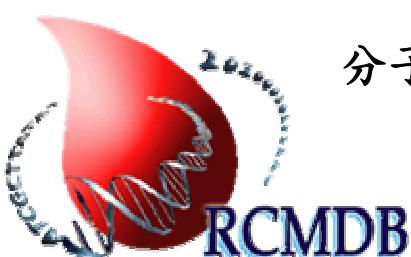


教育部顧問室「生物技術科技教育改進計畫」

生物資訊 I



分子檢驗與生物資訊 教學資源中心 主編

教育部顧問室 補助

中華民國九十二年七月出版

序

在人類基因體計畫完成後，生物技術科技邁入了一個新的紀元，成為二十一世紀最具發展潛力的科技，各科技大國莫不投入大量人力及經費從事教育、研究及產業開發。為因應生物科技突破發展與應用，行政院推動「挑戰 2008—國家重點發展計畫」，由經濟部擬定「兩兆雙星產業發展計畫」，明確勾勒出我國核心與新興產業政策方向。其中生物科技產業就是雙星產業之一，是政府規劃的未來明星產業。為配合政府推動「加強生物技術產業推動方案」，培育生物科技產業人才，教育部顧問室於八十七年七月提出「生物技術科技教育改進計畫」，輔導補助各大學院校提升生物技術教育之內容與層次，全力培育生物科技人才。在第一階段「生物技術科技教育改進計畫」執行四年的成果之基礎上，教育部顧問室從九十年度起推動第二階段的教育改進計畫，期望能引進後基因時代之基因體與蛋白質體，將我國生物技術科技教育與國際接軌。目標是要能提昇我國生物技術科技人才的質與量，並透過產學合作培育出理論與實際兼具的人才，以因應產業發展之需求。為加強第二階段計畫的推展，成立了五個特色產業策略聯盟教學資源中心，分別負責針對相關主題，邀請學界及產業界各領域之專家學者規劃課程並編撰教材。在計畫辦公室及各資源中心主持人努力下，總共完成八本各具特色之教材，這些教材內容涵蓋各項新的技術原理、操作及應用，將提供全國各校作為教學之重要參考。在這關係我國生物技術科技教育發展之教材付梓之時，本人在此感謝各位撰稿的先進賢達辛苦的付出與努力，同時也感謝諸位諮詢委員與顧問們的熱心指導以及教育部顧問室同仁的鼎力協助，使得這些專書能順利出版，特以為序。

第二階段「生物技術科技教育改進計畫」總計畫主持人
國立台灣大學醫學院生化暨分子生物研究所 教授 林榮耀

林 榮 耀

中華民國九十二年三月十二日

教育部顧問室「生物技術科技教育改進計劃」

生物資訊 I

總 編 輯 楊永正

助理編輯 黃素喜、蕭至君

作 者 王聿泰、汪詩海、范廷佳、許玉璇、
陳淑美、黃彥華、楊士德、楊永正、
葉昌偉、蔡健偉、賴俊吉

(依姓名筆劃為序)

分子檢驗與生物資訊教學資源中心 主編

教育部顧問室 補助

中華民國九十二年七月 出版

目 錄

	頁次
目錄.....	I
表格目錄.....	IV
圖目錄.....	V
方盒目錄.....	IX
範例目錄.....	X
練習目錄.....	XII

第一章 緒論

第一節 電腦應用在生物學研究上的重要性.....	1-1
第二節 使用電腦程式的正確態度.....	1-6
第三節 使用電腦程式套組的好處.....	1-7
第四節 生物資訊學的未來.....	1-9

第二章 作業系統與連線方式簡介

第一節 簡介.....	2-1
第二節 連線方式.....	2-1
第三節 UNIX 作業系統的簡介.....	2-3
第四節 文件編輯器 vi 的使用.....	2-11
第五節 顯示圖形與列印.....	2-12
第六節 EMBOSS 使用介面.....	2-14
第七節 結語.....	2-20

第三章 自學 EMBOSS 套組的基本技能

第一節 線上輔助系統使用法.....	3-1
第二節 程式的執行.....	3-6
第三節 資料庫的簡介.....	3-7
第四節 資料庫的搜尋.....	3-7
第五節 結語.....	3-8

第四章 SRS 資料庫查詢

第一節 簡單查詢.....	4-1
第二節 連結其他資料庫.....	4-4
第三節 View 的設定.....	4-5

第四節	SRS 分案查詢.....	4-12
第五節	SRS 資料庫查尋.....	4-19
第五章	資料庫搜尋與多序列並列分析	
第一節	簡介.....	5-1
第二節	FastA.....	5-2
第三節	Blast.....	5-11
第四節	以蛋白質序列搜尋 DNA 資料庫.....	5-16
第五節	多序列排比.....	5-17
第六節	結語.....	5-17
第六章	由應用例中學 EMBOSS 程式的功能	
第一節	基因分析之標準步驟.....	6-1
第二節	選殖基因時可能用的序列分析步驟.....	6-14
第三節	結語.....	6-14
第七章	問題導向學習(I)：轉錄因子 TFIIIA 的選殖	
第一節	序列的輸入.....	7-1
第二節	反轉譯.....	7-1
第三節	開放讀架的搜尋.....	7-3
第四節	序列資料庫的搜尋.....	7-8
第五節	結語.....	7-9
第八章	問題導向學習(II): 轉錄因子 TFIIIA 的性質分析	
第一節	模組資料庫的搜尋.....	8-1
第二節	蛋白質性質分析.....	8-3
第三節	多序列排比.....	8-7
第四節	結語.....	8-10
第九章	問題導向學習(III)：鋅指模組的發現與分析	
第一節	重覆序列的尋找.....	9-1
第二節	以模組搜尋序列資料庫.....	9-7
第三節	立體結構的預測.....	9-10
第四節	結語.....	9-11
第十章	問題導向學習(IV)：TFIIIA 基因體序列的分析	
第一節	尋找 mRNA 的起點.....	10-1
第二節	基因上游序列的分析.....	10-4

第三節	搜尋序列中字串的快速方法.....	10-5
第四節	預測 intron-exon 之交界點.....	10-6
第五節	序列排比的原理.....	10-10
第六節	結語.....	10-13
第十一章 問題導向學習(V)：5S rRNA 的二級結構分析		
第一節	由 RNA 序列預測 RNA 的二級結構.....	11-1
第二節	利用親緣分析預測 5S RNA 的結構.....	11-16
第三節	預測 RNA 結構的問題.....	11-19
第四節	結語.....	11-21
第十二章 序列資訊之應用		
第一節	簡介.....	12-1
第二節	EST.....	12-2
第三節	應用 EST 資訊的實例.....	12-3
第四節	結語.....	12-13
附錄 1	PuTTY, X Window 軟體的取得與安裝.....	A1-1
附錄 2	習題組：定位訊號 RNA 接合蛋白的選殖與分析.....	A2-1
附錄 3	習題參考解答.....	A3-1
索引.....		I-1

表格目錄

	頁次
表 2-1 與管理檔案有關的指令整理.....	2-8
表 2-2 與執行背景工作相關的指令整理.....	2-10
表 2-3 與程式執行有關的指令整理.....	2-10
表 5-1 參數 ktup 的設定對 FastA 輸出的影響.....	5-9
表 5-2 BLAST 典型輸出結果形式.....	5-12
表 7-1 Blast 之不同子程式的運作原理.....	7-9
表 8-1 自 tfiiia.p2s 中所得可能具有 α -螺旋的區域.....	8-5
表 10-1 序列 A 的索引對照表.....	10-6
表 10-2 TFIIIA 基因體序列中 exon、intron 之範圍.....	10-8
表 10-3 改變參數對用 needle 預測 Exon-Intron 連接處的影響.....	10-10
表 11-1 25 °C下，RNA 的鹼基堆疊自由能.....	11-3
表 11-2 <i>Xenopus laevis</i> 卵母細胞 5S rRNA 的切割酵素位置.....	11-15
表 12-1 各種不同來源的 eIF4G 相關序列的相似性比較.....	12-4

圖 目 錄

	頁次
圖 1-1 基因體分析計畫將造成生物學的下一次突破.....	1-2
圖 1-2 表示 DNA 序列的方式，對解讀 DNA 語言的影響.....	1-4
圖 2-1 使用 PuTTY 程式連接陽明生物資訊中心安裝 EMBOSS 的工作站.....	2-2
圖 2-2 使用 telnet 連接到陽明生資中心工作站，使用 EMBOSS 的畫面.....	2-2
圖 2-3 Xstart 操作介面.....	2-12
圖 2-4 在 Windows 可以利用 Internet Explorer 作為圖形瀏覽器.....	2-14
圖 2-5 Jemboss 網頁，此處可下載 Java Web Start 並且可 launch Jemboss.....	2-15
圖 2-6 Jemboss 簽入畫面，輸入使用者帳號和密碼.....	2-15
圖 2-7 出現連結錯誤.....	2-15
圖 2-8 Public Server Setting 視窗.....	2-16
圖 2-9 將 Proxy 表單中的打勾處移除，方能使用 Jemboss.....	2-16
圖 2-10 Jemboss 使用者介面.....	2-17
圖 2-11 Jemboss 使用者介面簡介.....	2-18
圖 2-12 工作管理員視窗.....	2-19
圖 2-13 檢視分析結果圖形.....	2-19
圖 2-14 檔案管理員視窗.....	2-20
圖 3-1 「命令列參數」顯示「-cfile Codon_usage_table_name」使用密碼使用頻率表的寫法...	3-5
圖 4-1 The Start page.....	4-1
圖 4-2 The Top page.....	4-2
圖 4-3 選擇 embl 資料庫.....	4-2
圖 4-4 embl 的標準查詢表格.....	4-3
圖 4-5 查詢特定欄位的資料.....	4-3
圖 4-6 查詢到的結果.....	4-3
圖 4-7 LINK page.....	4-4
圖 4-8 選擇連結的資料庫.....	4-4
圖 4-9 顯示查詢的結果.....	4-5
圖 4-10 新增一個自定表單.....	4-5
圖 4-11 勾選資料庫的資料欄.....	4-6
圖 4-12 利用自定選單觀看查詢結果.....	4-7
圖 4-13 在自定表單呈現的資料.....	4-8
圖 4-14 勾選一個分析工具.....	4-8
圖 4-15 可使用預設值送交執行.....	4-9
圖 4-16 執行中的工作狀態.....	4-9
圖 4-17 執行結果.....	4-10
圖 4-18 存檔格式設定.....	4-10

圖 4-19 存檔格式設定.....	4-11
圖 4-20 下載視窗.....	4-11
圖 4-21 開始頁面.....	4-13
圖 4-22 資料庫選擇頁.....	4-14
圖 4-23 暫時性專案管理.....	4-14
圖 4-24 SRS 安全型登入視窗.....	4-15
圖 4-25 SRS 非安全型登入視窗.....	4-15
圖 4-26 永久性專案管理頁.....	4-16
圖 4-27 個人化你的專案名稱.....	4-17
圖 4-28 檔案下載視窗.....	4-18
圖 4-29 儲存對話視窗.....	4-18
圖 4-30 瀏覽專案檔案.....	4-19
圖 4-31 開始頁面.....	4-21
圖 4-32 選擇資料庫.....	4-21
圖 4-33 填入關鍵字.....	4-22
圖 4-34 勾選查詢資料庫.....	4-22
圖 4-35 標準查詢表格.....	4-23
圖 4-36 組合查詢方式.....	4-24
圖 4-37 進階查詢模式.....	4-25
圖 4-38 進階查詢模式.....	4-26
圖 4-39 Expression 按鍵與文字方塊.....	4-26
圖 4-40 輸入 Q1&Q2.....	4-27
圖 4-41 Q3 所有資料為輸入值，再到 swissprot 查詢.....	4-27
圖 4-42 欄位資料頁中可以提供 “Description” 欄位的查詢.....	4-27
圖 4-43 瀏覽 Description 資料欄的索引.....	4-28
圖 4-44 進階查詢模式中點選 “Info” 鍵將可進入下拉式選單中所顯示的欄位.....	4-28
圖 4-45 在進階查詢中資料欄都是以超連結方式呈現.....	4-29
圖 4-46 在資料庫資訊頁中，資料欄將以超連結方式呈現.....	4-29
圖 5-1 FastA 的柱狀圖.....	5-2
圖 5-2 不同 ktup 值時，FastA 程式所找到的相似序列清單.....	5-5
圖 5-3 序列排比和區域性序列排比之輸出完全相同.....	5-7
圖 5-4 不同 ktup 值時，第一階段所找到的片段長度不同.....	5-10
圖 5-5 Blast 程式所列出的序列排比結果.....	5-14
圖 5-6 使用 Blast 程式中的過濾器除去低複雜性的或重覆性的序列的意義.....	5-16
圖 6-1 以 TFIIIA 為例，顯示 plotorf 程式的輸出結果，第二個讀架上的 ORF 才是真正表現的蛋白質.....	6-2
圖 6-2 廣域與區域性序列排比的比較.....	6-3
圖 6-3 程式 showalign 的蛋白質序列排比輸出格式，突顯守舊的區域.....	6-4

