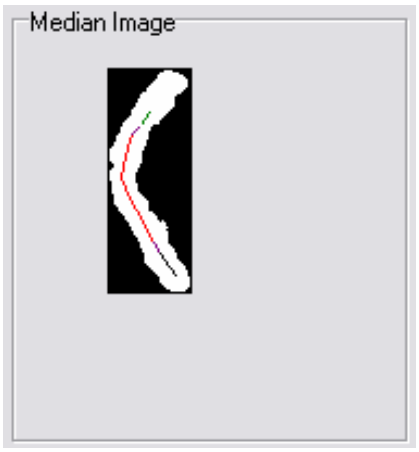



圖 18 圖 16 經過連線的結果

完成以上的步驟之後，染色體的中軸雖然已經擷取出來，但其實各線段之間並沒有順序性，因此接下來的動作便是排列各線段的順序，此外，由圖 17 以及圖 18 可看出，中軸連線的頭尾兩端仍然與染色體真正的端點有一段距離，因此必須延伸中軸的兩端直至與邊界相交。在上一個段落，也就是連線的部分，除了整條染色體中軸的頭尾兩個端點之外的各個端點應該都要與其他線段做連線，因此最後沒有與其他線段做連線的兩個端點即為整條染色體中軸的兩端。藉由這個資訊，便可以在影像中找出這兩個端點，並且令當中 $y$ 值較小的點為染色體中軸的頭端，$y$ 值較大的為中軸的尾端，最後由頭端開始對其他線段建立順序並且計算中軸的長度。

由於染色體的兩端常有彎曲的現象，因此在找端點的時候不能只是依中軸兩

端的斜率做延伸，必須在此延伸的兩邊一定範圍內做搜尋，先找出灰階值變異數最小的方向，而後再沿其垂直方向延伸與染色體邊界相交，其交點即可用來當作染色體真正的端點。以下將敘述詳細的過程。首先，將先前得到的染色體中軸連線頂端第一段連線取出（尾端的作法也一樣）且沿其斜率方向往外做延伸至染色體的邊界，並且在其左右各 45° 角之內每隔 5° 角之處從中軸連線頂端往外做一次延伸直至染色體邊界。在這十九條線段的中點上分別取垂直的線段直至與染色體邊界相交處，計算這些線段上染色體灰階值的變異量，將變異量最小的一段延伸當作正確的染色體端點。圖 19 及圖 20 分別為圖 17 以及圖 18 做中軸延伸後之結果，圖中染色體兩端紅色的部分為有做搜尋的區域。

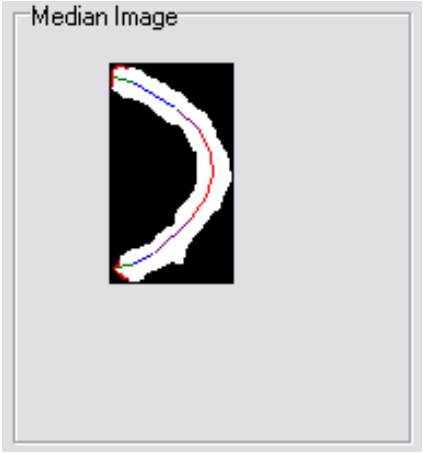



圖 19　圖 17 經過中軸延伸的結果

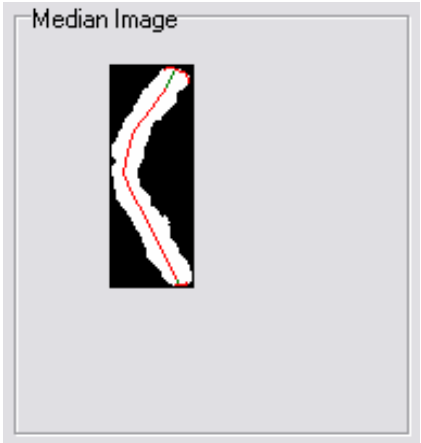



圖 20　圖 18 經過中軸延伸的結果

## 4.3 染色體紋理特徵之擷取

一般在做染色體影像辨識時重要的特徵包括了染色體的相對長度，著絲點的相對位置，以及染色體黑白條紋的分佈。但是由於影像中的染色體常會發生擠壓以及交錯的情況，因此染色體的相對長度並不是一個非常可靠的辨識依據；而我們所使用的影像資料庫中，著絲點的相對位置十分不明顯，同樣不能拿來做辨識依據。因此染色體的黑白紋理分佈在此是最重要的辨識特徵，在此節將會介紹擷取此項重要特徵的方法。

在第三章曾經提到，我們得到的染色體影像有一個較特別的特徵，即中軸部分顏色較淺，中軸兩側的顏色較深。如果只使用前一節所得到的中軸來做灰階值變化的偵測，可能會得到不準確的結果，因此在這裡除了前一節取出的中軸之外，還參考了染色體寬度 1/4 處以及 3/4 處的灰階值變化。實際的作法為，在所有前一節得到的斜率和緩之切片上各取 1/4、1/2、以及 3/4 寬度處的點之灰階值，得到如圖 21 的灰階值分佈曲線圖，圖中的紅色線段為染色體從頂端至尾端所取出的寬度 1/4 處之灰階值變化曲線，而綠色與藍色線段則分別為寬度 1/2 處與 3/4 處之灰階值變化曲線。

