第九屆旺宏科學獎

成果報告書

參賽編號：SA9-012

作品名稱：以人工智慧方法探討 DNA 序列比對

姓名：楊鎮榮

關鍵字：模糊聚類、鹼基序列、種屬鑑定

摘要

本文藉由模糊理論及模式識別的理論研究，開發設計出一套 DNA 自動分類辨識演算法，再由此方法進行生物數值分析，之後再以農業改良場所提供的資料，判斷雜交前後的托鞋蘭的親緣關係作為測試樣本，進行生物實驗檢定。

在此我們利用了不同原生種的拖鞋蘭分別作為學習樣本，再以雜交過後的拖鞋蘭樣本作為未知樣本進行分類辨識，並檢驗我們的方法是否具有實際臨床實驗的正確性及價值，本方法主要是以人工智慧中的模糊理論作為架構設計出一組數學模型，並利用電腦程式進行模擬。

首先提取拖鞋蘭基因中核醣體序列中的 ITS1~ITS2 進行分析，並利用(聚合酶連鎖反應法)PCR 技術大量複製，最後再由膠體電泳進行解序。在量化分析之前必須先將電泳儀分析之 ATCG 做前置處理，在此 DNA 序列會先轉換成胺基酸序列，並將各種胺基酸作為特徵，計算其出現頻率，以利數學模型進行 DNA 序列區別，最後利用歐式距離，來對雜交種的樣本與原生種的樣本進行檢核。

在 DNA 解序後，DNA 序列中出現了非 ATCG 字元，在與農改場博士以及老師討論之後，發現是解序過程中因為 DNA 分子量大小過於接近，導致儀器無法判讀是何種鹼基，在此會事先進行交叉檢驗，若還是有部分錯誤，將在數值分析時，作為極小誤差不予計算。

在程式模擬之前，我們也討論到了 FCM(Fuzzy C-means)以及 KM(K-means)的差異性，並在文中有做優缺點比較，在分類時選用準確性較高的方式，且將所有的特徵參數都考量進去，增加分析準確性。

由於我們由農改場所獲得的資料已經過處理，是片段 DNA(ITS 序列)且已知起點位置，所以在數值分析時，跳過滾動的步驟，也就是將 DNA 序列轉換成胺基酸序列時，只會轉換成一條(已知起始點)。

第一次先取 6 筆原生種 *P. acmodontum*、*P. charlesworthii*、*P. concolor*、*P. randsii*、*P. conco-bellatulum*、*P. rothschildianum* 作為及一筆雜交種 *P. rothschildianum* X *P. delenatii* 做為測試樣本進行分析。第二次我們擴大範圍以 14 筆原生種 *P. armeniacum*、*P. bellatulum*、*P. chamberlainianum*、*P. concolor*、*P. glaucophyllum*、*P. haynaldianum*、*P. lowii*、*P. micranthum*、*P. purpuratum*、*P. rothschildianum*、*P. sukhakulii*、*P. urbanianum*、*P. victoria-mariae*、*P. villosum* 進行測試檢定。

本次研究中，發現利用此方法，可以分辨出拖鞋蘭親代子代雜交後相互的親緣關係，此點驗證此本案所提方法之有效性。並相較以往只針對蛋白質、胺基酸等作區別或各級結構分類，本次結果可以說此次的研究有著突破的結果。

壹、研究動機

過去對於動、植物種屬鑑定所使用之方法，多以「型態觀察」為主，即根據動、植物之外觀與特殊形狀構造或功能來判別其為何種動、植物，但此種鑑定所需的條件極為嚴格，必須有完整的動、植物外型；或具有該種類動、植物特徵部位，才得以鑑定。而以現今生物科技應用於動、植物鑑識，應可期待具有極高的價值。因 DNA 記錄著所有生物體之遺傳特徵，不同物種間具有不同的基因組成，甚至於同種間之不同個體亦可以 DNA 分析加以區別，近幾年利用 DNA 序列資料作為動、植物種系之鑑定，已逐漸成為一種趨勢。

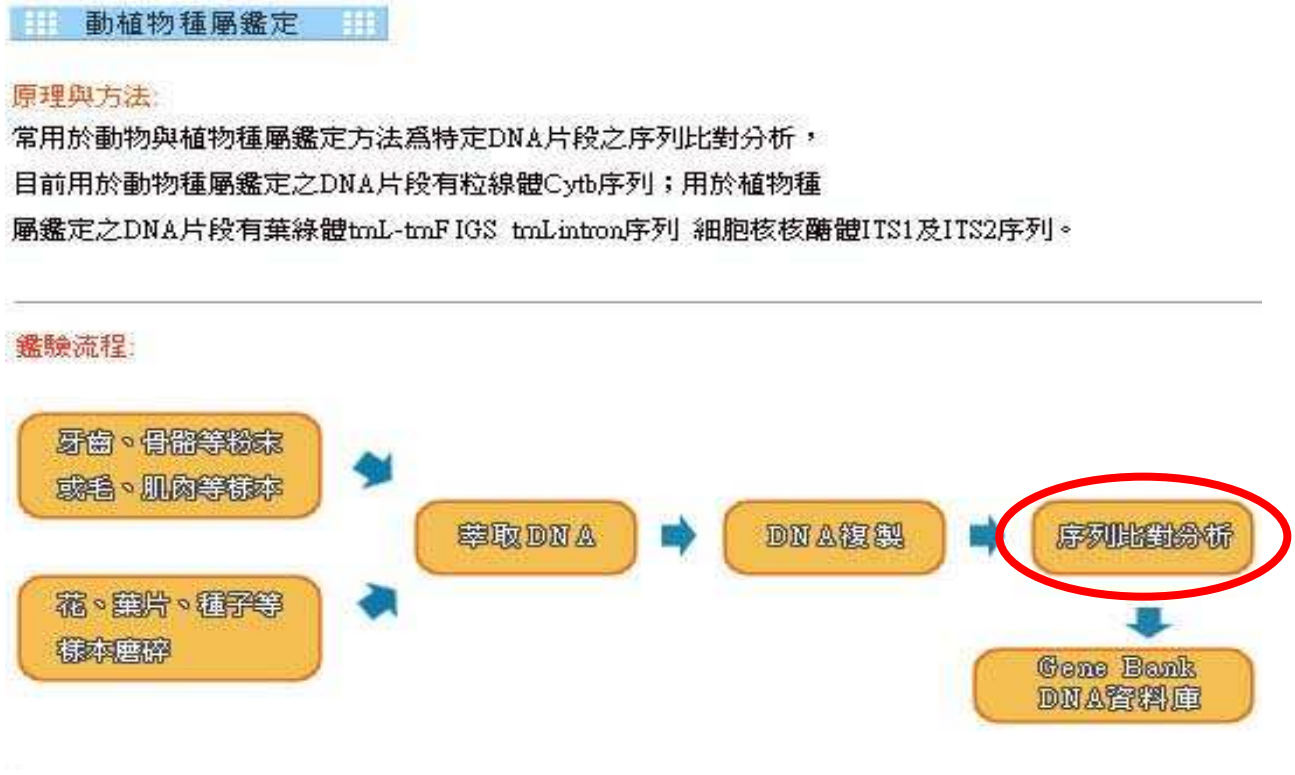
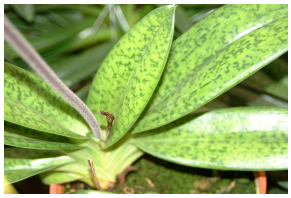


圖 1-1：動植物種屬鑑定流程圖

本研究即是利用序列比對開始進行改良[8]，並透過台中縣農業改良場所提供的 DNA 樣本來檢驗所提方法的正確性。目前在這項研究中最普通的思想是省略序列的某些細節，突出特徵，然後將其表示成適當的數學對象[1]。本文嘗試以模組化作為研究 DNA 序列的架構，並提出對序列集合進行分類的數值分析，也在 2010 年台灣區國際科展電腦科學中發表，且獲得電腦科學科的大會三等獎[9]。

現在我們更結合實際的植物樣本**拖鞋蘭**，如下圖 1-2 至 1-9，來進行臨床檢驗，以驗證本方

法在實際進行基因體分析及親緣關係的判斷時，具有良好的性質，並將相關結果發表於此次研究。

			
圖 1-2：*Beren-Genev-3*	圖 1-3：*Beren-Genev-4*	圖 1-4：*Hirsutissimum-3*	圖 1-5：*Hirsutissimum-4*
			
圖 1-6：*Hookerae-3*	圖 1-7：*Hookerae-4*	圖 1-8：*Sukha-roths-4*	圖 1-9：*Sukha-roths-5*

貳、研究目的

種屬鑑定在法醫學及生物學中佔有重要地位，現行用於法醫學種屬鑑定的方法主要有形態學、血清學和 DNA 分析技術，本文就種屬鑑定的技術做一改善並提出新的 DNA 分析方法，以改良現有的分析技術，並利用經濟植物拖鞋蘭的雜交，來驗證我們方法在實務上的可行性。

拖鞋蘭別名：仙履蘭、芭菲爾鞋蘭、兜蘭，英文名稱：Lady's slipper，花型異於其它蘭類，獨樹一格，花型花色的變化也多，拖鞋蘭最早被發現時，養蘭人士均傾倒於拖鞋似的奇特花瓣，同時也以此外型作為命名的依據，像是瑞典的植物學之父－林奈(Carl Linnaeus)將它命名為『女神之足』，英國人則稱它為『淑女的拖鞋』(Lady's slipper)，國人則稱它為拖鞋蘭或仙履蘭，其中以拖鞋蘭的稱呼最為普遍。因此國內、外不少農業研究機構均嘗試以雜交及其他品種改良方式來進行品種改良。

在此我們利用生物資訊學的系統聚類方式(K-Means)設計出一組分類演算法[1][7][10]，結合人工智慧理論的模糊理論應用於本研究中[9]；即利用模糊聚類方法 Fuzzy C-Means(FCM)[3][7][9][11]，藉由行政院台中農業改良場所提供的 6 種原生種拖鞋蘭及雜交後的 1 種雜交種拖鞋蘭，設計出程式及進行相關模擬分析，並且獲得結果：

(1)設計一組目標函數使其具有分類 DNA 序列功能的模糊分類演算法。

(2)驗證結合模糊理論的分類方法的有效性是否具體提昇[9]。

(3)臨床實驗樣本分析由 6 種原生種拖鞋蘭來雜交出 1 種雜交種的拖鞋蘭，再利用本研究所提的雜訊過濾方式，將非 ATCG 的其他因電泳分析過程所產生的儀器誤差予以修正，再依本文所提的人工智慧理論予以數值分析，期望驗證本方法於遺傳鑑定及種屬的鑑定有效性。