圖 6-4 程式 showalign 的蛋白質序列排比輸出格式，突顯有差異的區域.....	6-4
圖 6-5 使用 Plotcon 程式，以圖形顯示多序列排比中相似的區域.....	6-5
圖 6-6 Profile 分析的步驟.....	6-6
圖 6-7 emma 程式所產生的圖不是親緣分析樹.....	6-7
圖 6-8 以 TFIIIA 為例，顯示 pepinfo 程式的輸出結果.....	6-10
圖 6-9 以 TFIIIA 為例，顯示 pepinfo 程式的輸出結果.....	6-11
圖 6-10 檢視 TFIIIA 蛋白質第四號鋅指中的 α -螺旋是否具有兩性的特性.....	6-12
圖 6-11 蛋白質 TFIIIA 的 hmoment 分析.....	6-12
圖 6-12 基因分析標準步驟間的關係.....	6-13
圖 7-1 pep1.pro 序列的反轉譯.....	7-2
圖 7-2 backtranseq 的輸出結果.....	7-2
圖 7-3 TFIIIA 在核酸序列上的開放讀架的結果(在個人電腦中觀察的結果).....	7-4
圖 7-4 利用 getorf 程式尋找開放讀架的精確位置.....	7-6
圖 7-5 轉譯出的 TFIIIA 蛋白質序列，粗體字部份是當初用來選殖 TFIIIA 基因的兩段 peptide.....	7-7
圖 7-6 Blast 的使用.....	7-8
圖 8-1 TFIIIA 序列中所含的蛋白質模組.....	8-1
圖 8-2 Garnier 的輸出檔案格式.....	8-4
圖 8-3 msf 檔案格式.....	8-9
圖 9-1 以「點矩陣」法尋找重覆序列.....	9-1
圖 9-2 利用移動窗使曲線變平滑的示意圖.....	9-3
圖 9-3 不同 Window size 與 Threshold 對點矩陣表示法的效應.....	9-4
圖 9-4 對 TFIIIA 內重覆序列做序列排比所需的檔名檔案.....	9-4
圖 9-5 TFIIIA 內各重覆序列的序列排比結果.....	9-6
圖 9-6 Mse 後的鋅指序列排比.....	9-6
圖 9-7 鋅指與 DNA 交互作用的可能方式.....	9-7
圖 9-8 MRC 研究群提出的鋅指結構.....	9-7
圖 9-9 利用 C ₂ H ₂ 為模組樣式所搜尋到的節錄結果.....	9-9
圖 9-10 Berg 所提出的鋅指模型.....	9-11
圖 9-11 各資料庫串接示意圖.....	9-12
圖 10-1 用 water 找尋 TFIIIA 的 cDNA 序列和基因體序列中共有之序列.....	10-2
圖 10-2 用 needle 找不到 TFIIIA 的 cDNA 序列和基因體序列中共有之序列.....	10-2
圖 10-3 以 Tfscan 進行 TFIIIA 的上游基因序列分析的結果.....	10-4
圖 10-4 序列 A 的兩種表示方式，(1)以鹼基形式表示；(2)以雙鹼基(dinucleotide)形式表示.....	10-6
圖 10-5 利用 water 程式找尋 exon-intron 交界.....	10-7
圖 10-6 利用 water 以找尋 exon-intron 交界(將 Gap opening penalty 改為 100·Gap extension penalty 變為 0.1).....	10-9
圖 10-7 序列 A 中不完美重複序列之序列排比.....	10-11

圖 10-8 動態程式設計的計分方式.....	10-12
圖 10-9 每個格子的得分都是相鄰的三個格子的最高得分與該格子的分數之和.....	10-12
圖 11-1 不同結構對鹼基堆疊自由能的影響.....	11-2
圖 11-2 在不同溫度預測 <i>Xenopus 5S rRNA</i> 二級結構之結果.....	11-5
圖 11-3 以 5S rRNA 的結構為例，說明 RNA 結構可用不同的方法表現.....	11-6
圖 11-4 分析 MFold 結果所用的 energy dotplot (A) 與 P-Num plot (B).....	11-11
圖 11-5 以 MFold 預測 5S rRNA 之最穩定二級結構的折曲圖.....	11-12
圖 11-6 生物體內的 RNA 不一定是最穩定的結構.....	11-13
圖 11-7 雙股區域(Stem)上 i, j, k 的定義.....	11-14
圖 11-8 RNA 二級結構的微調是一個遞迴的過程.....	11-16
圖 11-9 產生協同變異的可能機制.....	11-17
圖 11-10 5S rRNA 的序列並列分析之結果，由此可看出 C18 與 G60, C30 與 G47 有協同變異 ...	11-18
圖 11-11 守舊的核苷酸(○、□)多位於 RNA 的單股區域.....	11-21
圖 12-1 根據序列的相似性圖示 EST W31907 及 W39270，以及 se16 的序列分別對應至小鼠的 eIF4G2 相關基因各個不同的區域，各項位置是以小鼠的 eIF4G2 相關基因由 5' 端至 3' 端的位置為準.....	12-6
圖 12-2 用 GCG 中的 GelMerge 程式來重組 96 個與小鼠 eIF4G2 相關基因間具高相似性的 EST，圖示各接合序列和小鼠 eIF4G2 相關基因之間的序列相似的位置.....	12-7
圖 12-3 第二次以 W39270 為關鍵字搜尋 UniGene 資料庫的結果.....	12-8
圖 12-4 第三次以 W39270 為關鍵字搜尋 UniGene 資料庫的結果(約在圖 12-3 的三星期之後)...	12-8
圖 12-5 將 296 個由 Blast 程式所搜尋到的 EST 序列，以 GelMerge 進行重組及分類的結果，圖示表示各接合序列和 p97 基因全序列的相關位置.....	12-9
圖 12-6 假設有一基因具有三個基因剪接點: J、J2、以及 J3，將 genomic sequence 分成 S1、S2、S3 及 S4 四個區段。由於選擇不同的基因剪接點，因而造成二種不同的基因剪接型式: 1st 和 2nd.....	12-10
圖 12-7 以 FastA 分析各接合序列可以和 p97 基因有良好並列的區域，縱軸表示在接合序列上的位置，橫軸表示各個不同的接合序列；方框區域表示各接合序列之中可以和 p97 有良好並列的區域，直線區域則是表示各接合序列之中不能和 p97 直接並列的區域.....	12-11
圖 12-8 由接合序列 aa096164 與 p97 基因的序列排比，所預測出的基因剪接點的位置。圖示 aa096164 與 p97 基因的序列相關位置，並列出在基因剪接點處的序列，以底線標示出守舊的區域.....	12-11
圖 12-9 預測基因剪接點的原理.....	12-12
圖 A1-1 PuTTY 的執行介面.....	A1-1
圖 A1-2 Xstart 視窗介面.....	A1-2
圖 A2-1 定位訊號 RNA 的決定.....	A2-1
圖 A2-1 pGEMT7Z(+)的圖譜.....	A2-3

方盒目錄

	頁次
方盒 2-1 vi 編輯器的按鍵使用法.....	2-11
方盒 3-1 wossname 的使用法.....	3-1
方盒 3-2 密碼使用頻率表.....	3-4
方盒 8-1 在 GCG/EMBOSS 環境下，寫模組樣式的規則.....	8-3
方盒 8-2 檔名檔案(list file)的產生與應用.....	8-8
方盒 9-1 序列重覆出現的意義.....	9-2
方盒 11-1 由序列預測能量最低二級結構之原理.....	11-1
方盒 11-2 加入限制條件的規則.....	11-13
方盒 11-3 以親緣關係預測二級結構的原理.....	11-16
方盒 12-1 生物學資料庫與 UniGene.....	12-5

範例目錄

	頁次
範例 2-1 請利用指定的教學帳號與密碼簽入工作站，然後更改密碼為工作站的號碼.....	2-3
範例 2-2 Windows 系統下 FTP 程式的使用.....	2-4
範例 2-3 查閱子目錄下有哪些檔案，並在螢幕上檢視「1.txt」.....	2-6
範例 2-4 請在更改使用權限前先顯示權限，再將 run.txt 改為可以執行之狀態，然後再顯示權限，確定更改成功.....	2-7
範例 2-5 試以「重新導向」的方式將「ls」指令的使用方法送到檔案「ls.doc」中.....	2-7
範例 2-6 假設你在一個 X-視窗中工作，請在背景中開兩個新的 X-視窗，並觀察其工作狀態.....	2-9
範例 2-7 接續範例 2-5，執行「run.txt」，再將其中斷，送到背景中執行，最後再列出在背景中執行的工作，並將此工作取消.....	2-9
範例 2-8 請使用 vi 建立一個檔名叫「pep1」的蛋白質序列檔，其序列為「FHNIKI」.....	2-11
範例 2-9 直接利用 X Window 顯示圖形.....	2-13
範例 2-10 直接產生圖形檔案.....	2-13
範例 3-1 使用 tpm 查詢 tpm 本身的程式使用手冊.....	3-2
範例 3-2 利用 wossname 查閱和「轉譯」相關的程式.....	3-3
範例 3-3 利用 tpm 查閱「反轉譯」程式的使用.....	3-3
範例 3-4 backtranseq 程式的使用範例.....	3-3
範例 3-5 使用 backTranseq 程式的「說明」，以瞭解使用程式時的注意事項.....	3-6
範例 3-6 USA(Uniform Sequence Addresses)表示法.....	3-7
範例 3-7 請取回 <i>Xenopus laevis</i> 的 TFIIIA 蛋白質序列，並將其命名為 tf3a.pro.....	3-8
範例 7-1 以 embossdata –showall 顯示合適的密碼使用表.....	7-1
範例 7-2 請找出 TFIIIA 在核酸序列上的開放讀架.....	7-3
範例 7-3 請找出 TFIIIA 核酸序列中最可能的開放讀架之起始密碼(AUG)與結束密碼的位置.....	7-5
範例 8-1 請檢查轉譯出的 TFIIIA 序列上，有哪些 ProSite 資料庫中的已知模組.....	8-1
範例 8-2 請用 Garnier 程式預測 TFIIIA 蛋白質的二級結構.....	8-3
範例 8-3 試練習使用 Hmoment 程式繪圖.....	8-6
範例 8-4 請利用 Emma 程式，將所有與 TFIIIA 有關的 DNA 序列做多序列排比.....	8-7
範例 8-5 請利用 Showalign 程式，將 TFIIIA 的蛋白質多序列分析結果表示成第 6-4 頁，圖 6-4 之形式.....	8-10
範例 9-1 利用點矩陣法來尋找 TFIIIA 中的重覆序列.....	9-2
範例 9-2 請將 Berg 所寫的模組樣式表示為 EMBOSS 的格式.....	9-8
範例 9-3 以 Berg 的模組樣式搜尋 Swiss-Prot 序列資料庫.....	9-9
範例 10-1 請利用已知的序列資訊找出 TFIIIA mRNA 的起始位置.....	10-1
範例 10-2 請估計 TFIIIA 基因上游序列中的 TATA 盒相對於 mRNA 起點的位置.....	10-4
範例 11-1 試以 Foldrna 程式預測 <i>Xenopus 5S rRNA</i> 的結構.....	11-4
範例 11-2 請繪出上述範例中所預測的結構之折曲圖(結果參見圖 11-2A).....	11-4

範例 11-3	請利用 PlotFold 繪出「xen5s.mfold」的 energy dotplot.....	11-9
範例 11-4	請利用表 11-2 中數據微調 25°C 下所預測出的二級結構.....	11-14
範例 11-5	自資料庫中取得 5S rRNA 序列.....	11-17

練習目錄

頁次

練習 2-1	請每人在共同帳號下，以自己的姓，建立一個屬於自己的子目錄，然後進入此子目錄，找出此子目錄在系統中的路徑.....	2-4
練習 2-2	請使用 anonymous FTP 的方式到「ymbc.ym.edu.tw」上取回本書所需之檔案.....	2-5
練習 2-3	將檔案 「1.txt」 更名為 「hum-tf3a.dna」，再將檔案 「2.txt」 拷貝成為 「yst-tf3a.dna」。此時目錄下將有兩個檔案具有同樣的內容，因此請刪除「2.txt」.....	2-6
練習 2-4	試將所有副檔名為「dna」的檔案資訊存入一個叫「dna.info」的檔案中.....	2-7
練習 3-1	請以 tfm 查詢 wossname 程式的使用手冊.....	3-2
練習 3-2	請查閱 backtranseq 中「數據檔案(Data Files)」這個副標題，以瞭解如何指定反轉譯所需的數據檔案.....	3-5
練習 3-3	如何利用 showdb 程式，自 EMBOSS 套組中搜尋已安裝的資料庫.....	3-7
練習 7-1	請重覆上述步驟(圖 7-2 與圖 7-3)，並從圖中估計最可能的開放讀架的開始和結束位置.....	7-5
練習 7-2	試將 TFIIIA 核酸序列中最長的開放讀架轉譯為蛋白質序列，並檢查 pep1 與 pep2 是否存在於轉譯出的蛋白質序列中.....	7-6
練習 8-1	請用 Antigenic 程式，將抗原性(antigenic index)標幟於其上(修改第 6-10 頁，圖 6-9 所示).....	8-4
練習 8-2	試利用 Pepwheel 程式檢視 TFIIIA 中第 137 到 155 個胺基酸的疏水性官能團分佈特性.....	8-6
練習 8-3	請利用 Emma 程式，將所有 Swiss-Prot 中與 TFIIIA 有關的蛋白質序列做多序列排比，並比較所用的 penalty 與 DNA 多序列排比之差別(第 8-7 頁，範例 8-4).....	8-9
練習 8-4	請將 TFIIIA 蛋白質的多序列分析結果表示成第 6-4 頁，圖 6-3 之形式.....	8-10
練習 9-1	請利用多序列排比顯示 TFIIIA 中各重覆序列的相似性.....	9-5
練習 9-2	試根據圖 9-5，將 TFIIIA 中各重覆序列的共有序列用符號表示出來.....	9-8
練習 9-3	試在指令行下條件，以 TFIIIA 的鋅指模組樣式搜尋 Swiss-Prot 資料庫.....	9-10
練習 10-1	Korn 的實驗室選殖到整個 TFIIIA 的基因體序列，可是只有決定了幾段序列，請利用其中的一段(accession # X03736)與 cDNA 序列比對，找出 intron-exon 邊緣的位置.....	10-7
練習 11-1	請找出一個方法，預測 <i>Xenopus</i> 5S rRNA 在室溫的結構，並繪出其折曲圖.....	11-8
練習 11-2	請利用 MFold 程式預測 <i>Xenopus</i> 5S rRNA 在室溫中之二級結構.....	11-8
練習 11-3	請利用(68,107)與(69,106)兩對協同變異為限制條件，預測出 5S rRNA 之二級結構	11-15
練習 11-4	編輯「xen5s.fas」，取得分最高的 29 個序列做多序列排比.....	11-18