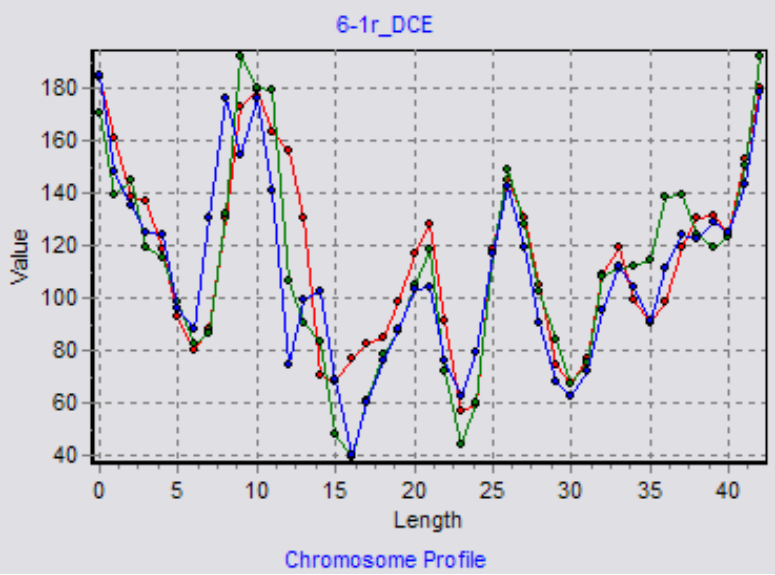



圖 21 一條屬於第六對的染色體之灰階值曲線圖

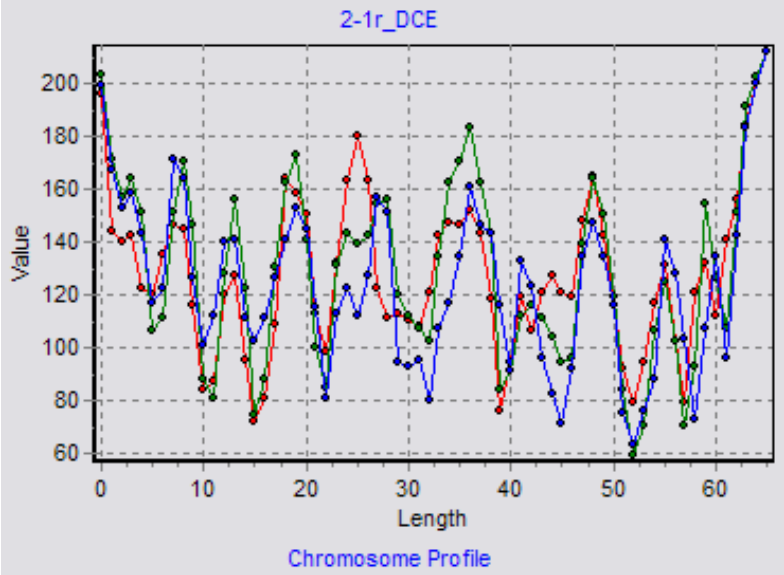



圖 22 一條屬於第二對的染色體之灰階值曲線圖

圖 22 是染色體黑白條紋分佈曲線圖的另一個例子，可以看到寬度分別為 1/4、1/2、以及 3/4 處取得的灰階值在某些地方會有很大的差異，例如圖中長度為 25 的地方，紅點灰階值（寬度 1/4 處）與藍點灰階值（寬度 3/4 處）相差 68；然而在某些地方灰階值又十分相近，例如在長度為 50 處，紅點、綠點、以及藍點的灰階值分別為 116、119、116。為了辨識所需，必須在每個長度上取得合理的代表性單一灰階值，在此我們應用多數決的概念來做處理。圖 23 是圖 22 中的三條灰階值曲線圖經過上述方法融合的結果，簡化後的曲線既可保留原本的資訊，也便於進行下一階段的比對步驟。

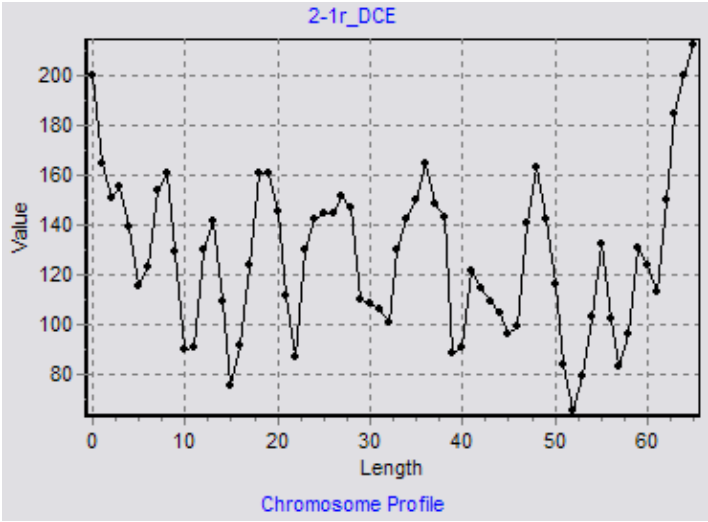



圖 23 圖 22 經過化簡的結果

## 4.4 染色體紋理特徵之比對

在上一節中得到一個細胞中 46 條染色體的灰階值曲線圖之後，便可以用這些資料來做交互比對的動作。在此的交互比對是為了對一個細胞中所有的染色體做配對，而非辨識某條染色體究竟屬於哪一對。由於每條染色體的長度都不同，灰階值範圍也不一樣，不能拿原始的資料來做比對，因此在比對之前必須對所有的資料做標準化的動作，包括了灰階值的標準化以及長度的標準化，其中長度的標準化是和字串對齊同時進行的。在灰階值的標準化方面，首先，算出一條染色體灰階值曲線圖上所有點的平均灰階值以及標準差，以每一點原本的灰階值減掉平均灰階值，所得到的結果再除以標準差，則會得到每一點的 *Z-Score*，在過程中，我們也記錄 *Z-Score* 的最大值與最小值。標準化動作完成之後，接著對所有的 *Z-Score* 做量化（quantization），對大於零的值取不小於它的最小整數，對小於零的值取不大於它的最大整數，而原本為零的值還是為零。圖 24 是某細胞中的第三對染色體灰階值曲線圖經過標準化以及量化之後的結果。

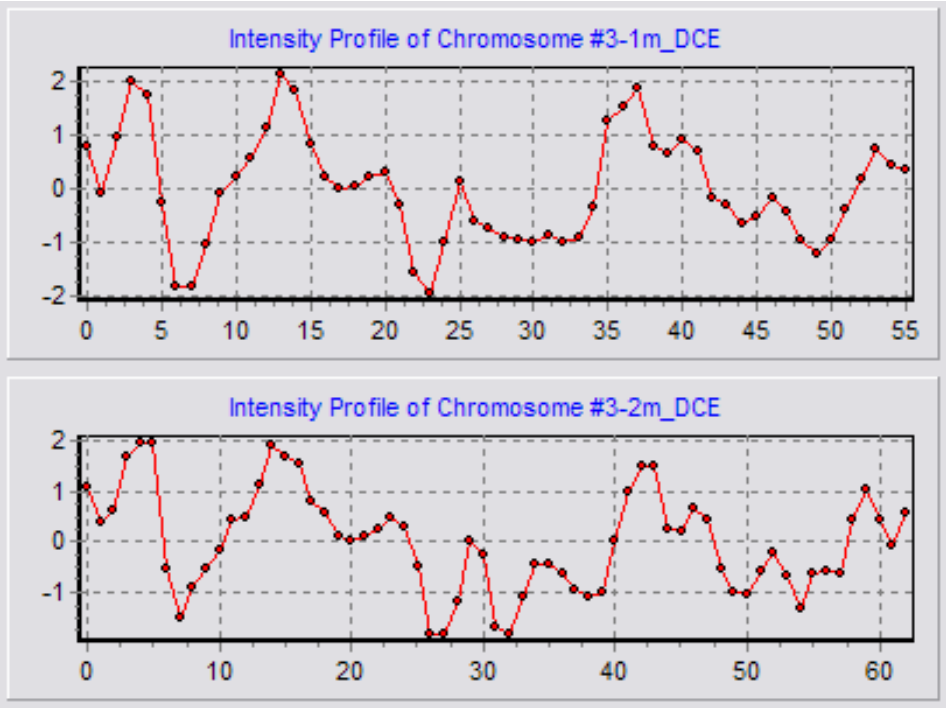



圖 24 某細胞中第三對染色體灰階值曲線圖經過標準化以及量化的結果

由圖 24 中可看出，雖然這兩條染色體為同一細胞中的同對染色體，但其長度並不相同，無法做點對點的交互比對，因此接下來的動作便是對長度做標準化。通常在做長度的標準化時，會採用內插的方式來增加點數，使得長度一致；然而在此我們發展基於字串對齊（substring matching）的演算法，一方面對齊兩條灰階值曲線的特定位置灰階值，同時也調整兩條灰階值曲線至一樣的長度。傳統的字串對齊演算法並沒有相似度的概念，因此不同處會盲目遞補入空格（gap），做出來的結果並不是很理想，如圖 25（$y$ 軸為零的點即為 gap），可以很明顯地看出來紅色線段 5～9 的區間應該對齊綠色線段 11～15 區間，然而並沒有對上。因此在這裡對這個方法加了一些改良的機制，引入相似度的觀念，相差較大的字元給予較低的比重，相差較小的字元給予較大的比重，且將其前後各一個字元也列入考量。我們的比對方法同時考慮到區域性的相似度以及全域的分佈，圖 26 是得到的結果之一例，我們可發現，效果明顯比傳統的字串對齊演算法來得好。圖 27 是改良過後的另一個例子。