(4)進而再增加到以 14 種原生種托鞋蘭分類一種雜交種的托鞋蘭，如本次分類結果。

參、研究過程

一、理論基礎

DNA 鑑定方法

DNA 鑑定是從動植物的生物個體內分泌物中採取出 DNA(含有基因情報的細胞內的染色體)，以進行分析識別確認同一性的方法。其鑑定方法如圖 3-1-1 及圖 3-1-2[8]，在本研究中我們實際前往行政院農委會台中農業改良場，跟隨改良場內的研究員學習 DNA 的相關分析方式。

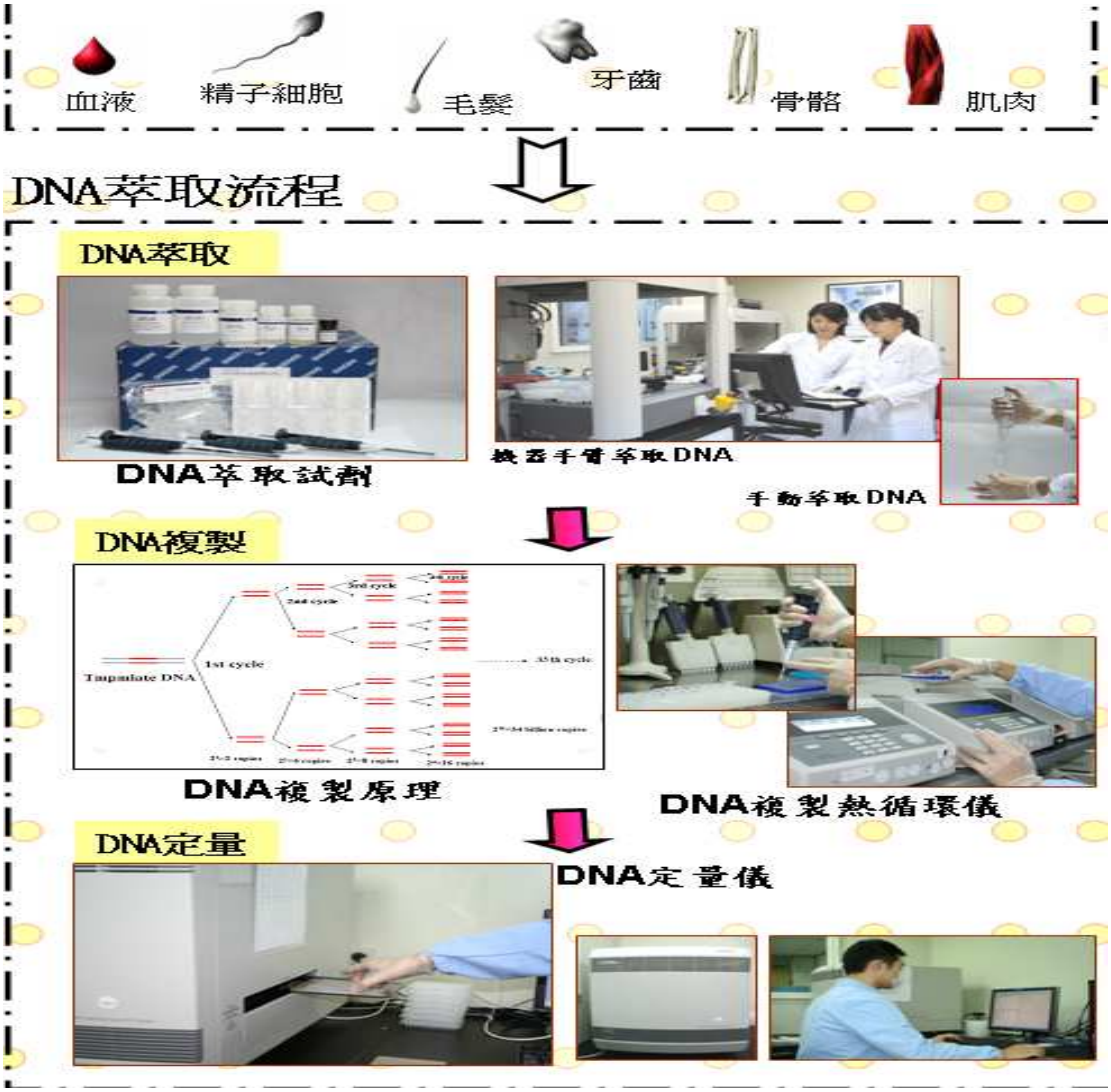


圖 3-1-1：DNA 鑑定方法 1

圖 3-1-2：DNA 鑑定方法 2

DNA 脫氧核糖核酸(Deoxyribonucleic acid，縮寫為 DNA)可組成遺傳指令，其組成為是腺嘌呤（A)、鳥嘌呤（G)、胞嘧啶（C）與胸腺嘧啶（T)。通常 DNA 序列具有高相似性，代表兩序列源自相同的祖先，並具有相同的空間結構，生物學家稱為同質性，而同質性序列通常具有類似的生化功能。生物學上定義，若蛋白質中超過 25%的胺基酸序列相同，或 DNA 中超過 75%的含氮鹼基序列相同，便幾乎可以確定蛋白質及 DNA 序列具有同質性，此點亦可作為親緣判定的參考[2]。

RNA 核糖核酸(Ribonucleic acid，縮寫為 RNA)是存在於細胞生物的遺傳訊息中間載體，RNA 的鹼基主要有四種，即腺嘌呤(A)、鳥嘌呤(G)、胞嘧啶(C)和尿嘧啶(U)，並參與蛋白質合成遺傳訊息的傳導流向為 DNA 到蛋白質，科學家認定 RNA 可以攜帶遺傳訊息。

蛋白質(Protein)是由胺基酸分子呈線性排列所形成，並透過形成肽鍵連接在一起，蛋白質的胺基酸序列是由對應基因所編碼。主要是遺傳密碼所編碼的 20 種「標準」胺基酸。在 DNA 遺傳序列轉變成為蛋白質序列的過程中，必須加入 RNA 序列，也就是 DNA 必須先轉錄(transcription)成 RNA，RNA 再轉譯(translation) 成蛋白質，如圖 3-1-3[4][13]。

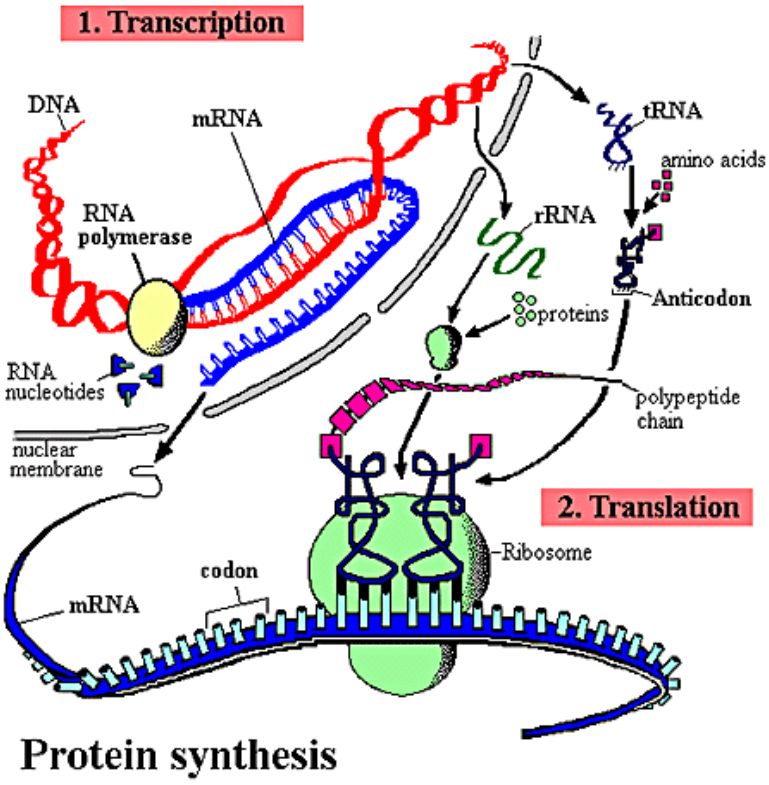
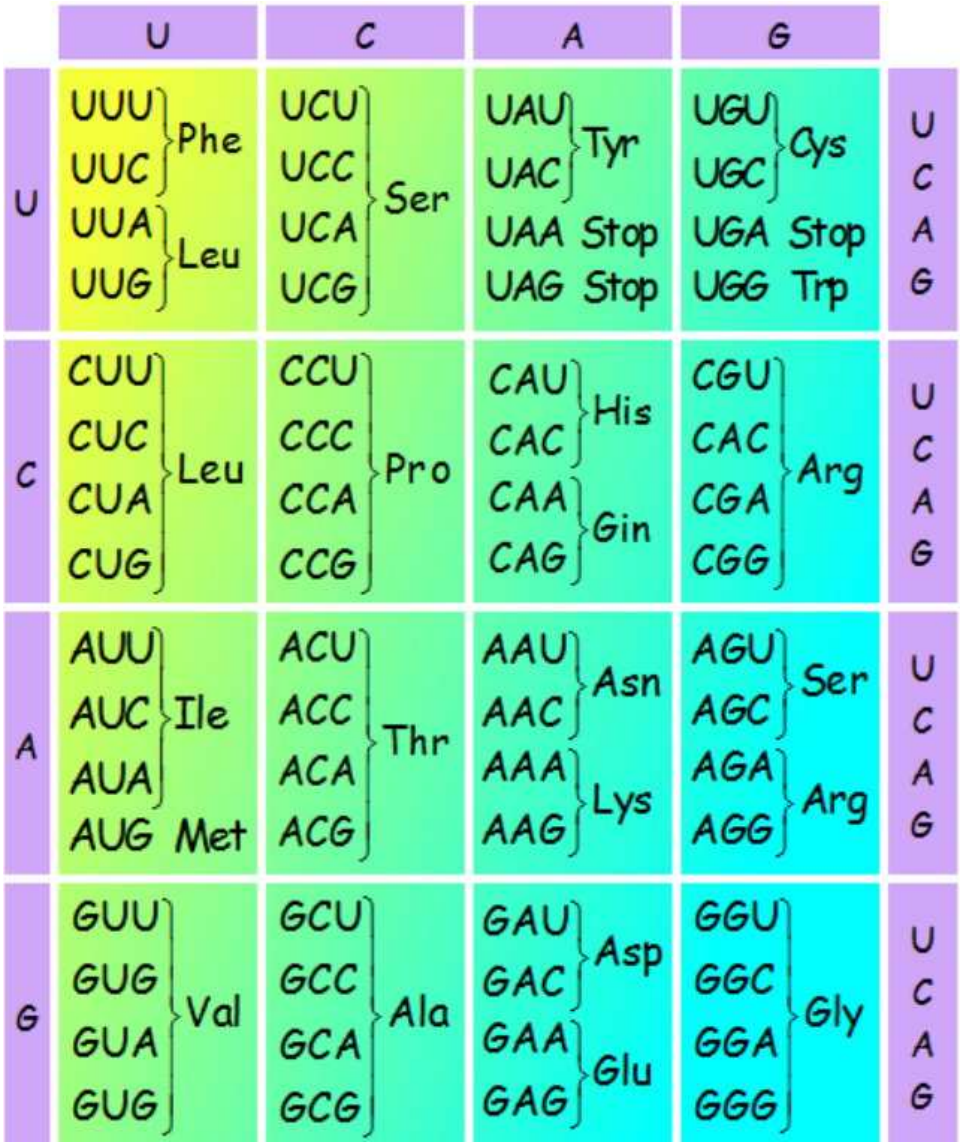