第一章 緒論

楊永正

陽明大學生物資訊研究所

一. 電腦應用在生物學研究上的重要性

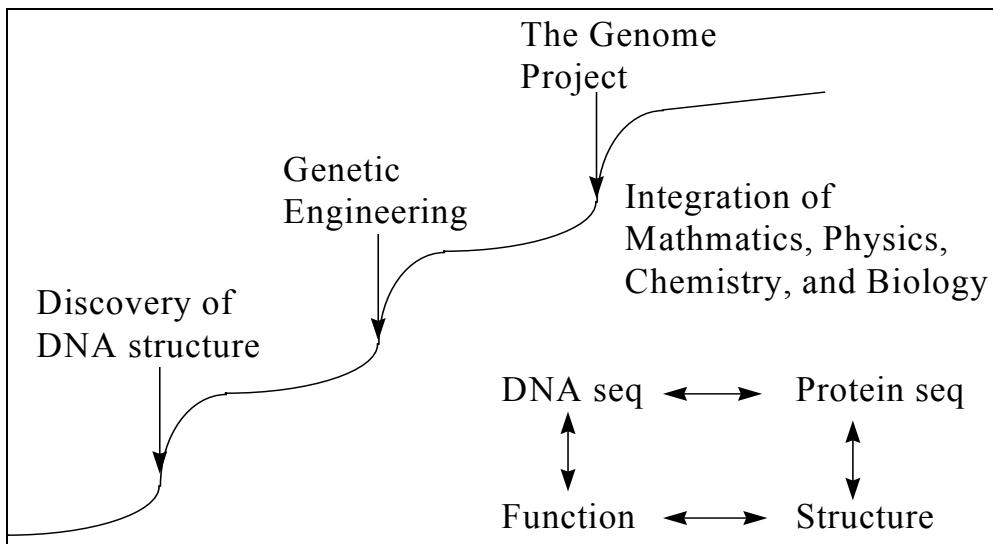
1988 年開始的基因序列分析計畫(genome project)，非但對生物學家研究的方式產生影響，也是電腦在生物學上應用的一個重要轉折點。因為有計畫的序列分析將產生大量的資訊，大到沒有電腦就無法處理這些資訊。在基因序列分析計畫中也支持生物資訊學(bioinformatics)的發展，因為研究所得的圖譜、序列資訊若不整理就無法有效的利用。能預見的是誰能有效利用這些圖譜、序列資訊，誰就能掌握先機。雖然目前大部份的資源都用於決定序列上，未來最大的挑戰將是如何了解這些序列中所存的資訊，也就是所謂的「DNA 語言」。

由上述的討論中可看出目前使用電腦工具的兩大方向，首先是如何得到資訊，其次是是如何分析資訊。早在 1945 年，Bush 在一篇名為「As we may think」的文章中⁽¹⁾就提出「梅米克斯(memex)」的觀念，其精義在於利用電腦工具替人迅速找尋所需資訊，以利思路的連續，而且找到的相關資訊更能使人觸類旁通。因為網際網路(internet)的通暢，使用者能非常方便地連接到遠處的資料庫，查閱所需的資訊。可是資料庫並未完全整合，所以使用者仍無法隨心所欲地遨遊於資訊之海，這離「梅米克斯」這理想還有一段距離，也是研究生物資訊學的人所需努力的方向。在另一方面，不斷有人寫一些小的電腦程式來協助我們處理一些繁瑣的工作，例如將不同 DNA 序列分析凝膠上的結果依重疊狀況組合成比較長的 DNA 序列(稱為 contig)、由決定的 DNA 序列中找尋限制酵素切割位置等。另一些人則希望利用電腦做預測的工作，所以他們發展出工具來比較親緣遠近、來預測基因的位置、甚至預測 RNA 或蛋白質的結構。雖然預測工具還是不理想，隨著資料庫的增大和計算能力的加強，預測的方法將會不斷改善，預測能力也將隨之趨於可以接受的程度。生物資訊學的主要的目的將是破解這「DNA 語言」，使我們了解生物體中的功能訊息是如何紀錄在 DNA 上的。

1. 基因體分析計畫將造成生物學的下一次突破

生物學在本世紀曾有兩次突破性的進展，一是發現 DNA 的結構，使我們有一個具體的模型去瞭解遺傳的原理。這個發現使分子生物學蓬勃的發展，所以對複製現象有深入的瞭解，進而發展出遺傳工程的技術，這技術是生物學的第二次突破。因為它使我們有改變活細胞性質的能力，也使某一段特定的 DNA 片段能被增殖出來做探針或大量表現為蛋白質，因此被廣泛地應用在各不同的生物學門。可預期在基因體分析計畫完成後生物學將再有一次突破性的進展(圖 1-1)，這是因為目前生物學發展的瓶頸在於知道 DNA 的序列後，卻不保證能推論出其蛋白質產物的功能；而不知道蛋白質的功能，就不知道它的作用方式，也無法了解生命現象。根據遺傳的原理，我們知道生命的奧秘盡藏於 DNA 序列中，所以如何解讀 DNA 序列中的資訊就成為目前生物學上最重要的問題。

圖 1-1. 基因體分析計畫將造成生物學的下一次突破



既然現在的問題是不知如何解讀 DNA 序列中的資訊，為何決定更多的序列就能協助我們去解釋 DNA 序列中的資訊呢？以汽車為例，當我們要了解車子各部份的功能時，固然可由分解一部汽車著手，亦可由了解設計藍圖著手。在有設計藍圖的輔助下，了解汽車原理的速度遠快於分解一部汽車的策略。這是因為藍圖所代表的是完整的資訊，在看不懂藍圖時，甚至可由比較不同車子藍圖著手來了解藍圖。可是在藍圖破損時，要想從藍圖了解車子的結構與功能就非常困難了。同樣的，在有完整資訊的狀況下，比較容易去了解 DNA 序列的意義；在沒有完整的序列資訊的情形下，則不易從序列推測蛋白質的功能。

基因體研究計畫(genome project)已在 2003 年完成，研究生物學的方法卻早已發生重要的改變。這種改變事實上在一些病毒的研究上早已發生，只是病毒無法自己獨立生存，所以只知道病毒的全部序列仍無法瞭解其生活之全貌。以對分子生物學發展有重要影響的 adenovirus 為例，早在取得其全序列之前，分子生物學家就利用核酸雜交的技術，找到在感染細胞中所表現的病毒基因。並發現 mRNA 剪接現象這一個重要的現象。這些工作顯示，即使沒有完整的 DNA 序列資訊，也能發現與解釋許多重要的生物現象。可是在沒有完整序列資訊之前，研究進展的速率遠低於知道序列後的速率。因為要找到具有明顯表現型的突變株不容易，若有序列資訊則可根據我們對蛋白質性質的認識，去設計突變株做測試。此外，在知道 DNA 全序列後，所有開放讀架的位置都很清楚，可用實驗確定哪一些開放讀架可真正轉譯出蛋白質。分子生物學家可用電腦尋找適當的限制酵素切割位置，以便做分子選殖或設計探針。於是利用核酸探針與對特定蛋白質產物產生的抗體，可輕易找出基因或蛋白質表現的先後次序。此外，也可根據序列設計突變，然後再利用探針與抗體來觀察突變所造成的影响，而不必單靠表現型的變化。利用這樣的方法可以找出各個不同基因的功能，這些都是在有序列之前很難做到的。

2. 解讀 DNA 「語言」的策略

DNA 語言之所以難懂是因為我們表示 DNA 序列的方式，並不是一個自然的表示法。如圖 1-2 所示，若將 Sidney Brenner 的一句名言改用表示 DNA 序列的方式來寫，就很難一眼看出這句話的意義。序列的自然表示法是以基因为單位，因為它像一個句子那樣表達一個完整的概念。可是在基因被調控時，或是表現成蛋白質而執行功能時，可能由一些獨立運作的小單元，例如 TATA box 或 leucine zipper 等，分別執行不同的任務。這些獨立單元就像字一樣，會重覆出現在不同的句子中。字在不同的句子中，有不同的組合方式，這就靠文法來規範。文法就像整合各種共有(consensus)序列或蛋白質模組(motif)成為基因的方式，文法也像是核酸序列與蛋白質之間交互作用的規則。想要瞭解一個未知的語言，就要由瞭解字的意義與文法的使用著手。

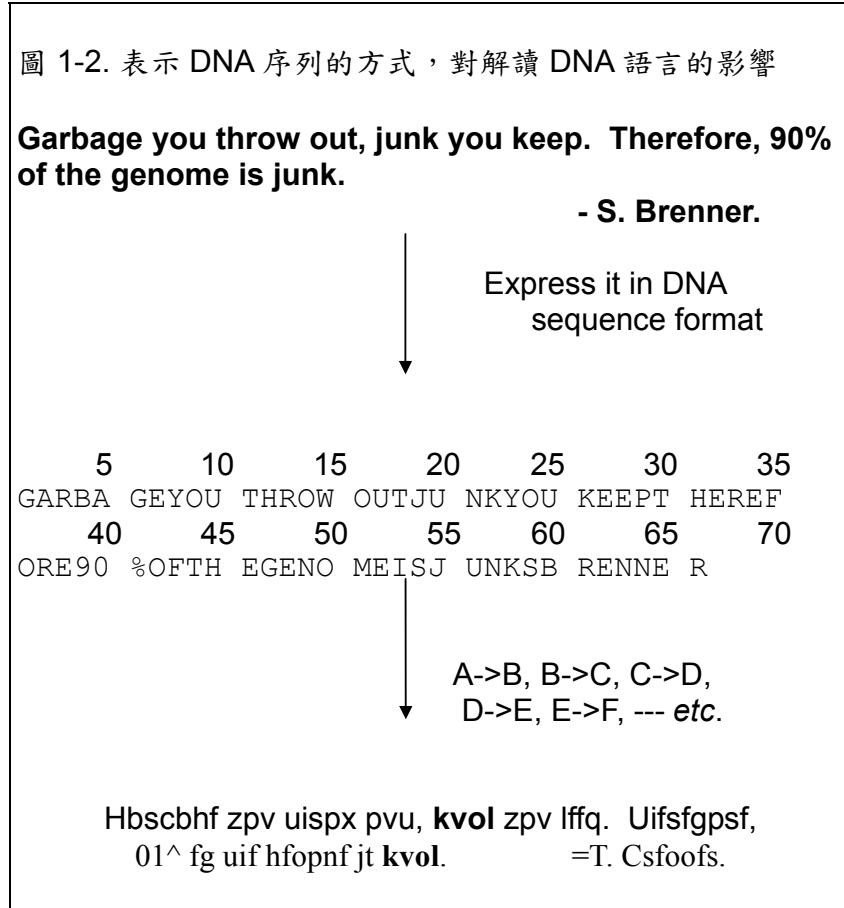
要了解如何破解一未知的語言，可以由瞭解破解密碼的方法入手。假設我將 Brenner 的句子中的每一個字母都轉化成為另一個字母，例如，將所有的 A 都變成 B, B 都變成 C, (圖 1-2)。任何不知道轉化規則的人，將無法了解這句話的意義。可是字母、字、詞的特性是它們會重覆的出現在文章中，因此只要文章夠長，就可以利用統計的工具來辨識什麼是字，例如在轉化過的句子中，kvol 出現兩次，它剛巧代表 junk 這字。戰時所用的密碼比這個例子複雜許多，可是破解的原理卻很相似，事實上由敵軍的反應中，可猜測字的意義。在二次世界大戰時，美軍就因為破解了日本的密碼，而在中途島重創日本海軍。這個比喻讓我們知道若能破解 DNA 語言的文法，共有序列與蛋白質模組的功能，就可以幫助我們瞭解基因的功能、基因間交互作用的方式，也就是生命現象是以「DNA 語言」記載在「生命」這一本「書」的章節中。