圖 25 某細胞中第五對染色體灰階值曲線圖做字串對齊後的結果

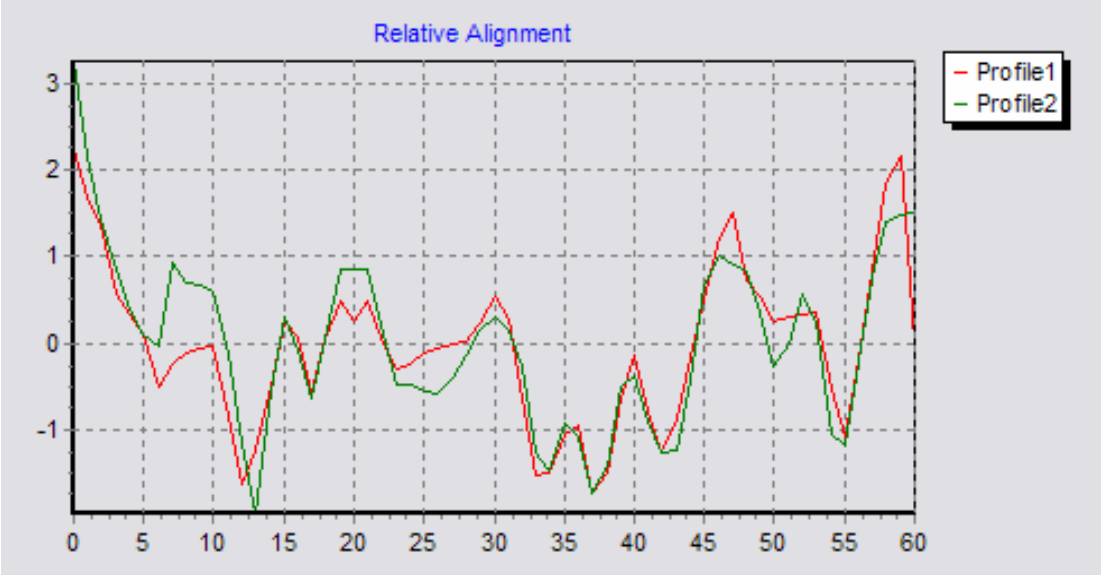



圖 26 圖 25 經過改良後的結果

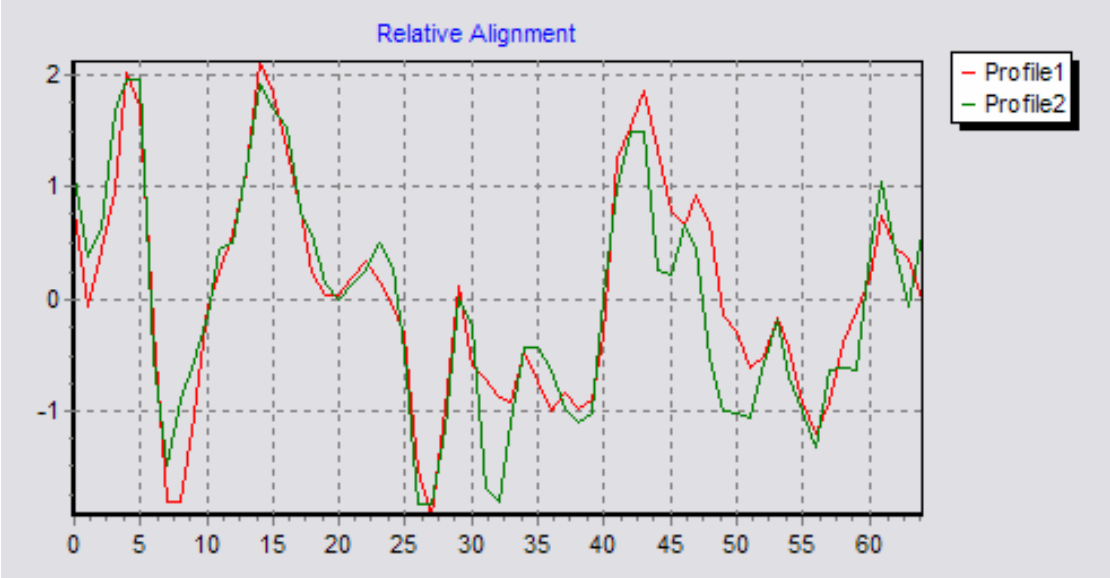
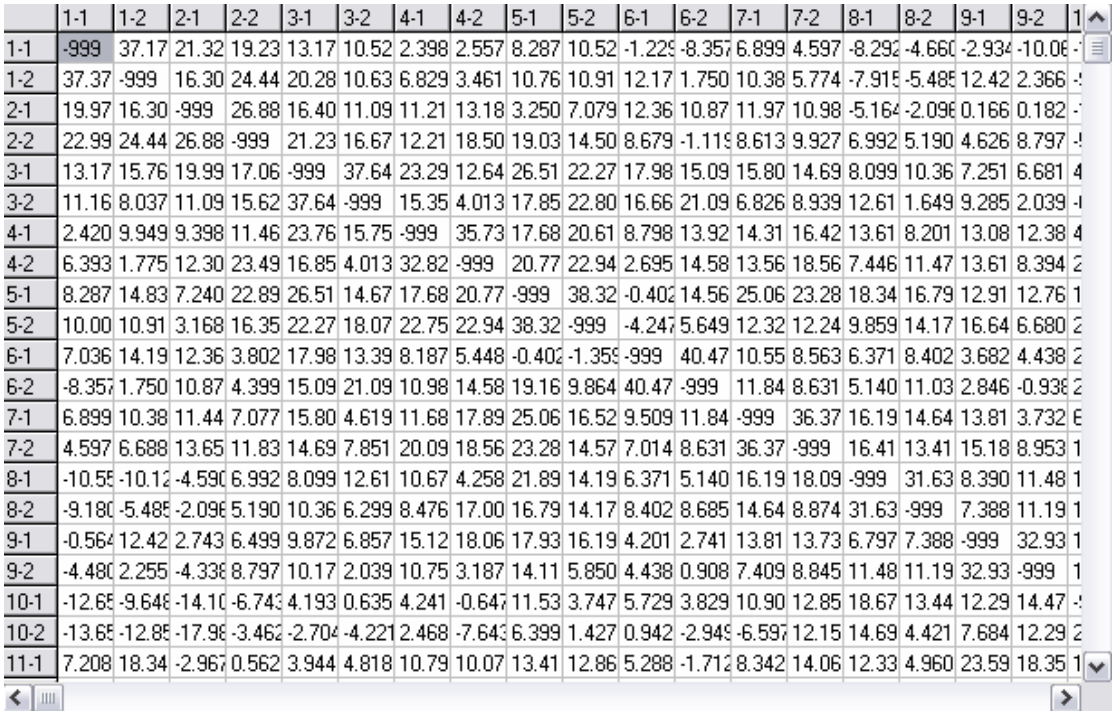



圖 27 某一細胞中第三對染色體灰階值曲線經過改良字串對齊演算法處理後的結果

從圖 26 以及圖 27 可以看出，用來比對的兩條染色體長度一致，且區域性的頂點也大致上都對齊了，因此接下來的步驟便是計算其關連性（correlation）。計算關連性的方法是將兩條曲線上每兩個相對應的點之值相乘，並且加總，便可以得到關連性係數。必須注意的是，在做字串對齊時是使用插補的方法來使得兩條曲線長度一致，若兩條染色體原本的長度相差很多，則插補的地方也就越多，而兩條染色體長度相差越多，實際上屬於同一對的可能性也就越小，因此計算出關