圖 3-1-3：轉譯及轉錄過程

鹼基序列對蛋白質序列的轉換機制

密碼子 Codon 是以三個 DNA 或 RNA 鹼基為一組，它暗示了鹼基與蛋白質互換的機制及生成原理。因胺基酸有 20 種，而 DNA 或 RNA 各有 4 種鹼基，它們之間要如何配對呢？若以兩個為一組來對應一個胺基酸，則只能產生 16 種組合(4^2=16)，顯然無法對應 20 種胺基酸；若以三個為一組，則可產生 64 種組合(4^3=64)；若四個為一組，則可產生 256 種組合(4^4=256)。所以有可能是以三個或四個為一組。最後生物學家利用噬菌體的交配來發現 DNA 的語言應該是以 3 個字為一組。所以 DNA 是以三個字為單位，來產生 64 種不同的組合，並利用多對一的函數映射關係對應到 20 種胺基酸[14]。

表 3-1-1：遺傳密碼表



	U	C	A	G	
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gin CAG Gin	CGU Arg CAC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUG Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

遺傳密碼表

遺傳密碼(表 3-1-1)所示，甲硫胺酸(Methionine)是一般通用的起始密碼子(initiation codon)，然而有極少數的生物例外，是使用 GUG 作為起始密碼子。UAA、UAG、UGA 是終止密碼子(stop codon)，它們並不對應任何胺基酸，就好像是句子中的「句點」一樣，當轉譯時遇到終止密碼子，轉譯就會停止。因鹼基有 64 個($4^3=64$)遺傳密碼子，卻只有 20 種胺基酸，所以一定有很多重複的對應，像精胺酸(Arginine)是擁有最多重複對應的胺基酸，可由六種不同的密碼子產生。

表 3-1-2：蛋白質常用的 22 組胺基酸

分類	名稱		縮寫	
1. 唯一對稱胺基酸	1. 甘胺酸	Glycine	Gly	G
2. 含飽和碳氫基團	2. 丙胺酸	Alanine	Ala	A
	3. 纈胺酸	Valine	Val	V
	4. 白胺酸	Leucine	Leu	L
	5. 異白胺酸	Isoleucine	Ile	I
3. 含芳香基團	6. 苯丙胺酸	Phenylalanine	Phe	F
	7. 酪胺酸	Tyrosine	Tyr	Y
	8. 色胺酸	Tryptophan	Trp	W
	9. 組胺酸	Histidine	His	H
4. 含額外酸基（及其醯胺）	10. 天冬胺酸	Aspartic acid	Asp	D
	11. 天冬醯胺酸	Asparagine	Asn	N
	12. 麩胺酸	Glutamic acid	Glu	E
	13. 麩醯胺酸	Glutamine	Gln	Q
5. 含額外胺基	14. 離胺酸	Lysine	Lys	K
	15. 精胺酸	Arginine	Arg	R
6. 含有醇基	16. 絲胺酸	Serine	Ser	S
	17. 蘇胺酸	Threonine	Thr	T
	18. OH-脯胺酸	Hydroxy Pro		
7. 含有硫	19. 甲硫胺酸	Methionine	Met	M
	20. 胱胺酸	Cysteine	Cys	C
	21. 雙胱胺酸	Cystine		
8. 環狀的亞胺酸	22. 脯胺酸	Proline	Pro	P

生物是一個資訊的系統，在細胞內的遺傳系統是由非常多不同的分子來運作。蛋白質就像是資訊系統的硬體，而 DNA 則是軟體，主要功能為儲存、複製與傳遞資訊。這兩者之間如何互相結合，主要關鍵就在於翻譯的系統[2][4][9][13]。我們將表 3-1-2 整理成 22 個特徵向量（表 3-1-3），以進行數據分析。

表 3-1-3：胺基酸變數設計

特徵	Ala (A)	Cys (C)	Asp (D)	Glu (E)	Phe (F)	Gly (G)	His (H)	Ile (I)	Lys (K)	Leu (L)	Met (M)
編號	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
特徵	Asn (N)	Pro (P)	Gln (Q)	Agr (R)	Ser (S)	Thr (T)	Val (V)	Trp (W)	Tyr (Y)	*	a+t
編號	x_2	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}

人工智慧理論-模糊理論

模糊理論是由模糊集合(Fuzzy Set)、模糊關係(Fuzzy Relation)、模糊邏輯(Fuzzy Logic)、模糊控制(Fuzzy Control)、模糊量測(Fuzzy Measure)…等理論整合而成。主要是將模糊概念予以量化的學問，起源於 1965 年扎德(L.A. Zadeh) 教授所發表的著名論文－「模糊集合」[7]。文中首次提出表達事物模糊性的重要概念：隸屬函數，並以模糊集合為基礎，奠定模糊理論的基礎，以研究不確定事物為目標，接受模糊現象存在的事實，根據不清晰訊息，透過近似推理(Approximation Reasoning)過程而得到正確結果，這與人腦「過程模糊，結論清晰」的思維方式極為類似，故已被廣泛應用於各種不同領域的智慧型系統中[5][6][7]。

傳統分類與人工智慧理論-模糊分類的比較

傳統分類演算方法可視為模糊分類的特例，其隸屬度在隸屬某一類時視為 1，在非此類時隸屬度視為 0。故兩者的目標函數差異，在傳統分類演算方法為 $J(V)$。而在模糊分類時多了隸屬度，即 $J(V,U)$。因此透過隸屬度的調整，強化了分類的有效性及合理性，此點正是本文利用來修正傳統 DNA 分類演算方法的最佳證明[7]。

特徵的提取

為有效地實現分類識別及親緣判定，故利用數學方法中的模式識別計算方法，先根據被識別的對象產生一組基本特徵，再對基本特徵進行變換，以得到最能反映分類本質的特徵，這就是特徵形成與提取的過程。因此列舉出儘可能完備的特徵參數集，如本文胺基酸變數 $X=[x_1,x_2,\cdots,x_{22}]$。

特徵的形成

因 A、T、C、G 這 4 種字符組成了 64 種不同的 3 字元串，且構成生物蛋白質的 20 種胺基酸，在參考文獻[14]的 Figure 2 中，將這 20 種胺基酸的編碼令為參數 $x_1 \sim x_{20}$。參數 x_{21} 是由 UAA、UAG、UGA 等終止字符串的出現頻率當作一個特徵。因在不用於編碼蛋白質的序列片段中，A 和 T 的含量特別多些，故將 A 和 T 是否特別豐富作為一個特徵，故參數 x_{22} =『A 和 T 出現的頻率之和』(表 3-1-3)。而在計算 3 字元串的出現頻率時，Brian Hayes 於論文中以同一種胺基酸的 3 字元串合成一類，只統計 20 類 3 字元串的出現頻率。若不考慮字元串在序列片段中

的起始位置，也就採用“滾動”算法[12]。(如 ACGUCC 中就有 ACG, CGT, GTC, TCC 共 4 個 3 字元串)。

表 3-1-4：第一筆 DNA 序列 22 個特徵出現頻率表

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	.08	.03	.08	.16	.00	.08	.05	.03	.08	.05	.03	.00	.03	.00	.08	.00	.11	.05	.03	.03	.00	.43
2	.00	.00	.00	.06	.00	.11	.03	.06	.06	.06	.00	.00	.03	.00	.31	.00	.11	.03	.03	.03	.11	.43
3	.08	.00	.03	.03	.00	.36	.03	.00	.06	.11	.00	.08	.00	.00	.08	.06	.06	.00	.00	.03	.00	.43

表 3-1-5：較完整的分析過程表

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
1	編碼	agg	cac	gga	aaa	acg	gga	ata	acg	gag	gag	gac	ttg	gca	cgg	cat	tac	acg	gag	gac	gag	gta	aag	gag	gct	tgt	cta
2	胺基酸	Arg	His	Gly	Lys	Thr	Gly	Ile	Thr	Glu	Glu	Asp	Leu	Ala	Arg	His	Tyr	Thr	Glu	Asp	Glu	Val	Lys	Glu	Ala	Cys	Leu
3																											
4	胺基酸	Ala	Phe	Cys	Asp	Asn	Glu	Gin	Gly	His	Leu	Ile	Lys	Met	Pro	Arg	Ser	Thr	Val	Trp	Tyr	Stop					
5	各數	3	0	1	3	0	6	0	3	2	2	1	3	1	1	3	0	4	2	1	1	0	37				
6	出現頻率	0.081	0	0.027	0.081	0	0.162	0	0.081	0.054	0.054	0.027	0.081	0.027	0.027	0.081	0	0.108	0.054	0.027	0.027	0					
7																											
8	編碼	a	ggc	acg	gaa	aaa	cgg	gaa	taa	cgg	agg	agg	act	tgg	cac	ggc	att	aca	cgg	agg	acg	agg	taa	agg	agg	ctt	gtc
9	胺基酸		Gly	Thr	Glu	Lys	Arg	Glu	Stop	Arg	Arg	Arg	Thr	Trp	His	Gly	Ile	Thr	Arg	Arg	Thr	Arg	Stop	Arg	Arg	Leu	Val
10																											
11	胺基酸	Ala	Phe	Cys	Asp	Asn	Glu	Gin	Gly	His	Leu	Ile	Lys	Met	Pro	Arg	Ser	Thr	Val	Trp	Tyr	Stop					
12	各數	0	0	0	0	0	2	0	4	1	2	2	2	0	1	11	0	4	1	1	1	4	36				
13	出現頻率	0	0	0	0	0	0.056	0	0.111	0.028	0.056	0.056	0.056	0	0.028	0.306	0	0.111	0.028	0.028	0.028	0.111					
14																											
15	編碼	ag	gca	cgg	aaa	aac	ggg	aat	aac	gga	gga	gga	ctt	ggc	acg	gca	tta	cac	gga	gga	cga	ggt	aaa	gga	ggc	ttg	tct
16	胺基酸		Ala	Arg	Lys	Asn	Gly	Asn	Asn	Gly	Gly	Gly	Leu	Gly	Thr	Ala	Leu	His	Gly	Gly	Arg	Gly	Lys	Gly	Gly	Leu	Ser
17																											
18	胺基酸	Ala	Phe	Cys	Asp	Asn	Glu	Gin	Gly	His	Leu	Ile	Lys	Met	Pro	Arg	Ser	Thr	Val	Trp	Tyr	Stop		a	t	a+t		
19	各數	3	0	0	1	3	1	0	13	1	4	0	2	0	0	3	2	2	0	0	1	0	36	33	15	48	
20	出現頻率	0.083	0	0	0.028	0.083	0.028	0	0.361	0.028	0.111	0	0.056	0	0	0.083	0.056	0.056	0	0	0.028	0		0.297	0.135	0.432	
21																											