語言的比喻使我們知道破解 DNA 語言的作法，可是 DNA 序列中重覆出現的區域是「junk」或是真的有意義？這可由遺傳和變異之間的平衡來討論：

基因體 DNA 雖具有保存遺傳特性的重責，仍須允許有限度的變異(variation)，以維持生物族群的多樣性(diversity)。基因突變(mutation)、複製(duplication)、重組(recombination)、---等都是產生變異的機制。因為基因重組所發生的位置並不固定，在基因重組的過程中，如果生物巨分子不能維持執行功能所需的結構，就有可能會被淘汰。在經過長時間的演化後，具有功能的區域在摺疊成三級結構時，就會逐漸具有獨立摺疊的性質。這是因為當一生物巨分子的重要功能區域具有獨立摺疊模組(IFM, Independent Folding Motif)性質時，若在此功能區域外部發生突變，功能區域因具獨立摺疊的性質，其結構較不受其周遭的突變所影響而能維持功能。當功能區域內部產生突變時，功能區域結構才有較大之機率遭到破壞，致使功能喪失而遭淘汰。而當生物巨分子重要功能區域不具有 IFM 性質時，在此區域外的突變較可能造成功能區域結構的破壞，而使此一基因遭到淘汰。兩相比較之下，在生物巨分子中具有 IFM 性質的功能區域實具有較高的突變耐受力，在演化的過程中要比不具 IFM 性質的功能區域要佔優勢。長久天擇下來，生物巨分子中具有重要功能區域將逐漸得到 IFM 的性質。

這些獨立摺疊的區域不因周圍序列的不同而影響其功能，所以即使被重組到其他的位置，仍能執行其原有的功能，例如在許多蛋白質中都可找到接合 DNA 或 ATP 的模組等。更重要的是在有了許多不同種的功能區域或 IFM 之後，生物體可以利用組合的方式產生各種新的蛋白質，或是增加族群的多樣性。當我們比較多個不同的蛋白質

序列時，有時會發現一小段相似的序列。這些序列通常都具有功能，而且是 IFM，一般稱之為模組樣式(pattern)(參見第 8-2 頁，方盒 8-1)。在 DNA 序列上也可能有重覆序列，通常是蛋白質的辨識位置，只是它們的功能不是以三級結構表現的。因此巨分子上的重覆序列所代表的是一些可能具有功能的區域，也是找尋 DNA 語言中的「字」的關鍵。



3. 交互作用圖譜才能表現出動態的生命

DNA 語言中的「文法」是規範模組之間交互作用的原則，使用「語言」的模型解釋如何由序列資訊來尋找具有功能的模組是很恰當的，但是這個模型卻不易解釋找尋「文法」的方法，因為它無法呈現出基因表現的階層性與時序性等特性。理想化的模型雖是幫助科學家思考一個複雜問題的輔助工具，在另一層面上亦可能會產生問題。針對系統的不同特性，經常需要找最適用的模型，才能真正發揮使用模型輔助思考的功能。

若使用前述汽車藍圖的模型，則蛋白質和核酸都像是生命藍圖中的元件組，是由多個元件(即巨分子中的模組)所構成的，其中每個元件都各有其獨立的功能。在粗看之下，藍圖中所記載的似乎只有各元件組的符號，其實各元件組在時、空中的交互作用

也都記載在藍圖中。只是描述生命的藍圖像一張 *halogram*，必須在有雷射光的照射下才會呈現立體的影像，元件組間的互動關係才可被呈現出來，在此姑且稱它為「交互作用圖譜(interaction map)」。在細胞中，某一群巨分子在做了適當的組合後，就會開啟另一類巨分子的表現。讓它們繼續交互作用，就會產生另一種動作，這一連串的動作雖不像引擎那樣有進氣、壓縮、點火、排氣等週而復始的動作，可是它造成的是生命的延續。經由另一個分子、胞器、細胞、…甚至個體的複製，又進入到「另一個」這樣的路徑。因為它不構成循環，所以有先後次序，因而產生層級(hierarchy)的概念。換句話說，交互作用不但發生於空間中，它還有時間上的意義，因此交互作用圖譜能動態的表現出生命現象。而序列資訊就像是沒有雷射光照射的 *halogram*，要想了解生命現象就要找出使 *halogram* 呈像的雷射光，也就是要由序列中找到交互作用圖譜間的資訊。

因此，若由「使用」資訊的角度來分析，我們似乎應將序列視為一個機械的藍圖。現在我們雖然不了解該怎麼讀這張圖，一旦學會了基本的識圖法則，不但可加速我們了解這機器的速度，也能據此修理、重建甚至重新設計，這裡雖然牽涉到許多道德的問題，那將是另一個討論的主題，在此完全將其忽略，而要討論在藍圖被拼湊出來後，我們要怎樣利用實驗與電腦來協助思考。

4. 電腦將是未來「輔助生物學者思考」的工具

一個細胞的生命藍圖是非常複雜的，若還希望考慮各細胞、各組織、甚至各器官間的交互作用，那麼就須將藍圖再繼續解析到更高的維度。事實上一般人處理多維空間問題的能力，遠不如處理三維空間問題的能力。因此，需要電腦的協助將部份的資訊投射到一個我們可以想像的空間。更重要的是要能由各種不同的角度來看這個物體，才不會忽視可能的交互作用，所以生物學者需要的是一種還不存在的資料庫軟體，以整理完整的序列資訊與交互作用圖譜。此時資訊學者所扮演的角色將是如何整理這些資料，使我們能隨心所欲地找出想判讀的資料，以便預測反應的路徑與調控方式。一個能動態呈現「交互作用」的資料庫軟體，對思考的幫助會遠大於一個指令、一個動作的傳統資料庫軟體。生物學者若說不出自己需要什麼，就很難得到資訊學者適當的協助，因此生物學者必須給出規格來，並集合物理、化學、數學、資訊學者的力量來破解生命的奧妙。

5. 學習如何應用序列資訊遠比學習使用電腦重要

國外的科學傳統是不容忽視的，別人在一個問題上已花了那麼多時間，累積出來許多經驗，我們想要馬上超越別人是非常困難的。雖然在網際網路上有許多免費的軟體，使用者必須自己安裝，如遇問題亦需自己解決，它們是沒有售後服務的。即使可以買到商用程式的執行檔，若無程式碼，仍無法改進其功能，永遠都是受制於人。在商業市場上電腦雖已走上開放系統，在學術界中因為市場不夠大，並沒人提供開發程式用的工具程式，各自發展的結果就是自己必須開發一套程式庫，支援自己所有的程式開發，這些後面的支援系統是買不到的，而我們若不自行發展，將來就會越來越落後。

對彎曲 DNA 極有研究的 Trifonov 博士剛由蘇聯出來時，曾主持 EMBL 生物資訊

小組。他在應徵時有人問他擅用哪一種電腦，他說「我不會用電腦」。別人問他為何來應徵這個工作，他回答說我相信你們要的是一個知道如何應用電腦解決生物學問題的人，而不是一個只會用電腦卻無法解決生物學問題的人，結果他被錄取了。因為EMBL的人深知他們要的是一個懂生物學，卻有能力與資訊學者溝通的人。

二．使用電腦程式的正確態度

使用電腦程式解決生物上的問題，屬於技術的層次，你只是將它當成一個工具，來幫助你確定某種想法是否正確。若對電腦程式有不正確的期望，則會產生不必要的失望，甚至誤解。因為學習態度對學習的效果有直接的影響，所以我要在開始解釋電腦程式的應用前，首先說明什麼是正確的學習態度。

1. 使用電腦程式屬於技術層次

在你做論文時，你可能跑了許多漂亮的凝膠電泳，可是仍沒有可用的實驗結果。這是因為在研究的層次上，我們依觀察設計實驗，並推測可能的結果以測試假說。電泳只是一種輔助測試的工具，技術純熟有助於研究，可是不代表你的假說是正確的。在使用電腦時，同樣應將技術與研究的層次分清楚。電腦只是根據人寫的指令運算，會使用電腦並不意味著你能得到有生物意義的結果。使用別人寫好的電腦程式就相當於在實驗室做實驗，不同的程式就相當於不同的實驗方法，比如 S1 mapping 與 RNase mapping 有時可達同樣的目的，可是各有其優缺點，該視狀況決定使用何種方法。若將使用程式比做凝膠電泳，程式會算出答案，就像凝膠能將分子量不同的物種分離開。會操作程式，得到印出的結果只不過像去做凝膠、連接電極；能讀輸出的結果也不過像由凝膠電泳估計分子量，都是技術層次的事。對新手而言，選用適當的程式是第一個難題。他將面對的第二個難題是選定程式中參數的設定，電腦程式中參數的設定就如同做凝膠電泳時選擇凝膠的種類與濃度，有時必須稍加嘗試才能找到最佳條件，但其規則是有資料可參考的。有時選錯凝膠的濃度或種類就無法將不同的物種分開，所以電腦若算不出觀察到的結果並不是電腦笨，而要問問自己是不是設錯了參數。

凝膠電泳只是達到實驗目標的一種工具，實驗的設計是在於凝膠上各欄結果的比較，而不是選定電泳的條件。因此利用電腦的工具來解決生物上的問題也同樣要經過設計，也要有適當的控制實驗來證明所觀察到的結果不是機率造成的影響。假設我們想找出與 disintegrin 有關的所有序列，以便經由多序列並列分析找到守舊的區域；再根據守舊的區域設計 PCR 的引子，來尋找細胞中新的、與 disintegrin 相關的序列。為了要找到資料庫中所有與 disintegrin 相關的序列，只靠關鍵字的搜尋是不夠的。較好的實驗設計是先找到一個典型的 disintegrin 序列，再找出在資料庫中所有和它相似的序列，這種方法乃是活用程式功能而產生的辦法，而不是死記各程式功能，而能想出的辦法。在資料庫中尋找相似序列也必須測試多個搜尋的條件，只用預設值可能會找不齊所有相關的序列，而找到的序列是否具統計上的意義，又需再經過控制實驗的測試。

對生物學家來講，還有一種問題是不知道如何解釋程式輸出的結果。過度相信或

依賴預測的結果則可能導致錯誤的結論，這就像是在 band-shift 凝膠上計算分子量是沒有實質意義的。因此事先閱讀電腦程式的使用與參考手冊是非常重要的，我們不應該因為不了解電腦而迷信電腦的能力，也不該因失望而否定電腦工具的重要，其實最重要的是要認清電腦程式的使用只是一種技術，它不能代你思考。

2. 如何學習技術？

(a) 略知各技術的原理，以知何處可用此技術

在第一次學習時應先求觀念(架構)的建立，以後有機會再學其間的細節(內容)。

(b) 不必學盡天下所有技術才開始做研究

做研究時，需要怎樣的技術才去學習；否則學後不用，很快就會忘記，用時仍然要再學，真正重要的是學習自學的方法。

(c) 選定方法前先做比較

不同的技術可能可以達到相同的目的。以實驗上決定 RNA 的相對濃度為例，RNase protection、Northern blot 和 primer extension 均可用，可是可能只有一種方法最適合目前的需求。因此應先選定一個最好的方法再深入研讀。徵詢有經驗的人的意見往往可以節省許多時間。

(d) 開始做前先詳讀參考資料

分子生物學上常用的電腦程式依使用與分析的難易程度可分為三大類，第一類是有明確的方法，使用非常簡單，例如尋找限制酵素切割位置等。第二類則需輸入最佳條件，使用時需經驗，但結果的解釋較直接，例如新序列與資料庫的比對。第三類則為預測，使用的方法很簡單，但分析結果時則需對理論有充分的了解，例如蛋白質與 RNA 的二級結構預測等。在學邏輯時，一個很重要的觀念是「雖然邏輯正確，錯誤的前提導致錯誤的結論」。電腦程式的前提在寫程式時已確定，誤用或超出程式在設計時所賦予的功能，可能產生錯誤的結果。了解各種技術的前提或極限，可以避免濫用或錯用某一技術。先讀使用手冊與參考手冊，就像你在實驗之前先讀別人的實驗步驟一樣。除非你已非常有經驗，參考別人的步驟通常比自己來找一切的最佳條件要迅速。深入的了解有助於靈活地運用這技術，甚至發展出新的用法。

三. 使用電腦程式套組的好處

使用 Internet 上的資源，最大的好處就是資料庫更新的頻率高。此外資料庫間的連結也做的比較好，例如 Entrez 系統不但可在序列資料庫間連結，更可連到文獻資料庫 Medline，或是反過來由文獻資料連到序列資料。可是只有資料庫是不夠的，許多全球資訊網的資源還可做序列分析。它們給人的印象是好學易用，而且圖形介面設計精美，讓人賞心悅目，他們的功能可能還比市售的程式更好，所以我們不禁要問，為什麼