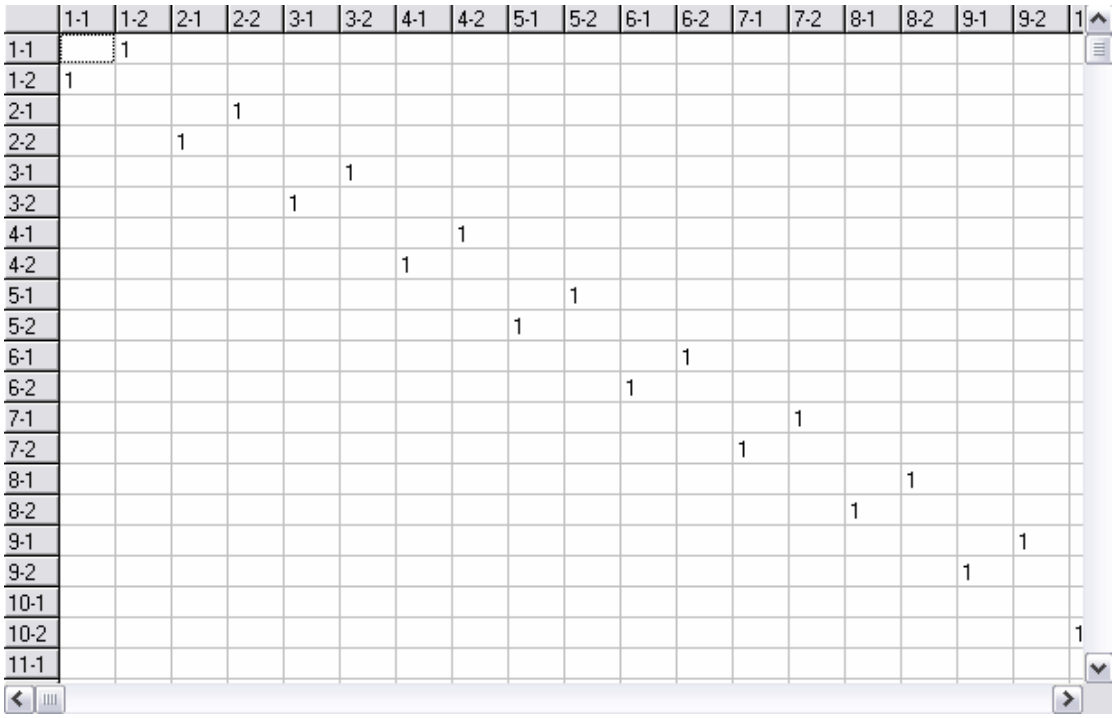
連性之後應該減去補零的影響，亦即補零的數目越多，關連性越小。

圖 28 是對某組資料計算關連性之後的結果，第一行以及第一列皆為此組資料中所有染色體影像的檔名，開頭數字相同的為同對染色體，但不一定是從同一個細胞中取出。格子中的資料便是兩兩染色體曲線計算關連性的結果。由於在此的目的是要找出一組一組屬於同對的染色體，彼此間關連性分數最高的可能就是同一對，為了方便觀察，將圖 28 中每一列的最大值以 1 表示，其他值則刪除，結果如圖 29。



| | 1-1 | 1-2 | 2-1 | 2-2 | 3-1 | 3-2 | 4-1 | 4-2 | 5-1 | 5-2 | 6-1 | 6-2 | 7-1 | 7-2 | 8-1 | 8-2 | 9-1 | 9-2 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-1 | -999 | 37.17 | 21.32 | 19.23 | 13.17 | 10.52 | 2.398 | 2.557 | 8.287 | 10.52 | -1.229 | -8.357 | 6.899 | 4.597 | -8.292 | -4.660 | -2.934 | -10.08 | - |
| 1-2 | 37.37 | -999 | 16.30 | 24.44 | 20.28 | 10.63 | 6.829 | 3.461 | 10.76 | 10.91 | 12.17 | 1.750 | 10.38 | 5.774 | -7.915 | -5.485 | 12.42 | 2.366 | - |
| 2-1 | 19.97 | 16.30 | -999 | 26.88 | 16.40 | 11.09 | 11.21 | 13.18 | 3.250 | 7.079 | 12.36 | 10.87 | 11.97 | 10.98 | -5.164 | -2.096 | 0.166 | 0.182 | - |
| 2-2 | 22.99 | 24.44 | 26.88 | -999 | 21.23 | 16.67 | 12.21 | 18.50 | 19.03 | 14.50 | 8.679 | -1.119 | 8.613 | 9.927 | 6.992 | 5.190 | 4.626 | 8.797 | - |
| 3-1 | 13.17 | 15.76 | 19.99 | 17.06 | -999 | 37.64 | 23.29 | 12.64 | 26.51 | 22.27 | 17.98 | 15.09 | 15.80 | 14.69 | 8.099 | 10.36 | 7.251 | 6.681 | 4 |
| 3-2 | 11.16 | 8.037 | 11.09 | 15.62 | 37.64 | -999 | 15.35 | 4.013 | 17.85 | 22.80 | 16.66 | 21.09 | 6.826 | 8.939 | 12.61 | 1.649 | 9.285 | 2.039 | - |
| 4-1 | 2.420 | 9.949 | 9.398 | 11.46 | 23.76 | 15.75 | -999 | 35.73 | 17.68 | 20.61 | 8.798 | 13.92 | 14.31 | 16.42 | 13.61 | 8.201 | 13.08 | 12.38 | 4 |
| 4-2 | 6.393 | 1.775 | 12.30 | 23.49 | 16.85 | 4.013 | 32.82 | -999 | 20.77 | 22.94 | 2.695 | 14.58 | 13.56 | 18.56 | 7.446 | 11.47 | 13.61 | 8.394 | 2 |
| 5-1 | 8.287 | 14.83 | 7.240 | 22.89 | 26.51 | 14.67 | 17.68 | 20.77 | -999 | 38.32 | -0.402 | 14.56 | 25.06 | 23.28 | 18.34 | 16.79 | 12.91 | 12.76 | 1 |
| 5-2 | 10.00 | 10.91 | 3.168 | 16.35 | 22.27 | 18.07 | 22.75 | 22.94 | 38.32 | -999 | -4.247 | 5.649 | 12.32 | 12.24 | 9.859 | 14.17 | 16.64 | 6.680 | 2 |
| 6-1 | 7.036 | 14.19 | 12.36 | 3.802 | 17.98 | 13.39 | 8.187 | 5.448 | -0.402 | -1.359 | -999 | 40.47 | 10.55 | 8.563 | 6.371 | 8.402 | 3.682 | 4.438 | 2 |
| 6-2 | -8.357 | 1.750 | 10.87 | 4.399 | 15.09 | 21.09 | 10.98 | 14.58 | 19.16 | 9.864 | 40.47 | -999 | 11.84 | 8.631 | 5.140 | 11.03 | 2.846 | -0.938 | 2 |
| 7-1 | 6.899 | 10.38 | 11.44 | 7.077 | 15.80 | 4.619 | 11.68 | 17.89 | 25.06 | 16.52 | 9.509 | 11.84 | -999 | 36.37 | 16.19 | 14.64 | 13.81 | 3.732 | E |
| 7-2 | 4.597 | 6.688 | 13.65 | 11.83 | 14.69 | 7.851 | 20.09 | 18.56 | 23.28 | 14.57 | 7.014 | 8.631 | 36.37 | -999 | 16.41 | 13.41 | 15.18 | 8.953 | 1 |
| 8-1 | -10.55 | -10.12 | -4.590 | 6.992 | 8.099 | 12.61 | 10.67 | 4.258 | 21.89 | 14.19 | 6.371 | 5.140 | 16.19 | 18.09 | -999 | 31.63 | 8.390 | 11.48 | 1 |
| 8-2 | -9.180 | -5.485 | -2.096 | 5.190 | 10.36 | 6.299 | 8.476 | 17.00 | 16.79 | 14.17 | 8.402 | 8.685 | 14.64 | 8.874 | 31.63 | -999 | 7.388 | 11.19 | 1 |
| 9-1 | -0.564 | 12.42 | 2.743 | 6.499 | 9.872 | 6.857 | 15.12 | 18.06 | 17.93 | 16.19 | 4.201 | 2.741 | 13.81 | 13.73 | 6.797 | 7.388 | -999 | 32.93 | 1 |
| 9-2 | -4.480 | 2.255 | -4.338 | 8.797 | 10.17 | 2.039 | 10.75 | 3.187 | 14.11 | 5.850 | 4.438 | 0.908 | 7.409 | 8.845 | 11.48 | 11.19 | 32.93 | -999 | 1 |
| 10-1 | -12.65 | -9.648 | -14.10 | -6.743 | 4.193 | 0.635 | 4.241 | -0.647 | 11.53 | 3.747 | 5.729 | 3.829 | 10.90 | 12.85 | 18.67 | 13.44 | 12.29 | 14.47 | - |
| 10-2 | -13.65 | -12.85 | -17.98 | -3.462 | -2.704 | -4.221 | 2.468 | -7.643 | 6.399 | 1.427 | 0.942 | -2.949 | -6.597 | 12.15 | 14.69 | 4.421 | 7.684 | 12.29 | 2 |
| 11-1 | 7.208 | 18.34 | -2.967 | 0.562 | 3.944 | 4.818 | 10.79 | 10.07 | 13.41 | 12.86 | 5.288 | -1.712 | 8.342 | 14.06 | 12.33 | 4.960 | 23.59 | 18.35 | 1 |