實驗步驟

1.將拖鞋蘭樣本萃取出 DNA

由植物體中提取 DNA，是先剪取植物體中片段組織，將其打碎之後，再進行 PCR 等反應。而我們主要是提取拖鞋蘭 DNA 中核糖體序列的 ITS1 到 ITS2，這是因為在種屬相近的物種，其全序列的相似度可達 99%，但在這段 DNA 序列中，其差異性特別大，所以一般在進行 DNA 分析時，此段 DNA 常被提取使用。

(1)採集欲取得 DNA 之植物樣本的鮮嫩葉片，取大約 100mg 並加入約 0.2～0.5ml 的萃取緩衝液(Tris-Buffer)開始研磨[註①]。

(2)研磨後，將含有植物樣本的緩衝液滴入試管，開始第一次離心。

(3)第一次離心後，取上層含有樣本 DNA 的澄清液(下層為搗爛的植物纖維素等廢液)滴入半截試管[註②]，並進行第二次離心。

(4)離心後，試管上層膜上殘留樣本 DNA，以清洗緩衝液 1(Washing-buffer1)及清洗緩衝液 2(Washing-buffer2)沖洗，其成分類似酒精，摒除掉一切雜質，並抑制聚合酶。下層為廢液便不再使用它。

(5)將試管再次放入離心機，但不加任何緩衝液，使其離心空轉，目的是使酒精揮發。

(6)取出試管，加入 Elluction-buffer 使 DNA 溶於其中從膜上分離，進行第三次離心。離心後下層之溶液即為植物樣本之 DNA 萃取液。

註①：每次反應所需之緩衝液量，均以生產該藥劑之公司建議為主。

註②：半截試管為一種下半部可拆卸試管，中間有一層膜，可吸附 DNA，將溶液中其他物質分離。

2.DNA 複製(進行 PCR 聚合酶連鎖反應)

PCR 反應需具備的條件和材料，分別如下：

(1)要被複製的 DNA 模板(template)。

(2)界定複製範圍兩端的引子(primer)：可分別和 DNA 的雙股配對結合，作為合成新股的起點。

(3)DNA 聚合酵素(taq polymerase)：將四種核苷酸催化聚合成一新的互補 DNA 鏈。

(4)合成 DNA 的原料：其中包括了 dATP、dTTP、dGTP、dCTP 等四種核苷酸，或是 dNTP(包含全部)。

PCR 是以熱循環進行 DNA 序列片段複製，每一次循環，可得到原數量兩倍的 DNA 序列，其中包含三步驟:

(1)變性：使用高溫(92～95℃)，使雙股 DNA 變性，分開成單股的 DNA，以作為往後複製的模板。

(2)黏合：使引子於一定的溫度下(通常在 40～55℃之間)，黏合於單股 DNA 模板上，作為合成新股的起點。

(3)延展：以單股的 DNA 作為模板，引子為起點，在 DNA 的聚合酵素的催化及適當溫度(約 72℃)的作用下，將四種核苷酸催化聚合成互補的 DNA 鏈。

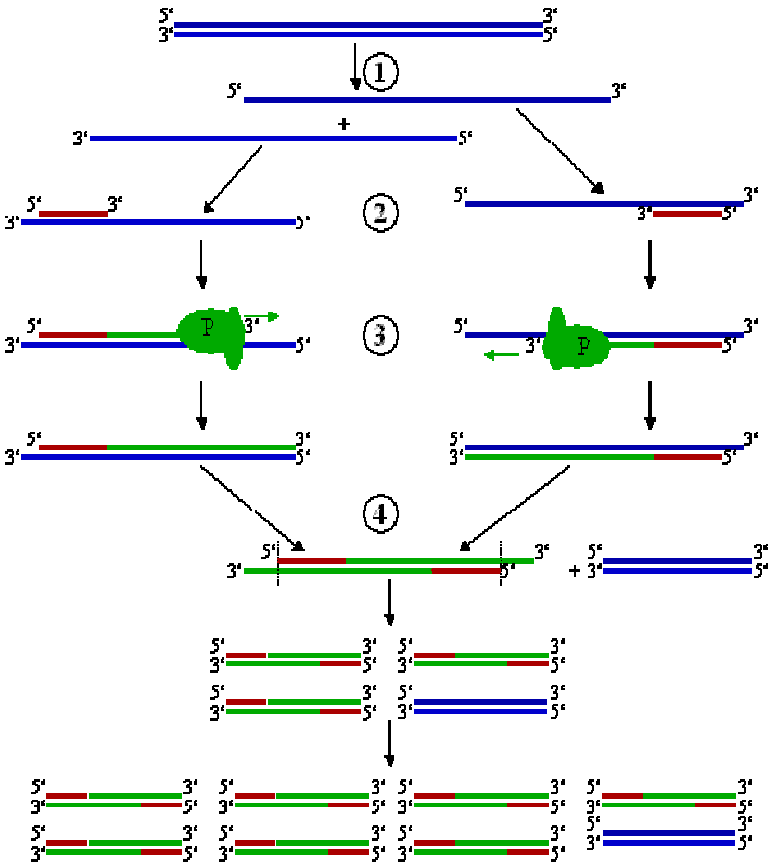


圖 3-1-4：PCR 過程簡圖

PCR 過程簡圖(圖 3-1-4)說明：

①95 度高溫下解開 DNA 雙股。

②約 55 度之下讓引子與 DNA 配對。

③在 72 度引子進行反應，開始合成互補 DNA。(P=聚合酶)

④一循環完成，做出兩段雙股 DNA，每一循環皆可得到雙倍雙股 DNA。如此循環 30 次便可得 2 的 30 次方倍，即 1073741824 倍，也就是大約十億倍。

3.DNA 解序(電泳儀分析)

原理：DNA 分子帶負電荷，當 DNA 置於電場中，它們會朝正極的方向移動。膠體電泳可以根

據 DNA 分子大小的不同來分離它們。

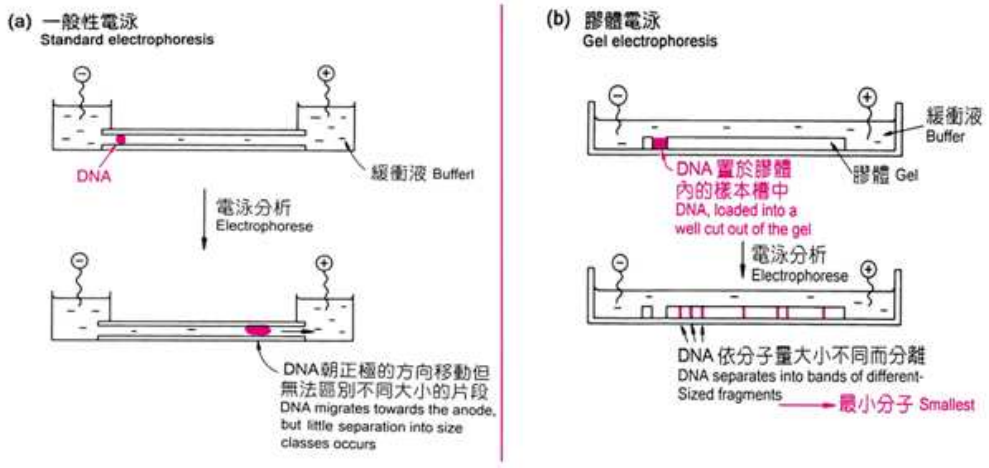


圖 3-1-5：膠體電泳圖

DNA 解序是利用 PCR 以及電泳分析方法，進行 DNA 序列解序之後利用自動定序儀進行序列分析，原理也是利用電泳方式，但是會在尾端放上雷射掃瞄器，掃描加入螢光標記(染色劑為 *IRDye*700^{nm} 或 *IRDye*800^{nm})的鹼基序列，再利用電腦判斷，它便會進行解序。