麼要學程式套組，而不學網際網路上的資源？

問題是即使像在網路上有「生物工作台(Biology workbench)」這樣受歡迎的序列分析工具，也無法提供像一般程式套組這樣完整的功能。若想完成一些較複雜的工作，就必須連到多個不同的站台去，這時就產生了格式互換的問題。這是因為電腦程式一旦寫出，輸入與輸出的格式就被限制。因此除非經特別的整合，一個電腦程式的輸出檔案可能無法直接輸入另一些電腦程式。而不同的資料庫也有不同的儲存格式，沒有一個電腦程式是能讀取所有資料庫的資料的，這時就產生了格式互換的需求。使用者必須花些精力在處理輸入的格式，初學者若分心去處理這些格式問題，將無法集中心思在解決生物學的問題。例如利用 FastA 搜尋資料後，若希望取部份序列做多序列並列分析，在 GCG 套組中只需在不想分析的序列前做一記號，即可做後續的分析，可是在網路上就必須將要分析的序列中的某一段，一個一個貼列下個分析軟體的網頁上。即使沒有格式互換的問題，你仍必須先由資料庫中取出序列，存到你的電腦上，再將其傳到另一站台上做分析。在網路通暢時不需花太多時間，可是在使用者多時，分析的速度就受到影響。此外，在分析較長的序列時，以圖形顯示結果要比用文字容易分析，可是以圖形顯示結果的程式，通常不容易在網路上找到。

市售的程式解決這個問題的方式是寫一幾乎含有所有的功能的程式套組，這樣在套組內的所有計算就不會有格式互換的問題。同時，資料庫也附在套組內，所以從查閱資料庫開始，到得到分析結果都不會有格式互換的問題發生。因為每一套組都是這樣設計的，從哪一種套組開始學習其實都可以，重要的是要學的紮實。比較知名的幾種套組包括 IG suite、GCG、PC/GENE、GeneWorks。前二者是安裝在工作站電腦上的，可經由網路連線使用，功能較完整，而且計算能力比較強；後二者則分別是 PC 與 Mac 環境下的套組，速度雖然慢但使用較方便且較易使用。

可是沒有一個套組是萬能的，以功能較強的 IG Suite 與 GCG 兩個套組而言，IG Suite 使用起來比 GCG 容易，而且在資料的搜尋、序列之獲取及分子中模組結構之尋找等功能上均比 GCG 好，但是支持援圖形顯示的程式比 GCG 套組少。此外，像 RNA 二級結構分析，親緣關係分析等功能則根本不提供。對於使用熟練的使用者而言，GCG 套組中的程式可利用指令方式執行，因此便於自動化；而 IG Suite 的環境中則必須做互動式之操作，非常耗時。GCG 程式套組自第八版起就增加了使用選單式的使用者介面(必須使用 X-視窗連線)，可是熟練的使用者仍可在 UNIX 環境下工作，享受自動化之便利。可是 GCG 的價格越來越貴，而開放軟體 EMBOSS 的功能已可逐漸取代 GCG。因此在本書中，我們將以問題導向的方式介紹 EMBOSS 套組，只有在 GCG 具有優勢之處，選擇性第介紹幾個 GCG 的軟體，並以實例說明各程式的應用場合。

在另一方面，網路上的序列分析使用的參數預設值(default)，並不見得是最佳的條件。因此在分析某些序列時可得到好的結果，應用到其他序列上時可能就不是很理想，這是使用者使用時必須瞭解的。而且因為網路上的資源的提供者，並不提供使用者任何訓練或諮詢服務，有關輸出結果的解讀必須查閱原始文獻，對許多使用者而言，讀這些文獻是非常困難的。相反的，使用套組不但可以調整程式中的參數，以找尋最佳的條件，更提供使用者訓練與諮詢服務。因此若想瞭解如何解讀輸出的結果，程式的前提與程式的適用範圍，使用程式套組是一個比較好的方法。對需要做大量分

析的人而言，程式套組更優於網路上的資源。因為只要熟悉 UNIX 這作業環境與套組的功能，就可將分析自動化。而為了安全上的考量，網路上的資源通常不讓使用者以指令控制程式的執行，所以無法自動化。

總之，如果你不是經常做序列分析，或是只用到序列分析中某幾種功能，使用網路資源是一個很好的選擇；可是如果要做較複雜、較深入的分析，或是做經常性的分析，那麼使用程式套組會比使用網路資源方便。

四. 生物資訊學的未來

若想進一步瞭解各程式的原理，可使用 Bairoch 所建的「序列分析文獻資料庫」(SEQ-ANALREF, sequence analysis bibliographic reference data bank)。本書中引用，但是未列出文獻均可在「<http://www.ym.edu.tw/bio/courses/>」上找到。

參考文獻

1. Vannevar Bush. (1945) The Atlantic Monthly. 176(1):101-108.

參考網站

<http://www.ym.edu.tw/bio/courses/>

第二章 作業系統與連線方式簡介

葉昌偉¹、楊永正²

¹ 國家高速網路與計算中心、²陽明大學生物資訊研究所

一. 簡介

在這個資訊化的時代，網際網路(Internet)的使用已成為新一代生活的一部份，就連科學研究也與網際網路脫離不了關係。因此在此假設讀者已經知道什麼是 FTP、Telnet、Gopher、與全球資訊網(www)，對這方面不熟悉的讀者，可閱讀本書最後「附錄」中的介紹後，再閱讀本章。

由於 EMBOSS 是一套相當方便也相當成熟的一套生物資訊分析工具，也提供了各種的連線使用模式，所以在使用前，就必須先了解連線的方式，其次要學會一些作業系統的指令，最後才去熟悉 EMBOSS 套組的使用方式。雖然這個學習過程，比使用網際網路上的資源困難許多，可是由第一章的討論中可知，在網際網路上的軟體未臻理想之前，EMBOSS 套組仍是目前最好的選擇。

二. 連線方式

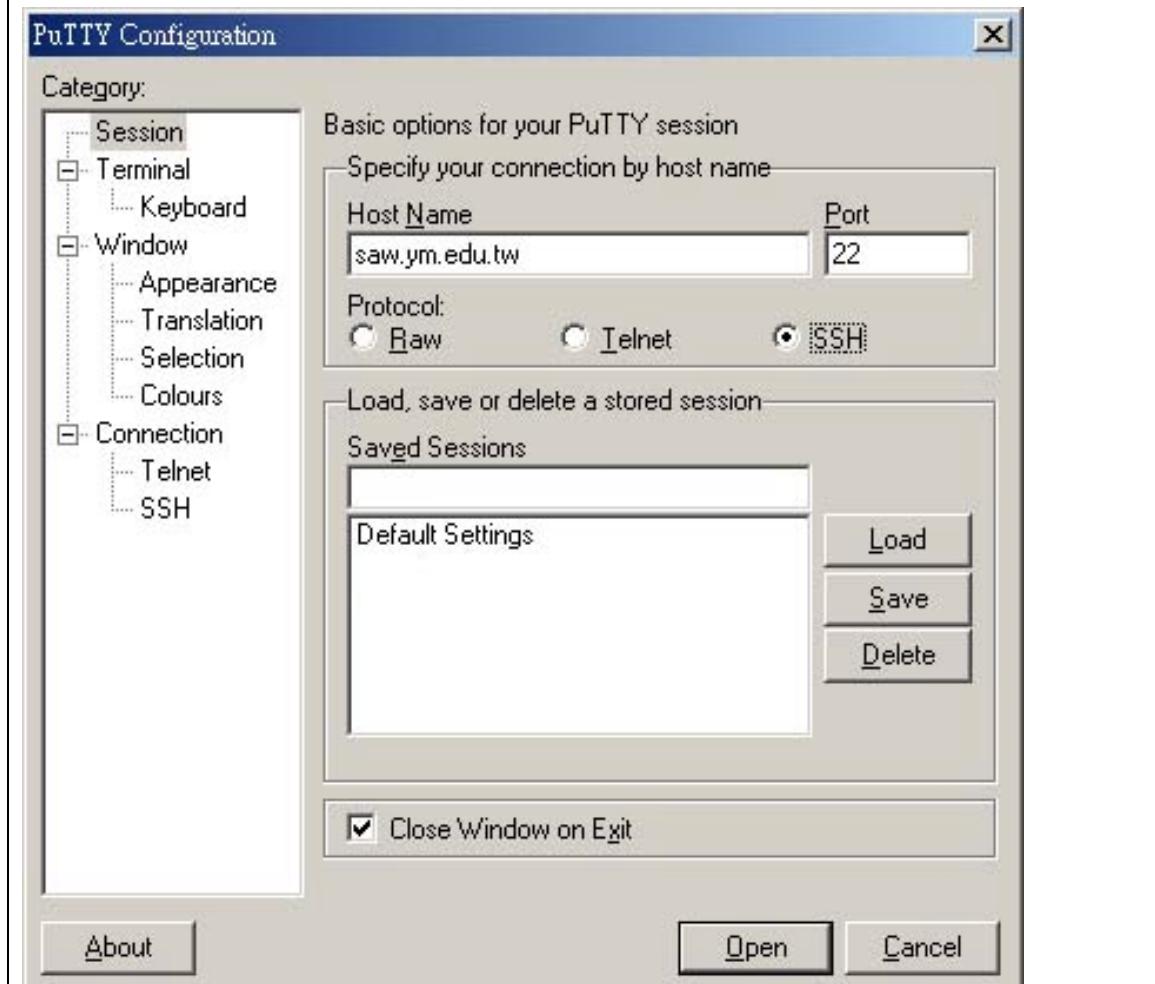
在寬頻網路的世代中，不論是何種性質的機構單位，都會有固接式網路的存在，這類固接式網路包含有台灣學術網路(TANet)，或 HiNet、SeedNet 等公眾網路系統提供的 ADSL 寬頻網路架構，有了這些寬頻網路的連接，便可以進入 EMBOSS 的世界。

1. 簽入 EMBOSS 主機

在此假設讀者已有可使用的固接式網路，對軟體安裝有疑問的讀者應向所在單位的資訊人員詢問。因為此處所舉例之 EMBOSS 套組是安裝在陽明生物資訊的電腦上，所以必須先向陽明生物資訊中心申請帳號，再以 X Window 的形式或是以 Putty 軟體簽入主機使用(X Window 與 PuTTY 軟體的取得與使用請參照附錄一)。

在此先以 PuTTY 做為連結主機的工具，啟動 Putty 程式後，在 Host Name 的欄位輸入主機名稱「saw.ym.edu.tw」(saw 是 sequence analysis workbench 的縮寫)，並且選擇 SSH 連線模式(如圖 2-1所示)，按下 Open 按鈕便會出現簽入畫面。

圖 2-1 使用 PuTTY 程式連接陽明生物資訊中心安裝 EMBOSS 的工作站



在「login as:」處，寫入帳號的使用者名稱(例如 xxx)，按輸入後系統會詢問使用者密碼。在此寫密碼時，螢幕上不會出現字，寫完後按輸入鍵，則可進入工作站主機，而利用 EMBOSS 內的 wossname 程式就可以看到我們能夠使用的 EMBOSS 程式有哪些。

圖 2-2 使用 telnet 連接到陽明生資中心工作站，使用 EMBOSS 的畫面

```

login as: xxx                                     (簽入系統)
Sent username "xxx"
xxxx@saw.ym.edu.tw's password:                  (輸入密碼)
%wossname                                         (查詢EMBOSS可用程式)
Finds programs by keywords in their one-line documentation
Keyword to search for, or blank to list all programs: sequence (輸入sequence)
SEARCH FOR 'SEQUENCE'
backtranseq    Back translate a protein sequence
biosed        Replace or delete sequence sections
btwisted      Calculates the twisting in a B-DNA sequence
chaos         Create a chaos game representation plot for a sequence
.....

```

```
.....
.....
.....
vectorstrip      Strips out DNA between a pair of vector sequences
wordcount        Counts words of a specified size in a DNA sequence
wordmatch        Finds all exact matches of a given size between 2 sequences
yank             Reads a sequence range, appends the full USA to a list file
```

如此便可以查詢此台工作站主機所具有的EMBOSS程式有哪些！

三. UNIX 作業系統的簡介

作業系統(Operation System, OS)是一種平台，他可以讓許多的程式軟體在其上運作，而 UNIX 是一種在學術界常用的作業系統。它功能強大，卻不太容易學，所以在此僅教需要用到的部份，使讀者在有了這些基礎後，一方面足以發揮 EMBOSS 套組的功能，另一方面也有了自學更多的 UNIX 指令能力。

UNIX 和一般個人電腦的作業系統最大不同之處，在於 UNIX 是一個多工及多使用者的環境，他可以提供給不只一人工作服務，除了在主機前面的操作外，使用者可以利用客戶端(client)的連線程式，將終端機(terminal)顯示在遠端的電腦上，但所有的工作依然是在主機上執行，客戶端只做呈現的工作。

每一個使用者所執行的每一個程式，都稱為一個在等待中的程序(process)，而不同程序在處理的過程中，UNIX 會依照順序給予適當的系統資源，每一個程式也都能將程式送至背景(background)工作，這些背景執行的工作，甚至可以在使用者簽出後繼續執行，待其下次簽入後再去檢查程式執行的結果。

由於許多硬體廠商的競爭，對於他們自己的產品都會推出一套較適合自己的 UNIX 系統，所以我們在坊間常常會聽到許多不同的 UNIX 系統，例如像是 Solaris、AIX、IRIX 等。而我們常聽到的 Linux 或 FreeBSD 都屬於在個人電腦上使用的「UNIX」系統，而後者與商業版的 UNIX 最大的不同在於它是完全免費使用的。如果讀者有興趣的話，可以自己試著取得免費的 Linux 自行安裝使用！