圖 28 對某組資料做關連性計算後的結果



| | 1-1 | 1-2 | 2-1 | 2-2 | 3-1 | 3-2 | 4-1 | 4-2 | 5-1 | 5-2 | 6-1 | 6-2 | 7-1 | 7-2 | 8-1 | 8-2 | 9-1 | 9-2 | 10-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-1 | | 1 | | | | | | | | | | | | | | | | | |
| 1-2 | 1 | | | | | | | | | | | | | | | | | | |
| 2-1 | | | | 1 | | | | | | | | | | | | | | | |
| 2-2 | | | 1 | | | | | | | | | | | | | | | | |
| 3-1 | | | | | | 1 | | | | | | | | | | | | | |
| 3-2 | | | | | 1 | | | | | | | | | | | | | | |
| 4-1 | | | | | | | | 1 | | | | | | | | | | | |
| 4-2 | | | | | | | 1 | | | | | | | | | | | | |
| 5-1 | | | | | | | | | | 1 | | | | | | | | | |
| 5-2 | | | | | | | | | 1 | | | | | | | | | | |
| 6-1 | | | | | | | | | | | | 1 | | | | | | | |
| 6-2 | | | | | | | | | | | 1 | | | | | | | | |
| 7-1 | | | | | | | | | | | | | | 1 | | | | | |
| 7-2 | | | | | | | | | | | | | 1 | | | | | | |
| 8-1 | | | | | | | | | | | | | | | | 1 | | | |
| 8-2 | | | | | | | | | | | | | | | 1 | | | | |
| 9-1 | | | | | | | | | | | | | | | | | | 1 | |
| 9-2 | | | | | | | | | | | | | | | | | 1 | | |
| 10-1 | | | | | | | | | | | | | | | | | | | |
| 10-2 | | | | | | | | | | | | | | | | | | | 1 |
| 11-1 | | | | | | | | | | | | | | | | | | | |

圖 29 圖 28 經過簡化的結果

從圖 29 中可以觀察到，第一對到第九對中的每兩條染色體都能找到與自己屬於同對的染色體，例如與 1-1 關連性最大的為 1-2，且與 1-2 關連性最大的為 1-1。在計算關連性的時候，我們還需引入雙向確認機制，以處理同一對染色體不一定會對彼此有最大的關連性的問題，例如 *a* 染色體應該與 *b* 染色體同對，卻與 *c* 染色體有最大的關連性。在此，我們用遞迴的方式處理這種情況，在第一輪得到的關連性結果中，將雙向都互為最大值的染色體兩兩挑出，剩下的染色體再做第二輪的關連性計算，以此類推，直至所有染色體都找到與其能夠雙向配對的染色體為止。如此，上述的問題或許可獲得解決，若 *c* 染色體與 *d* 染色體互為彼此關連性的最大值，便會被排除在下一輪的計算之外，那麼 *a* 染色體重新對剩下的染色體做關連性計算，可能就會找到真正與其同對的 *b* 染色體。

將所有完成配對的染色體與標準資料庫中的染色體做同樣的關連性計算，便可完成染色體影像辨識的工作。

# 第五章　實驗結果

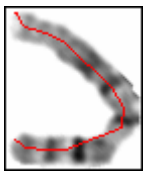
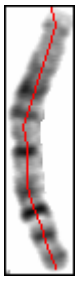
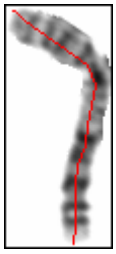
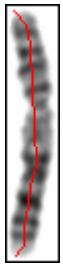
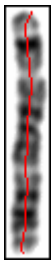
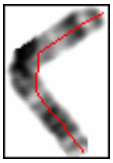
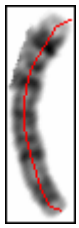
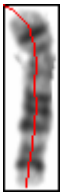
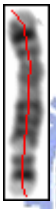
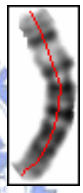
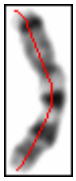
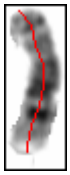
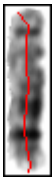
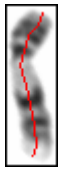
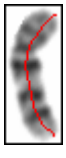
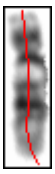
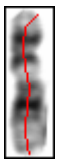
我們所提出的染色體影像自動辨識方法，在本章經過許多組染色體影像的實驗之後，可以證明此方法的確具有一定的可行性與正確性。在實驗中所使用的影像資料是來自國泰醫院所採集的細胞樣本，經過捷揚光電公司所研發之電子顯微鏡下所攝得的影像，且捷揚光電公司提供了影像的前置處理，因此在我們資料庫中的染色體影像已經過切割，亦即一個影像檔案中只有一條染色體。而整個實驗的硬體環境是在 Pentium Ⅳ 1700 MHz，作業系統為 Windows XP，程式作業環境為 Borland C++ Builder 6.0 的平台下操作。本實驗在時間上的數據僅提供一粗略的估計，在實際使用上仍有許多調整的空間。本章實驗分為兩個部分：

- 染色體中軸擷取
- 紋理特徵比對

下面分別針對這兩個主題做詳細的實驗與分析探討。

## 5.1　染色體中軸擷取

在 4.2 節中，我們詳細敘述了中軸擷取演算法，其過程包括了四個方向切片的產生、同方向且連續的切片中點相連、去除重複的連線、線段相連以及中軸的延伸。圖 30 為某一細胞中所有染色體經過中軸擷取之後的結果，紅色線段為中軸的大概位置。由於尚未經過微調的步驟，因此某些彎曲較嚴重的染色體（例如 1-1 染色體與 3-2 染色體）在彎曲部分的中軸線段可能會偏離真正的中軸一段距離。從圖 30 的四十六個影像中可以看出，除了彎曲得比較嚴重的幾條染色體之外，其他染色體擷取出來的中軸結果都不錯。

| | | | |
|---|---|---|---|
|  |  |  |  |
| 1-1 | 1-2 | 2-1 | 2-2 |
|  |  |  |  |
| 3-1 | 3-2 | 4-1 | 4-2 |
|  |  |  |  |
| 5-1 | 5-2 | 6-1 | 6-2 |
|  |  |  |  |
| 7-1 | 7-2 | 8-1 | 8-2 |
|  |  |  |  |
| 9-1 | 9-2 | 10-1 | 10-2 |
|  |  |  |  |
| 11-1 | 11-2 | 12-1 | 12-2 |

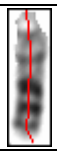
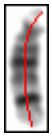
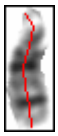
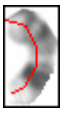
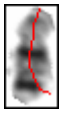
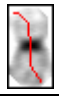
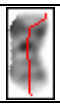
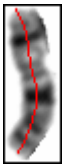
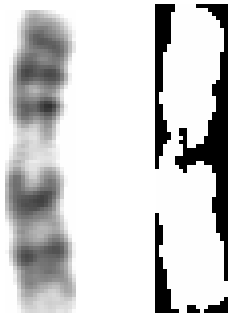
| | | | |
|---|---|---|---|
|  |  |  |  |
| 13-1 | 13-2 | 14-1 | 14-2 |
|  |  |  |  |
| 15-1 | 15-2 | 16-1 | 16-2 |
|  |  |  |  |
| 17-1 | 17-2 | 18-1 | 18-2 |
|  |  |  |  |
| 19-1 | 19-2 | 20-1 | 20-2 |
|  |  |  |  |
| 21-1 | 21-2 | 22-1 | 22-2 |
|  | |  | |
| x | | y | |