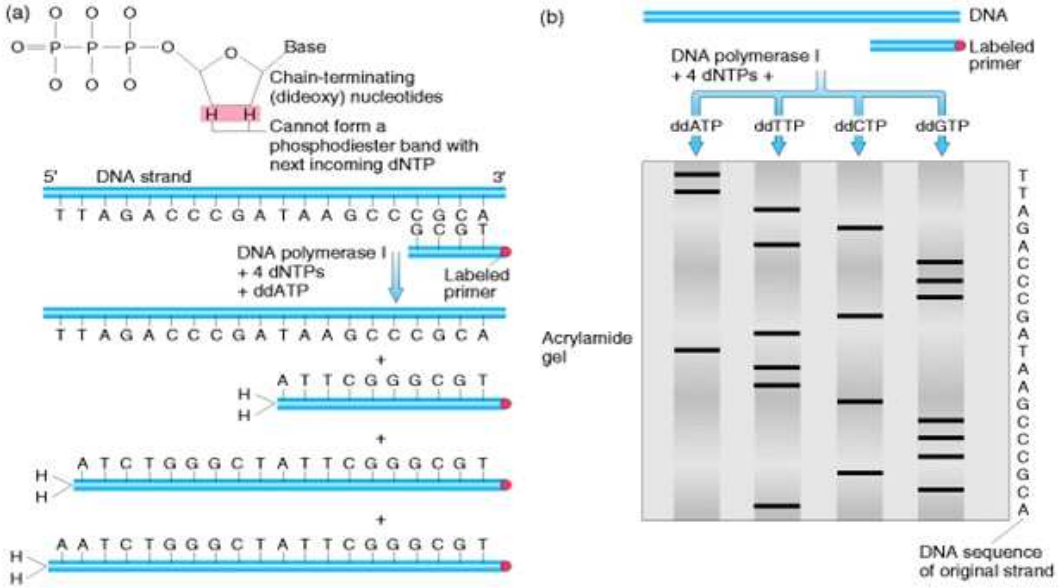


圖 3-1-6：自動定序儀進行序列分析圖

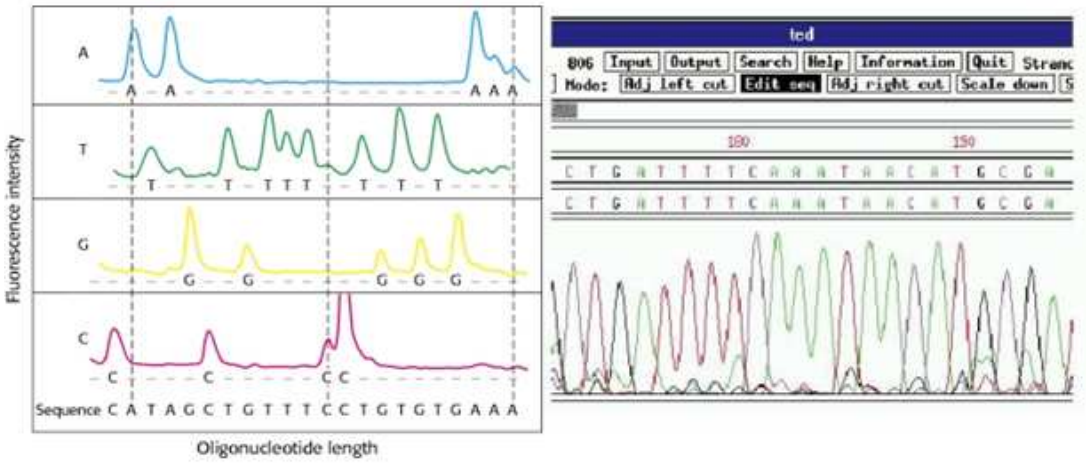


圖 3-1-7：電腦 DNA 定序軟體實際情況

4. 將 DNA/RNA 序列轉成胺基酸序列並將 22 筆特徵進行資料量化

因我們所獲得的資料(如附件 1)是已知序列中的起始點，故**本次不以滾動方式**，直接先將所有鹼基轉成蛋白質中的 20 組胺基酸，並各增加 1 組鹼基轉胺基酸的基因終止符及 A 和 T 的鹼基配對組，以表 3-1-3 胺基酸變數設計，做為本研究的 22 筆特徵向量，再統計其出現的相對次數頻率作為特徵提取之用，將原始字串轉為可供分析之數據。其中前 20 組 DNA/RNA 序列加上基因終止符及 A 和 T 的鹼基配對組，轉成胺基酸序列資料經滾動分析以矩陣表示(式 3.1-1)，透過相對次數轉換方式，將(式 3.1-1)以 66 筆數值表示，先利用 $\mathrm{Sort}(\min:(x_k^i - x_l^j)^2)$、$i \neq j$、$k \neq l$，$i=1,2,...,c$，$k=1,2,...,n$，$j=1,2,...,c$，$l=1,2,...,n$。來求出親緣性最佳的前 22 組滾動參數再加上終止符及 A 和 T 的鹼基配對組，以判斷最佳的變數 $x_k^{(i)}$，即(式 3.2-1)。

$$\begin{bmatrix} x_{1,n}^{(1)} & x_{2,n}^{(1)} & x_{3,n}^{(1)} & \ddots & & x_{22,n}^{(1)} \\ x_{1,n}^{(2)} & x_{2,n}^{(2)} & x_{3,n}^{(2)} & & \ddots & x_{22,n}^{(2)} \\ x_{1,n}^{(3)} & x_{2,n}^{(3)} & x_{3,n}^{(3)} & & & \ddots x_{22,n}^{(3)} \end{bmatrix}$$、n 為樣本數----------(式 3.1-1)

$$\begin{bmatrix} x_{1,n}^{(i)} & x_{2,n}^{(i)} & \cdots & x_{22,n}^{(i)} \end{bmatrix}$$、$i=1,2,...,c$、n 為樣本數----------(式 3.2-1)

經程式分類之後發現，若將終止符及 A 和 T 的鹼基配對組視為段落標號，共有 22 個參數存在，故令 $X_{k,n}^{(i)} = \{x_{1,n}^{(i)}, x_{2,n}^{(i)}, x_{3,n}^{(i)}, \cdots, x_{21,n}^{(i)}, x_{22,n}^{(i)}\}$，$x_{k,n}^{(i)}$ 表示第 k 個特徵在分類中出現的頻率，並將變數的個數調整，使參數維度降至足以代表整體的樣本參數。

5.模糊聚類方式設計

確定代表訊息的胺基酸序列的每一個胺基酸，利用模糊聚類的方式來確定鹼基序列訊息的胺基酸序列，並利用歐氏距離來量測彼此間的實際差距。因模糊聚類 Fuzzy C-Mean(FCM)演算法是利用模糊群集的分類方法，一般在進行模糊分類時，都是假設目標函數為：

$$J(V,U)=\sum_{i=1}^{d}\sum_{j=1}^{c}\sum_{k=1}^{n}u_{i,j,k}^{m}\,|\,x_{i,k}-v_{i.j}\,|^{2}\text{ ----------(式 3.3-1)}$$

其中

$x_1,\cdots,x_n$ 是樣本變數；

d 表示樣本胺基酸序列變數的參數個數；

$V_i=\{v_{i,1},\cdots,v_{i,c}\}$ 是第 i 筆資料，分類的平均值(群心)向量集；

$U_i=[u_{i,k}]$ 為一 $c\times n$ 矩陣，這裡 $u_{i,j,k}$ 是第 k 次輸入樣本的第 j 個隸屬值，且滿足

$$0\le u_{i,j,k}\le 1\quad i=1,2,...,d\ ;\ j=1,2,...,c\ ;\ k=1,2,...,n\text{ ----------(式 3.3-2)}$$

$$\sum_{j=1}^{c}u_{i,j,k}=1\quad i=1,2,...,d\ ;\ j=1,2,...,c\ ;\ k=1,2,...,n\text{ ----------(式 3.3-3)}$$

$$0<\sum_{k=1}^{n}u_{i,j,k}<n\quad i=1,2,...,d\ ;\ j=1,2,...,c\ ;\ k=1,2,...,n\text{ ----------(式 3.3-4)}$$

$m\in[1,\infty)$ 為一指數加權因子。

在此目標函數是每一個輸入樣本的歐氏距離平方和，而每一個歐氏距離的加權值是由模糊隸屬度調整。上述的 $x_{j,k}$ 表示胺基酸序列中的第 j 筆的第 k 個特徵值，$v_{i,k}$ 代表序列中的第 i 個類別的第 k 個特徵，因此兩個相減取距離范數 $\|\ \|$，計算其資料到群心的歐式距離。而演算法的迭代是利用式 3.3-5 及式 3.3-6：

$$v_{i,j}=\frac{1}{\sum_{k=1}^{n}u_{i,j,k}^{m}}\sum_{k=1}^{n}u_{i,j,k}^{m}\cdot x_{i,j,k}\qquad j=1,2,...,c\text{ ----------(式 3.3-5)}$$

$$\frac{\sum_{k=1}^{n}u_{i,j,k}^{m}\cdot x_{i,j,k}}{\sum_{k=1}^{n}u_{i,j,k}^{m}}=\frac{(u_{i,j,1}^{m}\cdot x_{i,j,1})+(u_{i,j,2}^{m}\cdot x_{i,j,2})+(u_{i,j,n}^{m}\cdot x_{i,j,n})}{u_{i,j,1}^{m}+u_{i,j,2}^{m}+\cdots+u_{i,j,n}^{m}}\text{ ---(式 3.3-5.1)}$$

$$u_{i,j,k}=\frac{\left[\frac{1}{\|x_{i,k}-v_{i,j}\|}\right]^{2/(m-1)}}{\sum_{j=1}^{c}\left[\frac{1}{\|x_{i,k}-v_{i,j}\|}\right]^{2/(m-1)}},\ j=1,2,...,c\ ;\quad k=1,2,...,n$$ ----------(式 3.3-6)

$$\frac{\left[\frac{1}{\|x_{i,k}-v_{i,j}\|}\right]^{2/(m-1)}}{\sum_{j=1}^{c}\left[\frac{1}{\|x_{i,k}-v_{i,j}\|}\right]^{2/(m-1)}}=\frac{\left[\frac{1}{\|x_{i,k}-v_{i,j}\|}\right]^{2/(m-1)}}{\left[\frac{1}{\|x_{i,k}-v_{i,1}\|}\right]^{2/(m-1)}+\left[\frac{1}{\|x_{i,k}-v_{i,2}\|}\right]^{2/(m-1)}+\cdots+\left[\frac{1}{\|x_{i,k}-v_{i,c}\|}\right]^{2/(m-1)}}$$ ----(式 3.3-6.1)

為了計算群心，因此所有的輸入樣本均被隸屬度考量為有等值的貢獻，且經迭代之後每一筆的樣本隸屬度，將被修正為樣本值與樣本所在群心的歐氏距離差距。

6.演算法設計與執行

FCM 演算法步驟：

1. 隨機初始化$U^{(0)}$、$V^{(0)}$、ε，設定迭代次數α由 1 開始，設定參數維度d、分類個數c、資料筆數n及指數加權m。

2. 由給定的$U_i^{(\alpha)}$、及(式 3.3-5)計算$V_i^{(\alpha)}$。

3. 由給定的$V_i^{(\alpha)}$、及(式 3.3-6)計算$U_i^{(\alpha)}$。

4. 若$\max|u_{i,j,k}^{(\alpha)}-u_{i,j,k}^{(\alpha-1)}|\le\varepsilon$----------(式 3.3-7)，則終止迭代；否則令$\alpha=\alpha+1$且回到步驟 2.，其中ε是事先設定的誤差容許值。