1. 簡單指令的使用

要使用一個多人共用的電腦，必須先簽入你個人的帳號。這與簽入終端機伺服器是非常相似的，需由使用者提供帳號名稱與密碼。由系統管理員所發放的密碼，可能對你沒有意義，也不容易記。在你第一次簽入時，可以更改簽入密碼，以避免使用一個你不熟悉的密碼，此時可用「passwd」指令(代表 password)來更改密碼。

範例 2-1 請利用指定的教學帳號與密碼簽入工作站，然後更改密碼為工作站的號碼。

Answer:

login: t001	假設的帳號名稱
Password:	輸入密碼時，密碼不會出現在螢幕上
%	%代表系統提示符號
%passwd	更改密碼
%old password:	避免你離開電腦時，別人更改你的密碼

第二章 作業系統及連線方式簡介

%new password:	輸入新密碼
%retype new password:	要確定新密碼沒有輸入錯誤，所以再輸入一次
Password has changed	告知密碼已用修改成功

在大多數 UNIX 主機內，應當都會安裝好 telnet 或是 SSH 的客戶端程式(telnet 與 SSH 通訊協定請參照附錄一)，所以假設你所簽入的工作站，並未安裝 EMBOSS 軟體，若想連接到有安裝 EMBOSS 的工作站，即可利用 telnet 或是 ssh 的指令來進行工作站的連結。

一旦開始做序列分析，即將產生的眾多檔案，為了整理這些檔案，最好建立子目錄來做分類，這可使用「mkdir」指令(代表 make directory)。若要除去已存在的子目錄，必須先清除子目錄中所有的檔案，再使用「rmdir」指令(代表 remove directory)除去子目錄。若要進入某一子目錄，可以使用「cd」指令(代表 change directory)。在進入另一子目錄後，若不確定現在的子目錄在系統中之位置，可使用 pwd 指令(代表 print working directory)。

練習 2-1 請每人在共同帳號下，以自己的姓，建立一個屬於自己的子目錄，然後進入此子目錄，找出此子目錄在系統中的路徑。

Answer:

% mkdir xxx	(xxx 是自己的姓)
% cd xxx	
% pwd	
??/home/chem/chem/xxx	(此路徑可能與教學用帳號不同)

不論在哪個子目錄下，若只鍵入 cd，而不加任何子目錄名，會自動回到簽入時所在之子目錄。若希望向上跳一層可在 cd 後空格，再加兩個點，即利用「cd ..」指令跳到上一層的子目錄。為了練習管理檔案的指令，必須先取得一些檔案，所以在此要利用 FTP 的功能來抓取一些檔案做練習用。

2. FTP 的使用

以 Telnet 的方式可以將個人電腦模擬成遠方電腦的終端機，以便使用遠方電腦做分析。要將分析結果取回列印，則必須使用 FTP 程式來存取檔案。在此將說明最原始的 FTP 的程式怎麼用，並在練習中讓你去取一個比較好用的 FTP 程式(CuteFTP)，這樣在未來，你可在自己的電腦上安裝此程式使用。

啟動 FTP 程式後，會進入一個互動式的(interactive)環境，每次系統執行完上一個指令，就會出現類似「FTP>」這樣的提示符號(參見範例 2-2)。若是想看有那些指令可用，只須在 FTP 的提示符號下，輸入「?」，即可看到範例 2-2 中列出之眾多指令。在 FTP 程式中所使用的部分指令與 UNIX 作業系統相似，例如列出目錄用 ls，更換目錄用 cd，建立或移除新的子目錄則分別用「mkdir」與「rmdir」。不過對 FTP 的功能而言，真正重要的是建立連線、切斷連線與存取檔案。

範例 2-2 Windows 系統下 FTP 程式的使用

```
C:\>ftp  
ftp> open saw.ym.gov.tw
```

第二章 作業系統及連線方式簡介

```
Connected to saw.ym.edu.tw.  
220 saw.ym.edu.tw FTP server (Digital UNIX Version 5.60) ready.  
User (saw.ym.edu.tw: (none)) :ymuyang (假設的帳號)  
331 Password required for ymuyang.  
Password: (密碼不會出現在螢幕上)  
  
230 User ymuyang logged in.  
ftp> ? (線上輔助系統)  
Commands may be abbreviated. Commands are:  
  
! delete literal prompt send  
debug ls put status  
append dir mdelete pwd trace  
ascii disconnect mdir quit type  
bell get mget quote user  
binary glob mkdir recv verbose  
bye hash mls remotehelp  
cd help mput rename  
close lcd open rmdir  
  
ftp>bye
```

不同的 FTP 程式可能會用不同的指令建立連線，最常用的是「open」，有時也會用「connect」。如範例 2-2 所示，在連線建立後，遠方電腦會要求你簽入，此時輸入你的帳號名稱與密碼即可。若不幸因為密碼打錯等小錯誤而未成功簽入，可輸入「user」再重新啟動簽入的畫面。如果你在對方的電腦上沒有帳號，可試用「anonymous FTP」的觀念(參閱參考文獻)，以「anonymous」或是「ftp」為簽入之使用者名，再以自己的電子郵件地址為密碼即可簽入，甚至有些 ftp 站不需要以電子郵件地址做為密碼即可簽入。在進入不熟悉的系統後，可先取回 readme 檔案，再由其中尋找使用此站的資訊。在使用完畢後即可輸入「bye」，跳離 FTP 系統。

在資料傳輸上，用「put」將檔案放到遠方電腦上去分析。用「get」來取得遠方電腦上的分析結果，如果要一次存或取多個檔案，則分別用「mput」與「mget」，其中 m 是代表「multiple」的意思。程式預設在每次傳輸時會要求使用者確認是否要進行傳輸，若你覺得沒有必要，可用「prompt」指令將此功能關閉。在關閉狀態下，若想啟動確認功能，也是用「prompt」指令開啟此功能。這種使用同一指令開啟或關閉某一功能的做法，有一特別的術語，稱之為「toggle」。

在檔案傳輸的格式上有「ascii」與「binary」兩種。前者在傳輸時有翻譯的功能，而後者則直接將 0,1 很忠實地傳到對方的電腦上。如果兩台電腦所用的編碼系統不同(例如工作站上用的 ascii 碼，與個人電腦使用的 ascii 碼稍有不同)，文字檔若未經適當的翻譯，就會產生錯誤。反之，若將圖形檔案等以 ascii 模式傳輸，在翻譯的過程中會使圖形起變化。所以傳文字檔(text file)時，應使用 ascii 模式；要傳圖形檔、程式檔(execution file)、或是經過文書處理所產生的檔案，則使用 binary 模式。如果兩台電腦使用完全一樣的 ascii 碼，那麼用 binary 模式傳輸文字檔也不會有問題。要設定傳輸的模式只需在提示符號後輸入「ascii」或「bin」即可。其他的指令使用較少，在此不一一介紹。

練習 2-2. 請使用 anonymous FTP 的方式到「ymbc.ym.edu.tw」上取回本書所需之檔案。

```
ftp ymbc.ym.edu.tw  
Connected to ymbc.ym.edu.tw  
-----
```

```
User: anonymous
331 Send your complete e-mail address as password.
Password:

ftp> prompt
200 PORT command successful.
150 Interactive mode off
ftp> mget *.*
200 PORT command successful.
150 Opening ASCII mode data connection for
/Windows/Winsock/Windows95/FTP/1.txt(95254 bytes).
226 Transfer complete.
95252 bytes received in 305.46 seconds (1.62 Kbytes/sec)
-----
```

3. 管理檔案的指令簡介

在作業系統下，若要管理檔案，就必須能查看檔案目錄，瀏覽檔案內容、複製檔案，將檔案改名或刪除檔案等。要查閱目錄，可用「ls」指令(代表 list)，要瀏覽檔案內容有兩種方式，一種是用「cat」指令(代表 catalog)，它相當於 DOS 下的「type」指令，其缺點是文字很快的在螢幕上向上捲動，使用者會來不及看。另一種是「more」指令，會一頁頁地顯示檔案內容。在螢幕左下角有一數字，表示已出現的檔案大小百分比，如果要看下一页就按空白鍵，要一次向上捲一行就按「enter」鍵，要停止顯示就按「q」。

範例 2-3 查閱子目錄下有哪些檔案，並在螢幕上檢視「1.txt」

```
% ls
1.txt          2.txt          5s.connect      run.txt        str-srch.ini
% more 1.txt
```

若覺得檔案名不妥，可用「mv」指令(代表 move)修改檔案名稱，在「mv」後的第一個檔名是要被修改的，第二個檔名是修改後的檔名。複製檔案時，使用「cp」指令(代表 copy)，需要在指令後分別提供來源與目標檔兩個檔案。如果要刪除檔案則用「rm」指令(代表 remove)。

練習 2-3 將檔案「1.txt」更名為「hum-tf3a.dna」，再將檔案「2.txt」拷貝成為「yst-tf3a.dna」。此時目錄下將有兩個檔案具有同樣的內容，因此請刪除「2.txt」。

```
% mv 1.txt hum-tf3a.dna
% cp 2.txt yst-tf3a.dna
% rm 2.txt
%
```

在一個多人使用的電腦上，為了安全上的考量，每一個檔案的使用權限都可個別設定。使用者可以分為三類，檔案擁有者(owner)權限最大，可以讀(read, r)、寫(write, w)、和執行程式(execute, e)；同一使用者群的人(group, 例如同一實驗室的人)，或世界上其他的人(world)則只有讀的權限。三類使用者各有「r」、「w」、「e」等三種權限，如果以「1」、

「0」分別代表具有、或不具有某種權限，則每一類的使用者可以用二位元的數代表權限。例如檔案擁有者的權限是「111」，同一使用者群的人的權限是「100」。若將此數以八位元的數表示，則上述兩個二位元的數分別等於 7 ($= 2^2+2^1+2^0$) 與 4 ($= 2^2+0+0$)。一般在表示使用者權限時，是依檔案擁有者、同一使用者群的人、與其他的人的先後次序以八位元的數寫出，因此 744 是讓檔案擁有者具有最多的權限，其他的人則只能讀，而不能修改或執行檔案。

使用者可用「chmod」指令修改檔案的使用權限，例如一般的檔案多只允許擁有者讀、寫，而不能執行。如果需改為可執行的檔案，可用「chmod 744 xxx」來變更使用者權限，其中 xxx 是要更改的檔名。要想看到檔案的使用權限，必須在「ls」指令後加上一個選項。一般的 UNIX 指令的格式為：指令名稱(選擇項、參數)。指令名稱、選擇項和參數間，至少需要有一個以上的空白隔開。每個指令一定要有名稱，但不一定需要選擇項與參數。在執行「ls」指令時，若用不同的選擇項，會以不同的格式顯示檔案名稱。如果不指定選擇項，即以預設值(短格式)執行指令。在下面的例子中選擇項是「-l」，代表以較長的格式顯示目錄。參數是用來指定指令作用的對象，例如範例 2-4 中的指令會列出副檔名為「txt」的檔案。因為「*」是表示任何一個檔案名，所以在執行上述的例子中的指令時，會列出所有副檔名為「txt」的檔案之全部檔案資訊。

範例 2-4 請在更改使用權限前先顯示權限，再將 run.txt 改為可以執行之狀態，然後再顯示權限，確定更改成功。

```
% ls -l *.txt
-rw-r--r-- 1 chem      c0000000 168 Jan 22 13:05 run.txt
% chmod 744 run.txt
% ls -l *.txt
-rwxr--r-- 1 chem      c0000000 168 Jan 22 13:05 run.txt
```

在學習這些簡單的指令後，你必須學會如何自學 UNIX 這作業系統，因此你必須學會如何用 man 指令(代表 manual)查線上輔助系統，例如「man ls」會說明「ls」指令的使用方法例如選項「-l」之意義。如果希望將說明文件列印到紙上，就必須先將其存入檔案，在 UNIX 下有一個「重新導向(redirection)」的功能，能將前一指令的輸出導入一個檔案。

範例 2-5 試以「重新導向」的方式將「ls」指令的使用方法送到檔案「ls.doc」中

```
% man ls > ls.doc
```

相同的觀念亦可應用到其它指令上，例如將箭頭的方向反過來，會使後面的檔案成為前面指令的輸入值。

練習 2-4 試將所有副檔名為「dna」的檔案資訊存入一個叫「dna.info」的檔案中

```
% ls -l *.dna > dna.info
```

若需將一指令的輸出變為另一指令的輸入，則可利用「piping」的觀念。最常見的就是將程式的輸出做為「more」指令的輸入，這樣使用者才有時間瀏覽輸出的內容。

以上所述各指令整理在表 2-1 中，供不熟悉這些指令的讀者參考。不過我要指出，學習這些指令最好的方法是練習使用它，而不是死背列出的各指令。

表 2-1 與管理檔案有關的指令整理

指令	使用範例	說明
passwd (password)	passwd	更換密碼
pwd (print working directory)	pwd	查看現在所在的子目錄
mkdir (make directory)	mkdir doc	建立子目錄 doc
cd (change directory)	cd doc	變換子目錄到 doc
rmdir (remove directory)	rmdir doc	刪除子目錄 doc
ls (list)	ls	查看檔案目錄
more	more ls.doc	一次展示一螢幕之檔案內容
cat (catalog)	cat	展示檔案內容
rm (remove)	rm ls.doc	刪除檔案 ls.doc
mv (move)	mv ls.doc x.doc	更改檔案名稱 ls.doc 為 x.doc
cp (copy)	cp 1.txt 2.txt	複製檔案 1.txt 為 2.txt
chmod (change mode)	chmod 744 x.doc	更改檔案使用權限
man (manual)	man ls	查閱手冊
> (redirecting std. output)	man ls > ls.doc	將前一指令的輸出導入檔案
< (redirecting std. input)	sort < 1.txt	將後一物件導為前一指令的輸入
(piping)	history more	將前一指令的輸出導入另一指令，成為其輸入