圖 30 某細胞中所有染色體經過中軸擷取的結果

另一方面，在我們的資料庫中也有少許染色體在經過中軸演算法處理之後，無法成功取得中軸的例子，亦即找不到中軸或是找到錯誤的中軸。如第三章所述，圖 31 中的染色體，因為在擷取其二值影像時閥值設定不適當，以致於其二值影像中的染色體邊界不正確，染色體某一部份寬度太細，所以演算法無法找到正確的中軸。圖 32 中的染色體也是一個不成功的例子，失敗的原因可能有兩個，一是長軸長度過短，一開始無法取得足夠數量的切片，一是切片中點連線的長度無法達到先前所設定的閥值。此外，在圖 33 中雖然有找到染色體中軸，但是因

為在做中軸兩端延伸的步驟時所取的角度不夠，因此無法找到染色體真正的端點。針對以上這些錯誤，可以藉由參數的調整，來得到更好的結果。



（a）　　（b）

圖 31 （a）無法成功擷取中軸之染色體 （b）（a）的二值化影像

圖 32 無法成功擷取中軸的另一例

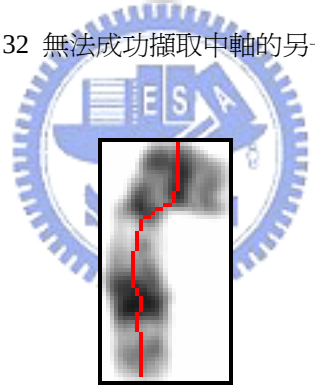
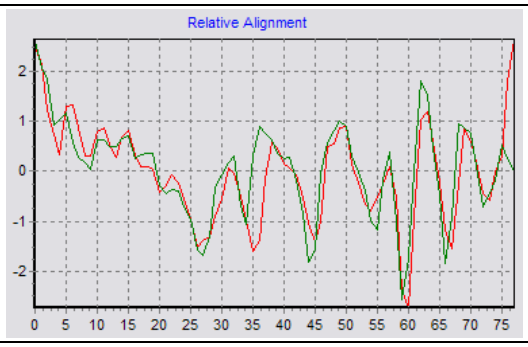


圖 33 中軸擷取錯誤之一例

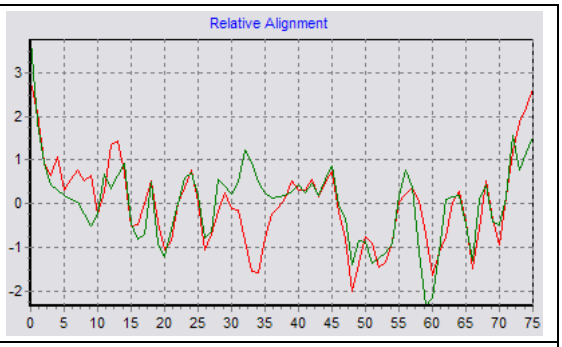
## 5.2　　紋理特徵比對

得到所有染色體的中軸並且微調之後，便可以在中軸上每隔固定距離取一次灰階值，以作為比對之用，有關於中軸的微調、灰階值曲線圖的繪製、以及比對的方法在 4.3 節和 4.4 節中有詳細的敘述。在這裡，我們將前一節所用的細胞拿來做比對，圖 34 是每一對染色體中的兩條（性染色體除外）經過改良後的字串