7.模糊聚類程式模擬部分

一般使用維度遞減的方式進行分類預測時，或多或少仍然有些樣本無法完全予以分類認定，但利用 Fuzzy C-Mean 的優點是將所有樣本參數d均列入分析，發現最後可正確分類出的序列組變多了，但為了確認分類結果是正確的，故利用連續三筆資料的結果均是同類時，才將其歸屬於成功的分類。因本次資料的原生種拖鞋蘭是由農業改良場中獲得(附件一>14 筆原生種)而且只獲得一段，而雜交的品項(附件一>雜交種)也是利用實驗室系統分析出來且也只獲得一段，因此 DNA 序列的起始和結束被固定，故我們在進行實驗時跳過滾動生成的程序，以避免使原本的序列有誤差產生。而進入 FUZZY 的分類過程，再以歐式距離比對所要測試的雜交種，

最後發現本方法可以正確地將 *P. rothschildianumXP. delenatii* 的雜交種子代，順利由 *P.armeniacum*、*P.bellatulum*、*P.chamberlainianum*、*P.concolor*、*P.glaucophyllum*、*P.haynaldianum*、*P.lowii*、*P.micranthum*、*P.purpuratum*、*P.rothschildianum*、*P.sukhakulii*、*P.urbanianum*、*P.victoria-mariae*、*P.villosum* 等十四類的原生種中分辨出來。

拖鞋蘭數值分析時的問題發現

在進行臨床檢驗時發現序列出現了些非 ATCG 字元，經與老師以數學分析的角度及改良場的博士用實驗的誤差討論之後，我們把那些非 ATCG 字元當作逗號，分段進行分類，因為蘭花序列仍有些非 ATCG 字元，必須分段進行分類討論，以供演算法進行數據分析。

在進行臨床檢驗時卻發現，出現了些非 ATCG 字元，再和老師以數學分析的角度及改良場的博士用實驗的誤差討論之後，我們把那些非 ATCG 字元當作逗號，分段進行分類，因為蘭花序列仍有些非 ATCG 字元，必須分段進行分類討論，以供演算法進行數據分析。例如：[…, …, …, …, ATT, NAC, GCA, …, …, …, …]。

其中出現 NAC 一個，又因 N 非 ATCG 字元，所以無法轉變成胺基酸，若是全序列中…個數為 n，而我們跳過 NAC 不考慮，則當全序列越長，其誤差值越小，如式子 $\lim_{n\to\infty}\frac{1}{n-1}=0$，則其他特徵出現頻率為 $\frac{x}{n-1}\fallingdotseq\frac{x}{n}$，而 DNA 序列都具有一定長度，(… 代表一個三字符串、胺基酸)。

肆、討論

模擬 FCM 與 KM 優缺點比較

假設：

九筆二維資料：(2,2)(2,4)(4,2)(4,4)(6,6)(8,8)(8,10)(10,8)(10,10)。

兩筆群心隨機初值：(5,3)(4,7)。

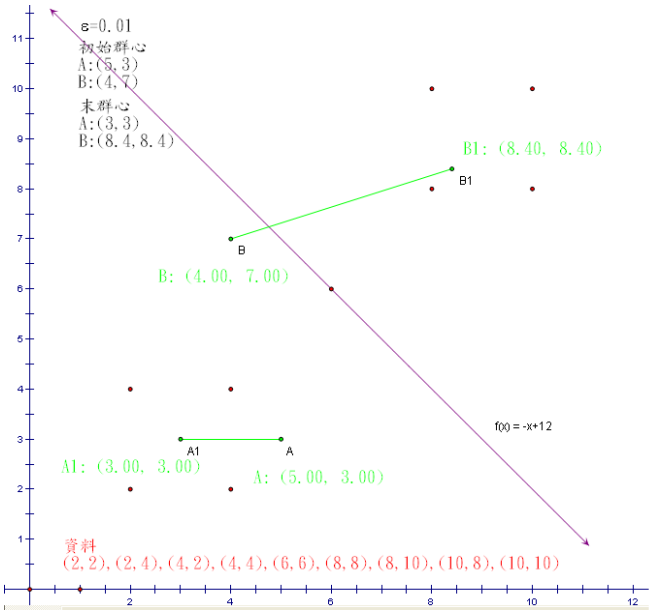


圖 4-1：KM

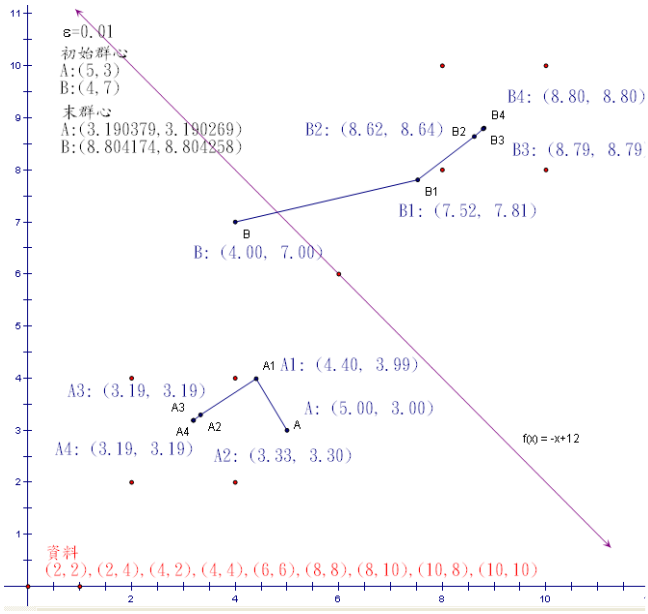


圖 4-2：FCM

結論：比較 KM 及 FCM，發現在 KM 時(圖 4-1)，群心的偏移比較快，但是卻比較不實際；而 FCM 的優點則是比較合理(圖 4-2)，但他的缺點是比較費時費工。

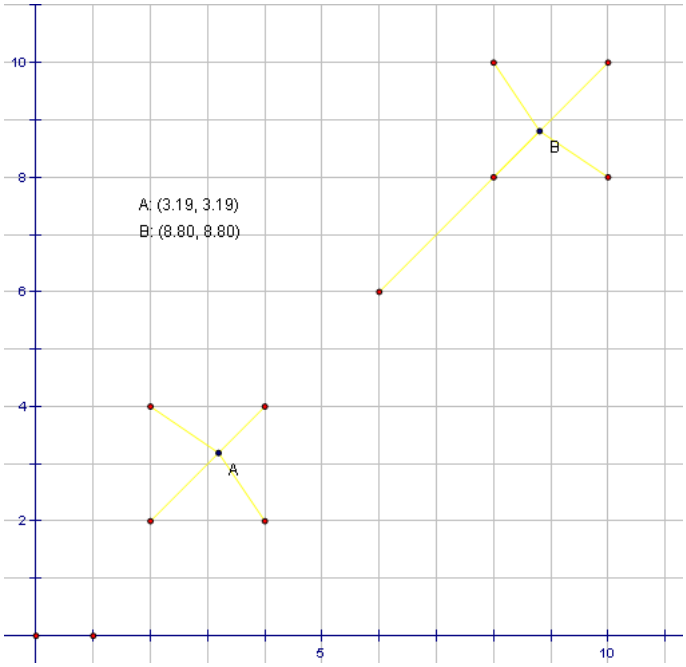


圖 4-3：目標函數

模擬資料是以直線方程式 $f(x)=-x+12$ 對稱，但在 KM 兩群心卻差的很多，FCM 的兩群心就趨近於對稱，這是因為隸屬度差異所形成，這也證明了 FCM 比 KM 更合理有效。最小化目標函數係指當 FCM 演算法反覆迭代，群心偏移達到最小誤差值(式 3.3-7)時，則此時的群心稱作「最佳群心」，最佳群心與所有資料的「距離總和」其值是最小的，如圖 4-3 黃色線部分，各個群心到資料的歐式距離總和為最小值。

模糊聚類程式模擬部分

利用 FCM 的優點是將所有樣本參數 d 均列入分析，因此最後一定會分類出來。且利用***連續三筆資料***的結果均是**同類**時，才將其歸屬於成功的分類，其結果得出 93.22%的正確率，再比對事先設計的人工樣本，發現此次分類的成功。

我們於去年的研究，事先假設 DNA 序列[9]樣本具有：(1)任意一段 DNA 序列樣本是隨機的被歸屬於 A 類或 B 類。(2)先不考量胺基酸排列順序對 DNA 序列所表示的生物訊息。(3)先不考量 DNA 序列中的鹼基三聯組的起始位置所表達的分類結果，但現在有考慮所以省略了滾動計算的步驟。(4)假設原題目所提供的參考樣本內含充分的資訊量[1]。

而本次分類方式，直接以原生種的 *P.armeniacum*、*P.bellatulum*、*P.chamberlainianum*、*P.concolor*、*P.glaucophyllum*、*P.haynaldianum*、*P.lowii*、*P.micranthum*、*P.purpuratum*、*P.rothschildianum*、*P.sukhakulii*、*P.urbanianum*、*P.victoria-mariae*、*P.villosum* 等十四類的原生種拖鞋蘭的 NCBI 資料庫序列各一筆，再結合將 *P.rothschildianumXP.delenatii* 的雜交種子代直接進行人工智慧分類計算結果如下：

To classify DNA sequences：

表 4-1：原生種資料學習

						
P.armeniacum	*P.bellatulum*	*P.chamberlainianum*	*P.concolor*	*P.glaucophyllum*	*P.haynaldianum*	*P.lowii*
						
P.micranthum	*P.purpuratum*	*P.rothschildianum*	*P.sukhakulii*	*P.urbanianum*	*P.victoria-mariae*	*P.villosum*

Classify unsegGroup:

表 4-2：人工智慧方法雜交種識別

						
		P.rothschildianum XP.delenatii				

先將上述十四類原生種資料進行量化數據，再利用人工智慧方法來識別雜交種資料；然後再利用十四類原生種量化資料的群心參數，與雜交種量化資料進行的歐式距離的參數比對，最後直接區分檢查出來的結果(表 4-2)，可見結合人工智慧-模糊理論的方式在 DNA 的分類鑑定中具有良好的效果。

伍、結論

一、本研究對傳統 DNA 分類的改進及應用

傳統 DNA 序列的分類，沒考慮序列的實際特性，當序列變得很多很長很複雜時，分類的準確性可能會降低，因此應增加對 DNA 序列的生物特性的考慮[9]，而本研究經由實驗室所提供的系統數據進行有系統的生物研究，並以拖鞋蘭的雜交實驗證明本方法的臨床價值－分類正確性及有效性，未來將推展至人體醫學及動物等多細胞基因分析，以使此方法更具有實務應用價值。