4. 執行背景工作

在了解到背景工作對序列分析的重要性後，我們將討論如何送出背景工作，如何檢視背景工作的狀態，與如何取消一個誤送或是該關閉的背景工作。在 UNIX 下有兩種方式將工作送入背景執行，第一種是在執行程式時即送入背景執行。第二種則是在程式開始執行後，先暫停執行，再將工作送入背景中執行。

執行程式時，只需將檔案全名輸入，即可啟動程式。若在檔名後加上「Apresand(即 & 符號)」，此程式將被送入背景執行，此時螢幕上會出現一個系統給定之工作號碼(job number)。若要查閱進度或終止程式的執行，就必須使用這獨一(unique)的編號。例如在使用 X-視窗連接時，如果需要開第二、甚至第三個 X-視窗，就必須在背景中執行「xterm」，才能繼續在前景中執行 EMBOSS 程式。雖然目前尚未講到 X Window 的部份，不過可利用此指令為例來說明背景工作的執行。

範例 2-6 假設你在一個X-視窗中工作，請在背景中開兩個新的X-視窗，並觀察其工作狀態

```
% xterm &  
[1] 29712      (此為 job number)
```

此時在螢幕上會出現另一X-視窗，猶如重新簽入一次，可是實際上只是新開一個視窗而已。若再執行一次「xterm」就會開第三個X-視窗。

```
% xterm &  
[2] 29730  
% jobs  
[1] +running    xterm  
[2] -running    xterm
```

另一種將工作送入背景執行的方式，是以 **ctrl-z** 中斷已送出的工作，此時螢幕上會出現「**stopped**」的訊息，此時若在提示符號後輸入「**bg**」，即可將重新啟動之工作送入背景執行。此法在執行的程式仍會不斷顯示進度或結果，恰與在第一種(即加上「&」符號)執行方式時，螢幕上不會出現任何訊息相反。

因為 UNIX 是一個讓多人都有多工便利的環境，為了公平地讓不同的人都有執行背景工作的機會，若想檢視有哪些背景的工作正在執行，可使用 **jobs** 指令。若想除去已送出的工作，則可用 **kill** 的指令。

範例 2-7 接續範例2-5，執行「run.txt」，再將其中斷，送到背景中執行，最後再列出在背景中執行的工作，並將此工作取消。

```
% run.txt  
% ^z  
stopped  
% bg  
% jobs  
% kill 23543
```

表 2-2 中列出與執行背景工作相關的指令，這個前景與背景的觀念有助於未來將 EMBOSS 程式的執行自動化。

表 2-2 與執行背景工作相關的指令整理

指令或符號	使用範例	說明
&	% fasta &	在背景中執行程式
Jobs	% jobs	顯示送入背景之工作
kill	% kill (job#)	移除指定之批次工作
ctrl-z		暫停前景中執行之程式
bg	% bg	將暫停之工作送入背景執行
fg	% fg	將背景之工作拉回前景執行

5. 簡化指令的執行

如果使用者覺得在 UNIX 下的指令不容易記憶，可用「alias」指令(中文是「別名」的意思)重新定義這些指令。若在提示符號後只輸入「alias」，會列出目前所設定的所有別名。我們將在「.log」檔中定義一些新的「別名」。在每次簽入電腦時，UNIX 會自動執行此檔案中的指令，因此每次簽入電腦時都會自動設定別名的意義。

為了避免重覆輸入不久前執行過的指令，尤其是一些長的指令，重覆輸入時可能有錯。此時可以利用作業系統會自動記住過去剛執行過的一些指令的特性。欲查閱已執行過的指令，可用「history」指令，顯示這些已執行過的指令，在這些指令之前有一序號做識別用。若要重複執行這些指令，可在想執行做指令的序號前加入「!」，即會自動執行。例如若要執行簽入後的第三個指令，可以用「!3」。如果要執行前一指令，則不必查閱序號，只需用「!!」即可。如果已執行過的指令很多，一頁看不完，可利用前述的「piping」功能，以「history|more」指令一次顯示一頁。

在執行程式的中途，如果發現輸入有錯或是等了很久都沒有反應，希望能重新開始，即可使用「ctrl-c」中斷程式的執行。此指令與「ctrl-z」的差異在於「ctrl-c」是永久中止程式的執行。而「ctrl-z」只是暫時停止程式的執行，仍可用「bg」或「fg」等指令重新從中斷點繼續，所以不能用來改錯。在表 2-3 中列出上述指令或按鍵，供初學者查閱之用。

表 2-3 與程式執行有關的指令整理

指令或按鍵	說明
Alias	指定指令的別名
history	使用指令的歷史
!	執行已執行過的指令
Ctrl-c	停止程式執行
Ctrl-z	暫時停止程式執行

四. 文件編輯器 vi 的使用

利用 PuTTY 進入工作站主機後，若需使用文件編輯器 vi 的話，可在提示符號後鍵入「vi filename」即可進入編輯器。此編輯器有兩種狀態，初進入時是在指令狀態，不能輸入。必須輸入「a」指令或「i」指令進入輸入模式(參閱方盒 2-1)，才能鍵入數據。如果要做修改又必須按「Esc」鍵回到指令模式，所以令人覺得它不好用。此外，它不能利用箭頭符號移動游標，必須記住許多鍵的意義才能自由移動，也是不方便之處。

方盒 2-1 vi 編輯器的按鍵使用法

移動游標：

h 鍵將游標向左移一格。

j 鍵將游標向下移一格。

k 鍵將游標向上移一格。

l 鍵將游標向右移一格。

在檔案中編輯文字：

x 鍵表示刪除游標所在位置的字母。**dd** 鍵 - 刪除游標所在的行。

Y 鍵 - 複製游標所在的行。

p 鍵 - 貼上複製的內容。

u 鍵 - 將前一個動作還原。

在檔案中輸入文字：

a 鍵 - 在游標後開始文字輸入狀態。

i 鍵 - 在游標前開始文字輸入狀態。

o 鍵 - 在游標所在的行之下加入新的空行，並進入文字輸入狀態。

O 鍵 - 在游標所在的行之上加入新的空行，並進入文字輸入狀態。

搜尋字串：

/字串 鍵 - 往後尋找具有該字串的行列。

?字串 鍵 - 往前尋找具有該字串的行列。

儲存已完成的編輯工作：

:w 鍵 - 將目前以修改的部份存檔。

結束 vi 程式：

:q! 鍵 - 終止編輯工作，放棄所做的修改。

:wq 鍵 - 將所做的修改存檔後離開。

取得 vi 的線上說明：**% man vi**

參考文件：Engineering Computer Network at Purdue University Computing manuals (<https://engineering.purdue.edu/ECN/Resources/KnowledgeBase/Docs/20020202121609/>)

在使用 EMBOSS 程式時，有時需要做一些簡單的編輯工作，可以利用 vi 來做。複雜的工作最好在個人電腦上先做好，再以 FTP 送到主機上使用。

範例 2-8 請使用 vi 建立一個檔名叫「pep1」的蛋白質序列檔，其序列为「FHNIKI」

% vi pep1

進入 vi 後，按「a」鍵，輸入「FHNIKI」

再按「:wq」鍵退出 vi 即可

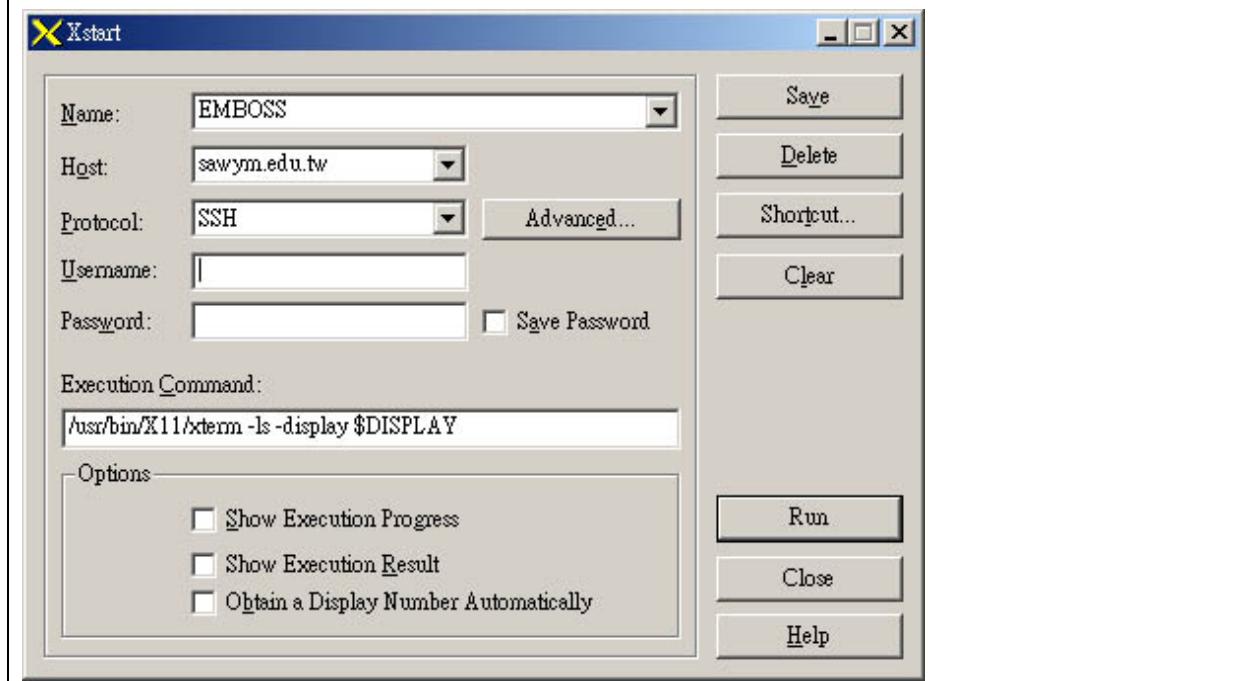
五. 顯示圖形與列印

在分析複雜的資訊時，分析圖形遠比分析文字檔案有效率。而當我們在使用 EMBOSS 套組時，大致上有兩種方式來觀看圖形的輸出，第一種方式是利用 X Window 來顯示圖形，此法較為方便，但圖形不能即時列印；第二種方式是產生輸出檔(如 PNG 圖形檔或是 postscript 文件格式檔案)，再利用 ftp 程式將輸出檔案抓回，利用影像處理軟體(ACDSee、Photoshop 等)或是 Internet Explorer 來顯示圖形檔或是利用 GSview 程式來看 postscript 檔案，此法雖較為複雜，可是列印圖形卻相當的方便。因此以下將以這兩種不同的方式為主體來討論如何使用與列印。

X Window 是 UNIX 系統下的一個視窗環境，只要在遠方電腦上的程式支援 X Window，連線到其上的電腦即可透過客戶與伺服器(client-server)的架構，來顯示遠方電腦的螢幕。它在自己的電腦上建立一個 X-server，然後在遠方電腦(客戶端)上利用「xterm -display」指令尋找這一台 X-server。一旦 client 找到 X-server，即可在使用者的電腦螢幕(X-server)上建立一個 X Window，顯示遠方電腦(客戶端)的螢幕。在此，視窗中的畫面，與你用 telnet 或是 ssh 方式模擬圖形終端機無異。可是 X-視窗可充分運用 UNIX 環境下的多工能力，開啟多個視窗，執行不同的程式，並且由於是圖形化的介面，同時可以顯示圖形。其實在一個視窗內，將工作依序丟到背景中執行也可以，只是在看結果時無法並列比較而已。

現今已經有許多在 Win32 架構下的 X Window 模擬器，較著名的有 Xmanager 與 X-Win32，在附錄一中會介紹 Xmanager 的取得與安裝，在此直接採用 Xmanager 實際操作圖形的顯示。執行 Xmanager 中的 Xstart 即可看見以下視窗(圖 2-3)：

圖 2-3 Xstart 操作介面



在此視窗中指定 Host 為「saw.ym.edu.tw」，並且選擇 Protocol 為「SSH」，輸入帳號密碼後即可取得 X Window 畫面，當進入了 X Window 的世界後，其實就和利用 PUTTY 所產生出的終端機並沒有太大的差別，所以便可以進行 EMBOSS 套組的使用。我們試著利用 dottup 這支程式來產生圖形，再此不贅述此程式的用途及意義，僅就圖形的產生做說明。

範例 2-9 直接利用 X Window 顯示圖形

```
%dottup                                     執行dottup程式  
Displays a wordmatch dotplot of two sequences  
Input sequence: embl:xl23808      指定第一條序列  
Second sequence: embl:xlrhodop     指定第二條序列  
Word size [10]:                      word size設為10  
Display as data [N]:                  不以data形式顯示資料  
Graph type [x11]:                    以X window方式顯示圖形  
如此一來，系統會自行產生一視窗顯示圖形。
```