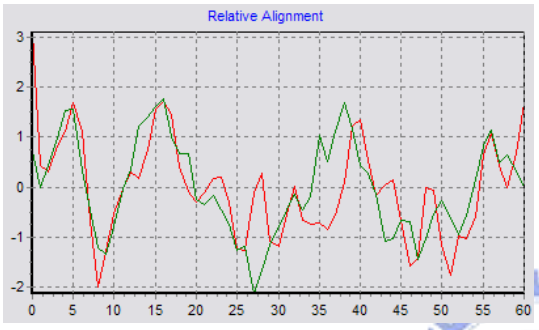
對齊演算法處理的結果。



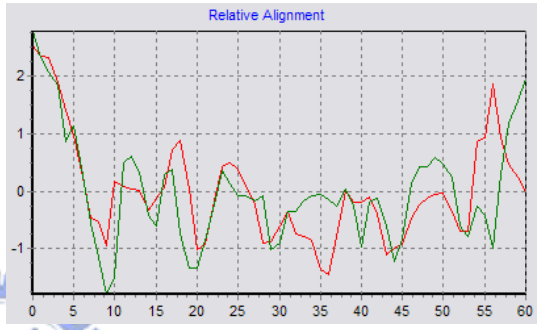

第一對染色體



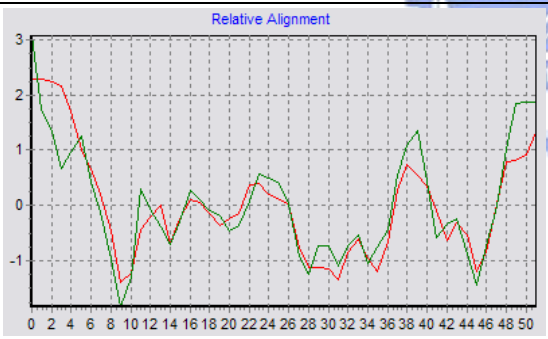

第二對染色體



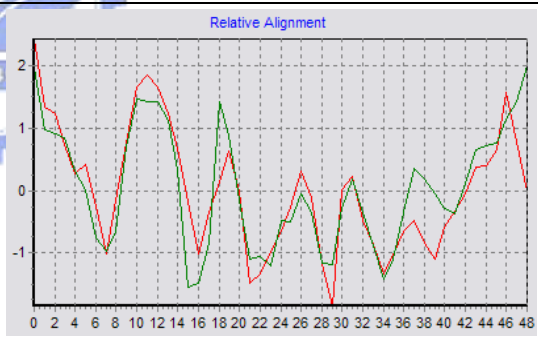

第三對染色體



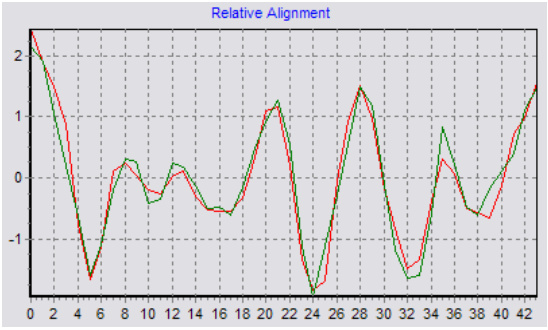

第四對染色體



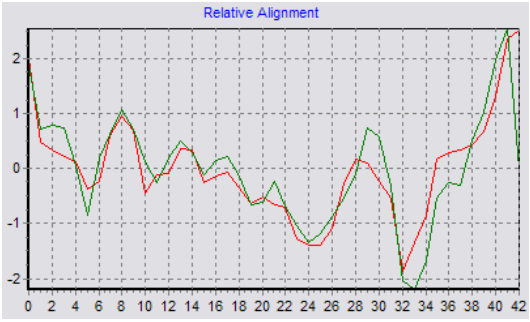

第五對染色體



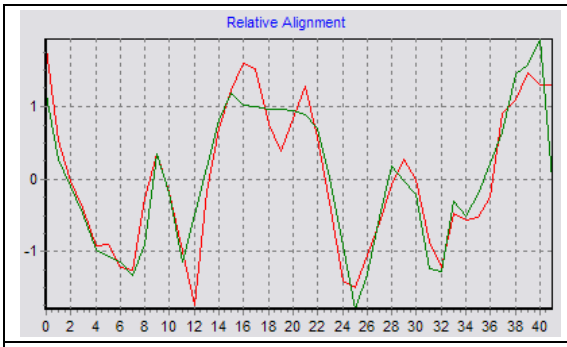

第六對染色體



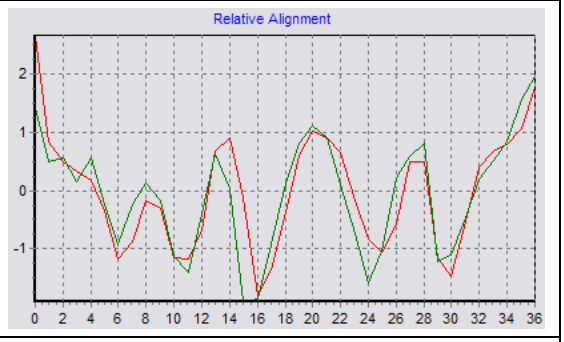

第七對染色體



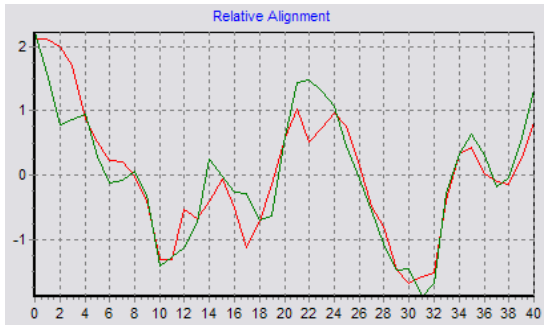

第八對染色體



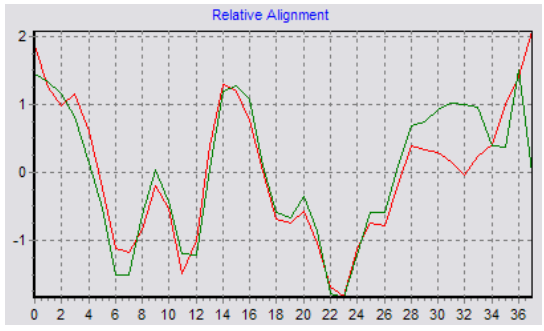

第九對染色體



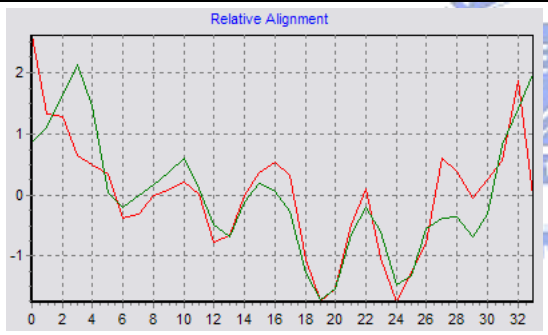

第十對染色體



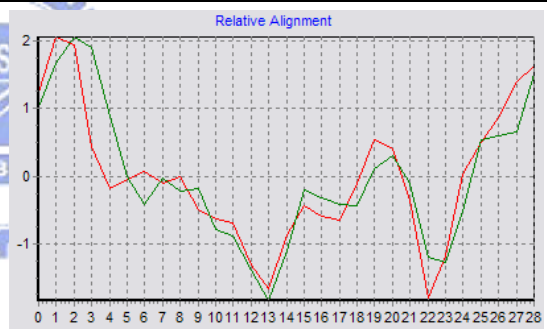

第十一對染色體



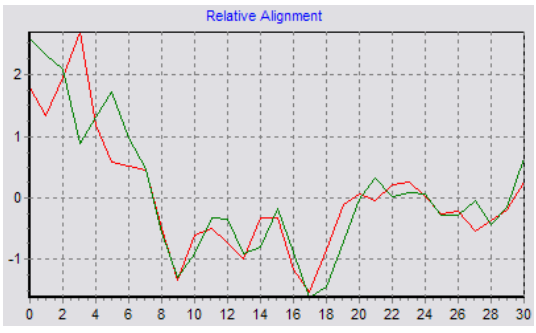

第十二對染色體



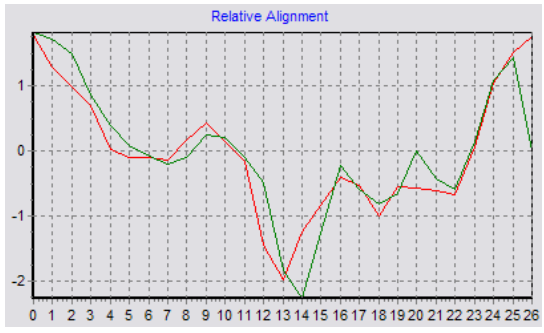

第十三對染色體




第十四對染色體




第十五對染色體




第十六對染色體

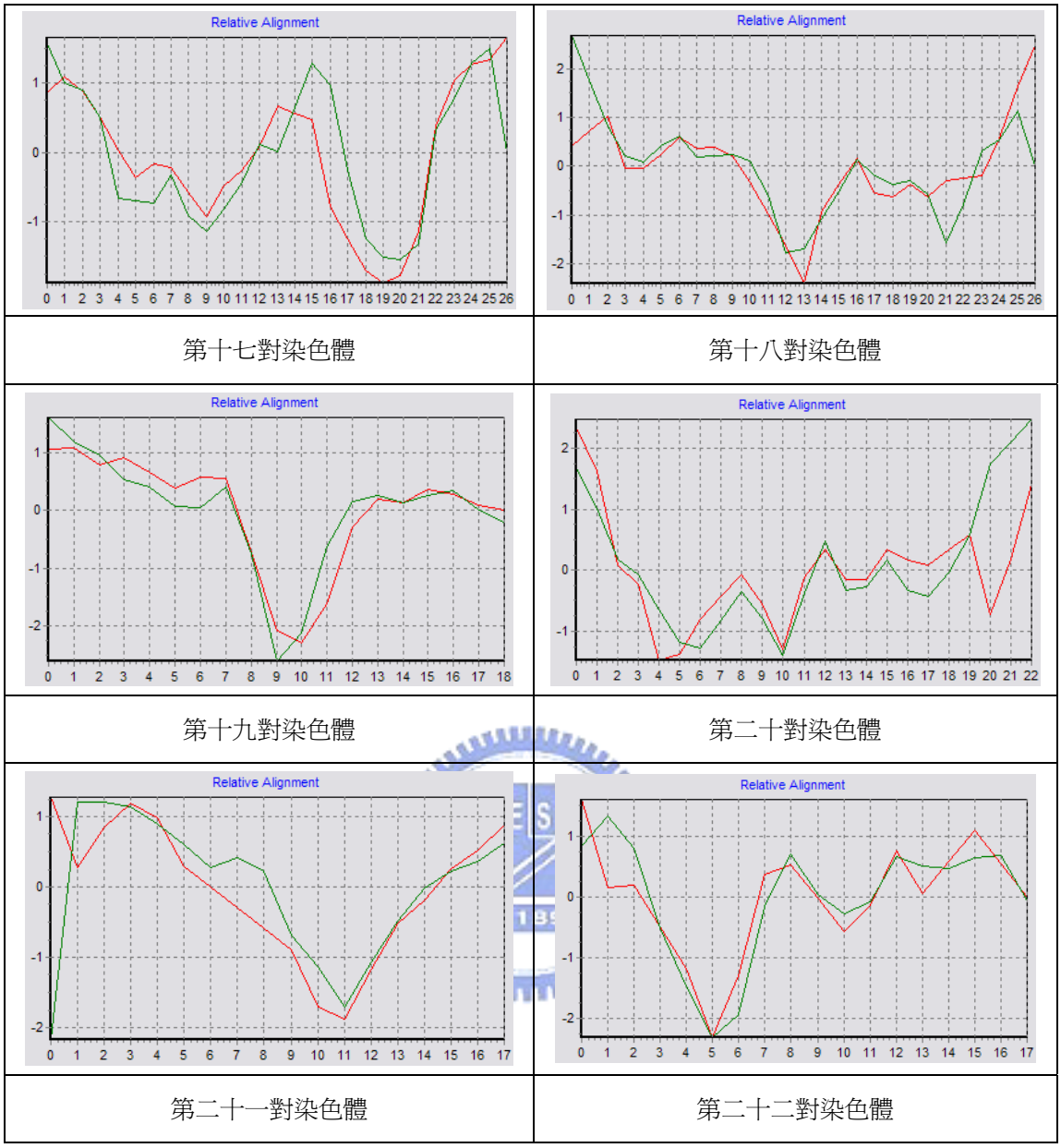



圖 34 各對體染色體做字串對齊的結果

由圖 34 可以看出來，各對染色體對齊的情況都不錯，區域性的最大值與最小值大體上都能夠對齊，因此接下來的步驟便是計算一個細胞中所有染色體之間的關連性。以這個細胞來說，做關連性計算的結果，二十二對體染色體中有二十對能夠得到正確的配對結果（第十六對與第十八對配對錯誤），因此，此細胞在配對的部分正確率為 90.91%。

完成了各細胞中染色體配對的步驟後，接著便是將每一對染色體與標準染色體做比對，以得到染色體組型。然而標準染色體資料庫的建置尚未完成，因此目

前我們由十八組細胞資料中經由人工挑出形狀較完整且灰階值較清楚的染色體組成一組標準染色體，當中有二十二條體染色體（每對體染色體取其中的一條）以及 x、y 染色體各一條。將這組標準染色體與資料庫中的十組染色體做比對，結果得到 82.56%的正確率，證明在此所使用的演算法有不錯的辨識率。表 1 為此十個細胞做比對的詳細資料，其中類別個數欄位裡的數字代表此細胞中可以成功擷取出中軸的染色體對數，由於性染色體可以分為兩個類別，因此類別個數的最大值為 24。

| | 類別個數 | 分類錯誤的染色體 | 正確率 |
|---|---|---|---|
| 1 | 24 | 6, 8, 10, 13, 16, x, y | 70.83% |
| 2 | 23 | 8, 10, 15, 20, y | 78.26% |
| 3 | 24 | 14, 18 | 91.67% |
| 4 | 24 | 14, 16, 17, 20, 22, y | 75.00% |
| 5 | 24 | 20, y | 91.67% |
| 6 | 21 | 13, 16 | 90.48% |
| 7 | 24 | 6, 8, 13, 14, 15, 16, 18 | 70.83% |
| 8 | 23 | 10, 15 | 91.30% |
| 9 | 23 | 8, 10, 14, 16, 22, y | 73.91% |
| 10 | 24 | 13, 16 | 91.67% |

# 第六章　結論與未來展望

在這篇論文中，我們提出了一套完整的流程，能夠自動化地辨識染色體影像，而不用人為的操控。這一套流程包括了染色體中軸的擷取和微調，黑白紋理特徵分佈的萃取，以及兩兩染色體之間的配對和辨識。由實驗結果可以看出來，這套演算法對於我們資料庫中的染色體之辨識效果不錯，無論在中軸的擷取或是關連性的計算部分，都有高度的準確性。

然而，資料庫中仍有少數的染色體在做中軸擷取時無法找到適當的中軸，或是在比對時無法找到同一對的染色體。造成這些錯誤的原因通常為染色體的長度太短，或是對灰階值影像做二值化時，閥值的設置不恰當，造成二值化影像有破碎的情形。針對前者，可以藉由參數的調整來得到較佳的結果；然而針對後者，因為二值化的影像是資料庫中給定的，除非資料庫中的影像有所改善，否則難以得到更好的結果。

由於目前我們的資料庫中細胞數目不足，無法做大量的實驗來驗證結果，也無法找出更多的例外情況來針對程式做改善，因此未來的目標便是建立更龐大的染色體影像資料庫，並且配合程式中各個參數的調整，以期使這套流程能夠有更好的染色體影像辨識率，並且能夠應用在形狀更多元的染色體影像上。

此外，在第二章中提到的類神經網路，可以藉由學習演算法不斷提升辨識率，因此若在我們的流程中加入學習的機制，會使得系統更完整，所以這也是未來展望中重要的項目。

# 參考文獻


[1] B. Lerner, “Toward a Completely Automatic Neural Network Based Human Chromosome Analysis,” *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 8, issue 4, pp. 544-552, Aug. 1998.

[2] R. J. Stanley, J. M. Keller, P. Gader, and C. W. Caldwell, “Data-Driven Homologue Matching for Chromosome Identification,” *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, Jun. 1998.