二、本研究所提新模型的改進方向及推展

本研究成功地提出以模糊分類來嘗試解決實際 DNA 序列樣本，並修正及對實際生物特性存有具體意義，並比較以往只針對蛋白質、胺基酸等作區別分類，本次結果可以說此次的研究有著突破的結果。

三、本研究所提出分析工具的優缺點分析

優點：

1. 提出解決臨床上的生物實驗問題，並成功地建立解決這類難題的數學模型，且可運用到實例中。
2. 解決複雜的分類問題並確定其最後結果的分類正確性，且模型假設條件少，因而能準確地反映實際情況，可靠性高。
3. 採用系統化分析，逐漸深入，提升了準確性，避免了在一些細節問題上的糾纏。
4. 改良了[9]未考慮了 DNA 樣本序列中字串出現的頻率及相對位置作為特徵，並列入基因突變(雜交前後)的考量。
5. 創新 DNA 分類的方式，比較其他生物技術分類方法，只針對蛋白質結構進行區別。

缺點：原本所提的 DNA 序列的分類方法可能與實際情況不一定完全相符，在本例中因為資料來源很明確(來自植物研究中心的實驗室)，而且參考資料少(各種都只有一條)，因此可以不用滾動方式生成，而直接進行 DNA 樣本序列的分析，但是如果遇到資料量大時，例如由 70 種原生種來區分 10 種雜交種時可能要再修正起始模擬參數，以免演算過程掉入數學理論中的相對極小值，而使 DNA 序列的分類可能與實際情況不一定完全相符。

陸、參考資料(References)

[1] 韓中庚、宋明武、邵廣紀，數學建模競賽，北京科學出版社，2007。

[2] 周純芬、彭洪文 編著，生物資訊輕輕鬆鬆學，合記圖書出版社，2005。

[3] 萬絢、閻嘉義，以亂數基礎分類法和 Fuzzy-C-Means 分群法分析土石流判釋問題，水保技術 4(1):37-46，2009。

[4] 遺傳密碼與基因表現-陽明大學普通生物學網路教材。

[5] 模糊理論筆記(Fuzzy Note)，http://irw.ncut.edu.tw/peterju/fuzzy.html。

[6] 林信成、彭啟峰，Oh!Fuzzy 模糊理論剖析 第 3 波出版社，1994。

[7] 楊敏生、楊鎮槐，模糊聚類及其應用，藍海文化，2009。

[8] 中華民國刑事鑑識科學 DNA 小百科，http://www.cib.gov.tw/science/Science0201.aspx?DOC_ID=00007

[9] 楊鎮榮，2010 年台灣區國際科學展覽會電腦科學科，2010。

[10] k-means clustering，http://en.wikipedia.org/wiki/K-means_algorithm，[英文]。

[11] Dunn, J. C.，A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated couster. J. Cybernet, 3,32-57，1974 [英文]。

[12] Bezdek，J. C.，Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum.，1980 [英文]。

[13] Brain Hayes，The Invention of the Genetic Code， American Scientist-Computing Science，Jan.-Feb.，1998。

[14] Zheru Chi、Hong Yan、Tuan Phan，Fuzzy Algoritms With Application to Image Processing and Pattern Recognition.，World Scientific Publish Co. Pte. Ltd，1996。

附件 1：拖鞋蘭 DNA 序列

雜交種>*Delr(P. rothschildianumXP. delenatii)*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGTGCGGTTCGCCGCGCACGACACAGCAAGAAGT

CCATTGAACCttATCATTTAGAGGAAGGcGAAGTCGTAACAAGGTTTCCGTAGGTGAACCtGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTGAATCCAGAGGATCAG

TTTACTTTGGTCGCCCATGGGCgCTTGCTATTGCGGTGACCTAGATTTGCCATGGAGCCTCCTTGGGAGCTTTCTTGCCGGCGATCTAACCCTTGCCCGGCGCAGTTTTGCGCCAAGTC

ACATGACACATAAATGGTGAAGGGCACGGCCCTTTGTGAATTCAAGGAGGTGAAGGGCATGTGGCCTTGAGCCTACACTCCCTCCCCCTCTCAAATTATTTTTTGAACAACTCTCAGCA

ACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGCGATAAGTGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCAAGGCCATCAGG

CCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATATCTCTCCCTTAACGAGGCTGTCCAGGCATACTGTTCAGCCGGTGCGGGTGTGAGTTTGGCCCCTTGTTCTTTGGTGCTGGGGGT

CTAAGAGCTGCAGGGGCTTTTGATGGTCCTAAATTCGGCAAGAGGTGGACGCAACGTGCTACAACAAAACTGTTGTGCGAATGCCCCGGGTTGTCGTATTAGATGGGCCAGCATAATCT

AAACACCCTTGTGAACCCCATTGGAGGCCCATCAACCCATGATCAGTTGATGGCCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

原生種 14 筆

01>*P. armeniacum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGTGCGGTTCGCCGCGCACGACACAGCAAGAAGT

CCATTGAACCTTATCATTCGAGAAGTCGTAACAAGGTTTCCGTAGTAGAACCTAACGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCCAGAGGATCAGTTTACTC

TGGTCACCCATGGGCGCTCGCTTATTTGCGGTGGCCTAGATTTGCCATGGAGCCTCCTTGGGAGCTTTCTTGCCGGCGATCTAACCCTTGCCCGGCGCAGTTTTGCGCCAAGACATATG

ACACATAATCGGTGAAGGGCATAGCCCTTCGTGAATTCAAGGAGGGGGCGGCATGTGGCCTTGACCCTACACTCGCTCCCCCTCTCAAATTATTTTTTGAACAACTCTCAGCAACGGAT

ATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGCGATAAGTGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCAAGGCCATCAGGCCAAGG

GCACGCCTGCCTGGGCATTGCGAGTCATATCTCTCCCTTAATGAGGCTGTCCGTGCATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTGCTAGGGGTCTAAGA

GCTGCAGGGGCTTTTGATGGTCCTAAATTCGGCAAGAGGTGGACGAAATTACMAACAACGCGAATGCTCCAGGTTGTCGTATTAGATGGGCCAAGCACAATCTAAAGACCCTTGTGAAC

CCCACTGGAGGCACATCAACCCGTGATCAGTTGATGGCCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

02>*P. bellatulum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATTGTTGTGATGTGGGCGGTTCGCGCgCACGACACAGCAAGAAGTC

CATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCGCATAATAATTGATCGAGTTAATCTGGAGGATCAGT

TTACTTTGGTCACCCATGGGCATTTGCTGTTGCAGTGACCTAGATTTGCCATCGGGCCTCCTTGGGAGCTTTCTTGCTGGCGATCTATACCCTTGCCCGGCGCAGTTTTGCGCCAAGTC

ATATGACACATAATTGGTGAAGGGGGGGAGGGGGCGTGCTGCCTTGACCCGCTCCAAATTATTTTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCA

GCGAAATGCGATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCGTAT

CTCTCCCTTAATGAGGCTGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATAC

GGCAAGAGGTGGACGAACTATGCTACAACAAAACTGTTGTGCGAATGCCCCGGGTTGTTGTGTTACATGGGCCAGCTAATCAGAAGACCCTTTTGAACCCCATTAGAGGCCCATCAACC

CATGATCAGTTGATTGGCCATTTGGTTGCGATCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

03>*P. chamberlainianum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCGCGCACGACACAGCAAGAAGTC

CATTGAACCTTATCATTTAGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCAGTTTACTTT

GGTCACCCATGGGCATTTGCTGTTGCAGTGACCTAGATTTGCCATCGAGCCGCCTTGGGAGCTTTCTTGTTGGCGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAGTCATATGAC

ACATAATTGGAAGGGGCGGCATGCTGCCTTGGCCCTCCCCAAATTATTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGCGATAAATG

GTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGACATATCTCTCCCTTAATGAGGC

TGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGCAAGAGGTGGACGAA

CTATGCGACAACAGAACTGTTGTGCGAATGCCCCGGGTTGTCGTATTAGATGGGCCAGCATAATCTAAAGACCCTTTTGAACCCCATTGGAGGCCCATCAACCCATGATCAGTTGACGG

CCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

04>*P. concolor*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTTCGCCGCGCACGACACAGCAAGAAGT

CCATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGTGAACCTGCGGAAGGATCATTGTTGAGATCGCATAATAATTGATCGAGTTAATCTGGAGGATCAGT

TTACTTTGGTCACCCACGGGCATTTGCTGTTGCAGTGACCTAGATTTGCCATCGAGCCTCCTTGGGAGCTTTCTTGCTGGCGATCTATACCCTTGCCCGGCGCAGTTTTGCGCCAAGTC

ATATGTCACATAATTGGTAGAAGGGGGGGAGGGGCGTGCTGCCTTGACCCSCTCCAAATTATTTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAG

CGCAAATGCGATAAATGGTGTGAATTGCAGAAtCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATAT

CTCTCCCTTTAATGAGGCTGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCACGGGCTTTTGATGGTCCTAAATA

CGGCAAGAGGTGGACGAATTATGCTACAACAAAACCGTTGTGCGAATGCCCCGGGTTGTTGTGTTACATGGGCCAAGCTTAATCAGAAGACCCTTTTGAACCCCATTGGAGGCCCATCA

ACCCATGATCAGTTGATTGGCCATTTGGTTGCGATCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

05>*P. glaucophyllum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCTCGACGCTCTACCGATTGATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCCGCGCACGACACAGCAAGAAGTCC

ATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATGATAATTGATCGAGTTAATCTGGAGGATCAGTT

TACTTTGGTCACCCATGGGCATTTGCTGTTGCAGTGACCTAGATTTGCCATCGAGCCGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAGTCATATGACACATAATTGGAAGGGGC

GGCATGCTGCCTTGGCCCTCCCCAAATTATTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGCGATAAATGGTGTGAATTGCAGAATC

CCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCCGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGACATATCTCTCCCTTAATGAGGCTGTCCATACATACAGTT

CAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGCAAGAGGTGGACGAACTATGCGACAACAGAAC

TGTTGCGCGAATGCCCCGGGTTGTCGTATTAGATGGGCCAGCATAATCTAAAGACCCTTTTGAATCCCATTGGAGGCCCATCAACCCATGATCAGTTGACGGCCATTTGGTTGCGACCC

CAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

06>*P. haynaldianum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCCGCGCACGACACAGCAAGAAGT

CCATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCAG

TTTACTTTAGTCACCCATGGGCATCTGCTCTTGCAGTGACCTGGATTTGCCATCGAGCCTCCTTGGGAGCTTTCTTGCTGGCGATCTAAATCGTTGCCCGGCGCAGTCTTGCGCCAAGT

CATATCACACATAATTGGAAGGGGGGCGGCATGCTGTCTAGACCCTCCCCAAATTATTTTTTGATAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAAT

GCGATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGCCATCAGGCCAAGGGCACGCCTGCCTGGCATTGCGAGTCATATCTCTCCCTT

AATGAGGCTGTCCATACATACTGTTCAGCCACAGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCCATGGGCTTTTGATGGTCCTAAATACGGCAAGAG

GTGGACGAAGTATGCTACAACAAAACTGTAGTGCGAATGCCCAGGGTTGTCGTATTAGATGGGCCAGCATAATCTAAAGACCCTTTGAACCCCATTAGAGGCCCATCAACCCATGATCA

GTTGATGGCCATTTGGTTGCGACCCCAAGTCAGGTGAGGCAACCCGCTGAGTTTAAG

07>*P. lowii*

TTGACTACGTCCCTGCCCTTTGAACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCAGATCGTTGTGATGTGGGCGGTCCGCGCGCACGACACAGCAAGAAGTC

CATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCAGTT

TACTTTAGTCACCCATGGGCATCTGCTCTTGCAGTGACCTGGATTTGCCATCGAGCCTCCTTGGGAGCTTTCTTGCTGGCGATCTAAATCGTTGCCCGGCGCAGTCTTGCGCCAAGTCA

TATCACACATAATTGGAAGGGGGGCGGCATGCTGTCTAGACCCTCCCCAAATTATTTTTTGATAACTCTCAGCAATGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGC

GATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATATCTCTCCCTT

AATGAGGCTGTCCATACATACTGTTCAGCCACAGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGTTTTTGATGGTCCTAAATACGGCAAGAGG

TGGACGAAGTATGCTACAACAAAATTGTAGTGCGAATGCCCAGGGTTGTCGTATTAGATGGGCCAGCATAATCTAAAGACCCTTTGAACCCCATTAGAGGCCCATCAACCCATGATCAG

TTGATGGCCATTTGGTTGCGACCCCAAGTCAGGTGAGGCAACCCGCTGAGTTTAAG

08>*P. micranthum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGGTGTGTGCGGTTCGCCGCGCACGACACAGCAAGAAGT

CCATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGTGAACCTGCGGAAGGATCATTGTTGAGATCGCATAATAATTGATCGAGTTAATCCAGAGGATCGGT

TTACTTTGGTCACCCTTGGGCGCTTGCTATTGCGGTGACCTAGATTTGCCATGGAGGGAGCCTCCTTGGGAGCTTTCTTGCCGGCGATCTAACCCTTGCCCGGCGCAGTTTTGCGCCAA

GTCATATGACACATAATTGGTGAAGGGCATAGCCCTTCGTGAATTCGAGGAGGGGGCGGCATGTGGCCTTGACCTCTCAAATTATTTTTTGAACAACTCTCAGCAACGGATATCTCGGC

TCTTGCATCGATGAAGAACGCAGCGAAATGCGATAAGTGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCAAGGCCATCAGGCCAAGGGCACGCCT
GCCTGGGCATTGCGAGTCATATCTCTCCCTTAATGAGGCTGTCCATGCATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTGCTGGGGGTCTAAGAGCTGCAGG
GGCTTTTGATGGtCCTAAATTCGGCAAGAGGTGGACGAATCATGCTACAACAAAACTGTTGTGCGAATGCTGTATTAGATGGGCCATCATAATCTAGAGACCCTTGTGAACCCCATTGG
AGGCCCATCAACCCATGATCAGTTGATGGCCTTTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

09>*P. purpuratum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCGCGCACGACACAGCAAGAAGTC
CATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCTGT
TTATTTTGGTCACCCATGGGCATTTGCTGTTGAAGTGACCTAGATTTGCCATCGAGCCTCCTTGGGAGCTTTCTTGTTGGCGAGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAG
TCATATGACACTATAATTGGTGAAGGGGGTGGCATCCTGCCCCGACCCTCCCCAAATTATTTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCG
AAATGCGATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATATCTC
TCCCTTAACGAGGCTGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGCTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGC
AAGAGGTGGACGAATTATGCTACAACAAAACTGTTGTGCCAAGGCCCCGGGTTGTCGTATTAGATGGGCCACCATAATATAAAGACCCTTTTGAACCCCATTGGAGGCCCATCAACCCA
TGATCAGTTGATGGCCATTTGGTTGTGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

10>*P. rothschildianum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCGCGcACGACACAGCAAGAAGTC
CATTGAACCTTATcATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGTGAACCTGCGGaAGGATCATTGTTGAGATCACATAATAATTgATCGAGTTAATCTGGAGGaTCAGTT
TACTTTGGTCACCCATGGGcATCTGCTCTTGCAGTGACCTGGaTTTGCCATCGAGCCTCCTTGGGAGCTTTCTTGCTGGCGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAGTCA
TATGACACAtAATTGGaAGGGGGCGGCATGCTGCCTTGACCCTCCCCAAATTATTTTTTTGACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGC
GATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGCCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATATCTCTCCCTT
AATGAGGCTGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGCAAGAGG
TGGACGAACTATGCTACAACAAAATTGTTGTGTGAATGCCCCGGGTTGTCGTATTAGATGGGCCAGCATAATCTAAAGACCCTTTTGAACCCCATTGGAGGCCCATCAACCCATGATCA
GTTGACGGCCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

11>*P. sukhakulii*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGACGCTCTACCGATTGATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCCGCGCACGACACAGCAAGAAGTCC
ATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCTGTT
TACTTTGGTCACCCATGGGCATTTGCTGTTGAAGTGACCTAGATTTGCCATCGAGCCTCCTTGGGAGATTTCTTGTTGGCGAGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAGT
CATATGACACATAATTGGTGAAGGGGGTGGCATCCTGCCCTGACCCTCCCCAAATTATTTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAA
ATGCGATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATATCTCTC
CCTTAACGAGGCTGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGCAA
GAGGTGGACGAACTATGCTACAACAAAATTGTTGTGCAAAGGCCCCGGGTTGTCGTATTAGATGGGCCACCGTAATCTGAAGACCCTTTTGAACCCCATTGGAGGCCCATCAACCCATG
ATCAGTTGATGGCCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

12>*P. urbanianum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGATGGTCCGGTGAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGCCGCGCACGACACAGCAAGAAGTCC
ATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGGGTTAATCTGGAGGATCTGTT
TACTTTGGTCACCCATGAGCATTTGCTGTTGAAGTGACCTAGATTTGCCATCGAGCCTCCTTGGGAGCTTTCTTGTTGGCGAGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAGT
CGTATGACACATAATTGGTGAAGGGGGTGGCATCCTGCCCTGACCCTCCCCAAATTATTTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAA
ATGCGATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGTCATATCTCTC
CCTTAACGAGGCTGTCCATACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGCAA
GAGGTGGACGAACTATGCTACAACAAAATTGTTGTGCAGAGGCCCCGGGTTGTCGTATTAGATGGGCCACCGTAATCTGAAGACCCTTTTGAACCCCATTGGAGGCCCATCAACCCATG
ATCAGTTGATGGCCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

13>*P. victoria-mariae*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCGCGCACGACACAGCAAGAAGTCCAT
TGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGTTGAGATCACATGATAATTGATCGAGTTAATCTGGAGGATCAGTTTG
CTTTGGTCACCCATGGGCATTTGCTGTTGCAGTGACCTAGATTTGCCATCGAGCCGATCTAAACCCTTGCCCGGCGCAGTTTTGCGCCAAGTCATATGACACATAATTGGAAGGGGCGG
CATGCTGCCTTGGCCCTCCCCAAATTATTTTTTAACAACTCTCAGCAACGGATATCTCGGCTCTTGCATCGATGAAGAACGCAGCGAAATGCGATAAATGGTGTGAATTGCAGAATCCC
GTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGGCATTGCGAGACATATCTCTCCCTTAATGAGGCTGTCCACACATACTGTTCA
GCCGGTGCGGATGGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCATGGGCTTTTGATGGTCCTAAATACGGCAAGAGGTGGACGAACTATGCTGCAACAAAATT
GCTGTGCGAATGCCCCGGGTTGTCGTATTAGATGGGCCAGCATAATCTAAAGACCCTTTTGAACCCCATTGGAGGCCCATCAACCCATGATCAGTTGACGGCCATTTGGTTGCGACCCC
AGGTCAGGTGAGGCAACCCGCTGAGTTTAAG

14>*P. villosum*

TTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGTTGTGATGTGGGCGGTCCGACGCGCACGACACAGCAAGAAGT
CCATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGTGAACCTGCGGAAGGATCATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCAGT
TTACTTTGGTCACCCATGGGGCATTTGCTGTTGAAGTGACCGAGATTTGCCATCGTGCCTCCTTGGGAGATTTCTTGCTGGCGAGATTTAAACCCTTGCCCGGCGCAGTTTTGCGCAAG
ATCATATGACACATAATTGGTGAAGGGGGCGGCATGCCGCCTTGACCCTCCCCAAATTATTTTTTAACAACTCTCAGCAACGGGGATATCTCGGCGTACTTGCATGATGAAGAACGCAG
CGAAATAGCGATAAATGGTGTGAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCTGAGGCCATCAGGCCAAGGGCACGCCTGCCTGGCCATTGCGAGTCATAT
CTCTCCCTTAATGAGGCTGTCCACACATACTGTTCAGCCGGTGCGGATGTGAGTTTGGCCCCTTGTTCTTTGGTACGGGGGGTCTAAGAGCTGCGTGGGCTTTTGATGGTCCTAAATAC
GGCAAGAGGTGGACGAACTATGCTACAACAAAACTGTTGTGCGAATGCCCCGGGTTGTCGTATTAGATGGGCCCAGCATAATCTAAAGACCCTTTTGAACCCCATTGGAGGCCCATCAA
CCCATGATCAGTTGATGGCCATTTGGTTGCGACCCCAGGTCAGGTGAGGCAACCCGCTGAGTTTAAG