[3] J. M. Cho, “Chromosome Classification Using Backpropagation Neural Networks,” *IEEE Magazine of Engineering in Medicine and Biology*, vol. 19, issue 1, pp. 28-33, Jan.-Feb. 2000.

[4] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, “A Comparison of Multilayer Perceptron Neural Network and Bayes Piecewise Classifier for Chromosome Classification,” *IEEE International Conference on Neural Networks*, vol. 6, pp. 3472-3477, Jun.-Jul. 1994.

[5] X. Wu, P. Biyani, S. Dumitrescu, and Q. Wu, “Globally Optimal Classification and Pairing of Human Chromosomes,” *IEEE Conference of Engineering in Medicine and Biology Society*, vol. 1, pp. 2789-2792, Sep. 2004.

[6] P. Mousavi, S. S. Fels, R. K. Ward, and P. M. Lansdorp, “Classification of Homologous Human Chromosome Using Mutual Information Maximization,” *IEEE International Conference on Image Processing*, vol. 2, pp. 845-848, Oct. 2001.

[7] Q. Wu and K. R. Castleman, “Automated Chromosome Classification Using Wavelet-Based Band Pattern Descriptors,” *IEEE Symposium on Computer-Based Medical Systems,* pp. 189-194, Jun. 2000.

[8] G. Ritter and G. Schreib, “Profile and Feature Extraction from Human Chromosome,” *IEEE International Conference on Pattern Recognition*, vol. 2, pp. 287-290, Sep. 2000.

[9] J. M. Conroy, R. L. Becker, W. Lefkowitz, K. L. Christopher, R. B. Surana, T. O'Leary, D. P. O'Leary, and T.G. Kolda, “Hidden Markov Models for Chromosome Identification,” *IEEE Symposium on Computer-Based Medical Systems*, pp. 473-477, Jul. 2001.

［10］M. Moradi, S. K. Setardhdan, and S. R. Ghaffari, “Automatic landmark detection on chromosomes' images for feature extraction purposes,” *IEEE International Symposium on Image and Signal Processing and Analysis*, vol. 1, pp. 567-570, Sep. 2003.

［11］P. Mousavi, R. K. Ward, P. M. Lansdorp, and S. S. Fels, “Multi-feature analysis and classification of human chromosome images using centromere segmentation algorithms,” *IEEE International Conference on Image Processing*, vol. 1, pp. 152-155, Sep. 2000.

［12］N. Sweeney, R. L. Becker, and B. Sweeney, “A comparison of wavelet and Fourier descriptors for a neural network chromosome classifier,” *IEEE International Conference on Engineering in Medicine and Biology society*, vol. 3, pp. 1359-1362, Oct.-Nov. 1997.

［13］G. Ramstein, M. Bernadet, A. Kangoud, and D. Barba, “A Rule-based Image Analysis System For Chromosome Classification,” *IEEE International Conference on Engineering in Medicine and Biology Society*, vol. 3, pp. 926-927, Oct.-Nov. 1992.

［14］J. Piper, E. Granum, “On Fully Automatic Feature Measurement for Banded Chromosome Classification,” *Cytometry*, vol. 10, pp. 242-255, 1989.