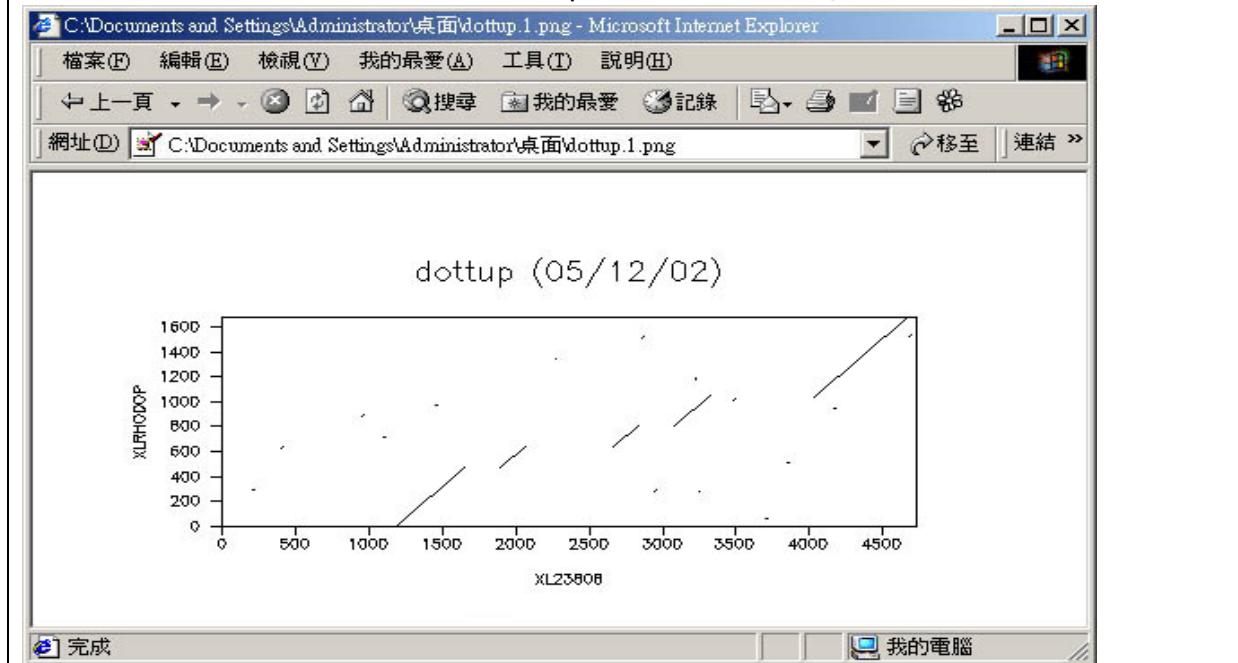
自從全球資訊網創建以來，JPEG(Joint Photographic Experts Group)與GIF(Graphics Interchange Format)是兩個主要圖像顯示的檔案格式，但是由於 GIF 檔案格式發生了一些演算法著作權的問題，以致於大家著手在公用格式的開發，而產生了一個更有效的圖形檔案格式來取代 GIF 檔，這個檔案格式名為可攜式網路圖形(Portable Network Graphics, PNG)，而 EMBOSS 採用的正是 PNG 格式的圖形檔案，下面範例 2-10 介紹如何產生 PNG 圖形檔。

範例 2-10 直接產生圖形檔案

```
%dottup                                     執行dottup程式  
Displays a wordmatch dotplot of two sequences  
Input sequence: embl:xl23808      指定第一條序列  
Second sequence: embl:xlrhodop     指定第二條序列  
Word size [10]:                      word size設為10  
Display as data [N]:                  不以data形式顯示資料  
Graph type [x11]: png                以png形式輸出圖形，若要以postscript形式輸出在此填入ps或  
postscript即可  
Created dottup.1.png                 dottup程式告知產生出dottup.1.png圖形檔
```

在產生完圖形檔後，即可利用先前介紹之 FTP 工具將檔案取回，利用 Internet Explorer 即可作為瀏覽圖形及列印的平台(如圖 2-4)。

圖 2-4 在 Windows 可以利用 Internet Explorer 作為圖形瀏覽器



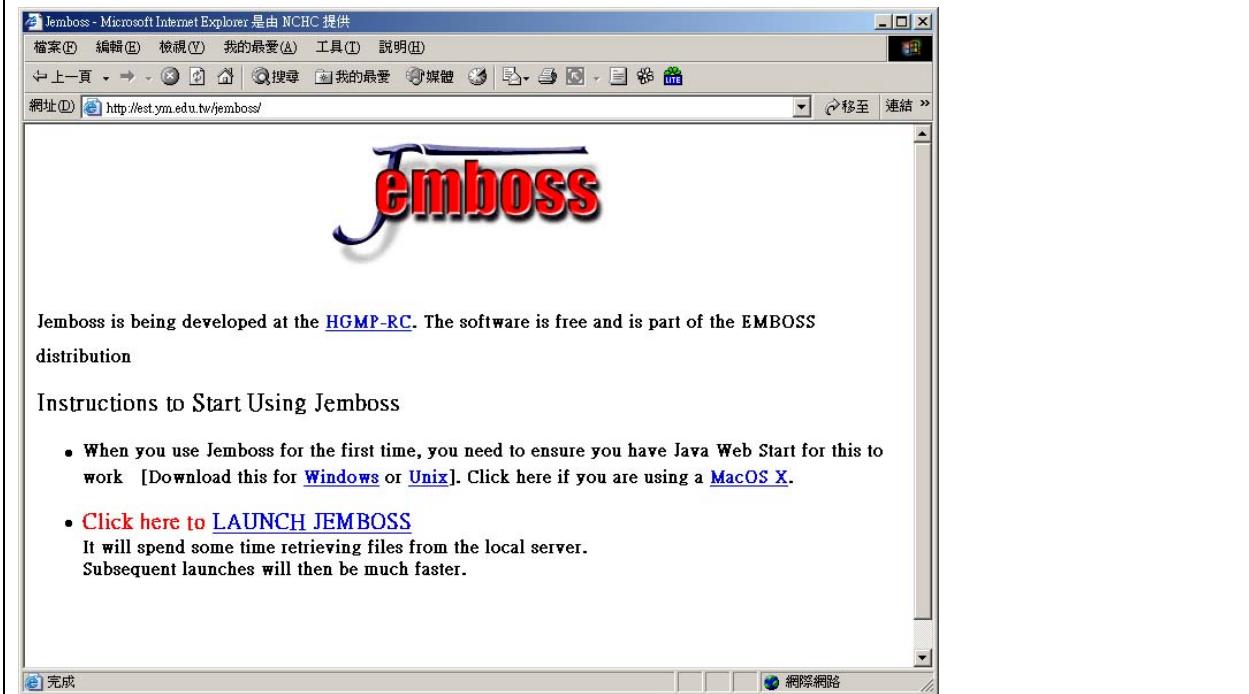
六. EMBOSS 使用介面

一般而言 EMBOSS 都是在 UNIX 系統下以指令模式(command mode)執行的程式，由於不是每個人都對於 UNIX 系統相當熟悉，所以現今也有其他不同使用介面的 EMBOSS 供使用者做選擇，目前較為通用的是網頁介面的 EMBOSS 以及最近才由英國 Human Genome Mapping Project Resource Centre(HGMP-RC)所推出的 Jemboss 介面。

在陽明大學生物資訊中心 (<http://saw.ym.edu.tw/emboss/>) 首先推廣 EMBOSS 之後，目前國內有許多機構安裝有網頁介面的 EMBOSS，例如行政院國科會國家高速電腦中心 (<http://saturn.nchc.gov.tw:9091/Pise/>)，國家衛生院以及清華大學 (<http://emboss.life.nthu.edu.tw/>)。

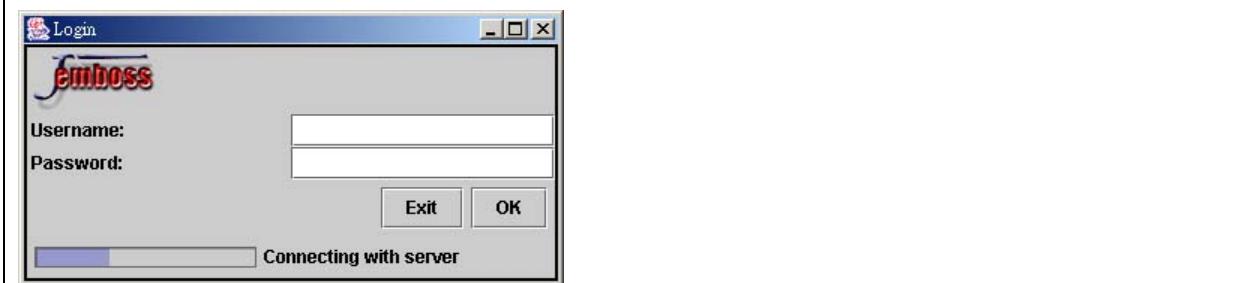
目前正由陽明大學生物資訊研究中心又開始推廣 Jemboss。Jemboss 是一套由 Java 程式語言寫的圖形化使用者介面，必須要申請帳號才能使用。只要有帳號，Jemboss 工作站也可以讓使用者以 X Window 的方式，使用其圖形使用者介面。此外，其它安裝有 Java client 端程式的電腦(UNIX、PC 及 Mac) 也都可以使用其圖形化介面。在使用 Jemboss 之前，Jemboss 網頁 (例如 <http://biomed.ym.edu.tw/jemboss/>，圖 2-5) 下載 Java Web Start。

圖 2-5 Jemboss 網頁，此處可下載 Java Web Start 並且可 launch Jemboss



在安裝完 Java Web Start 後，只要按下網頁(如圖 2-5)中的 LAUNCH JEMBOSS 連結，即可安裝 Jemboss 的客戶端程式(系統會自行將 Jemboss.jnlp 檔抓回來執行)，這一個檔案會幫助你安裝最新版的 Jemboss。在安裝的過程中，Jemboss 伺服器端會要求使用者輸入帳號密碼(圖 2-6)。

圖 2-6 Jemboss 簽入畫面，輸入使用者帳號和密碼



在輸入完密碼後，可能會遇到一個問題，此時會出現以下畫面(圖 2-7)：

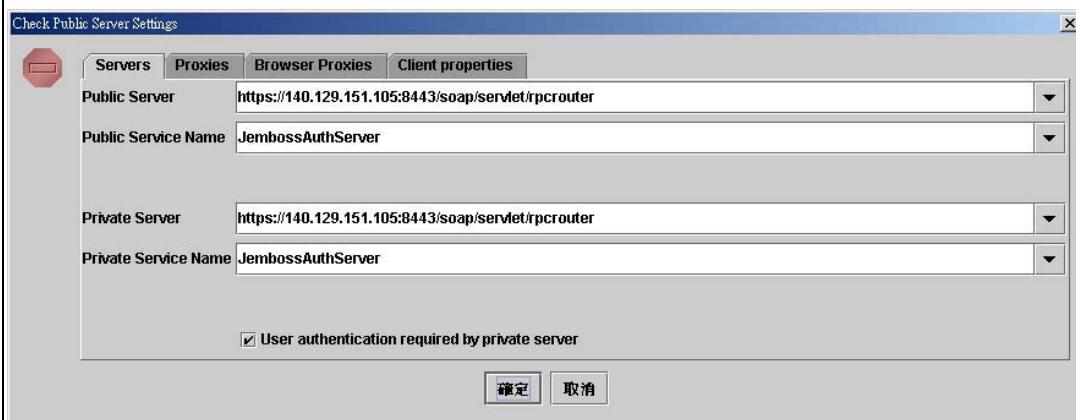
圖 2-7 出現連結錯誤



第二章 作業系統及連線方式簡介

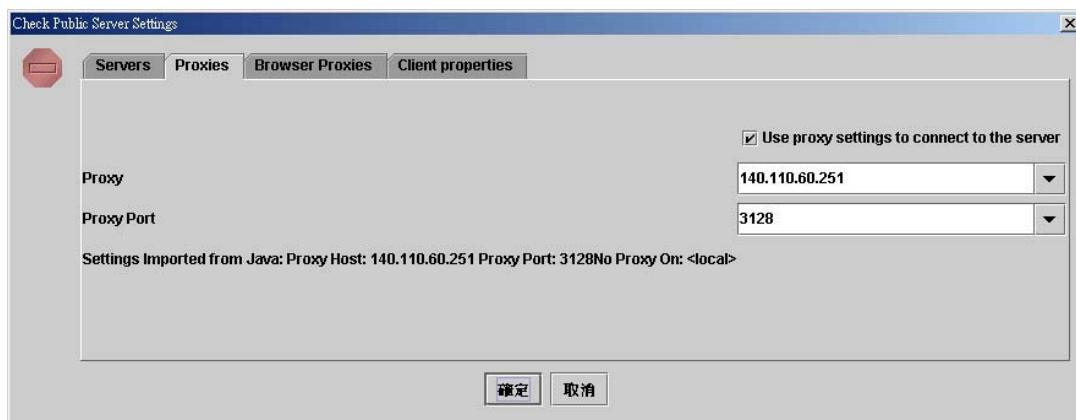
同時另外還會出現一個 Jemboss Public Server Settings 的視窗(圖 2-8)：

圖 2-8 Public Server Setting 視窗



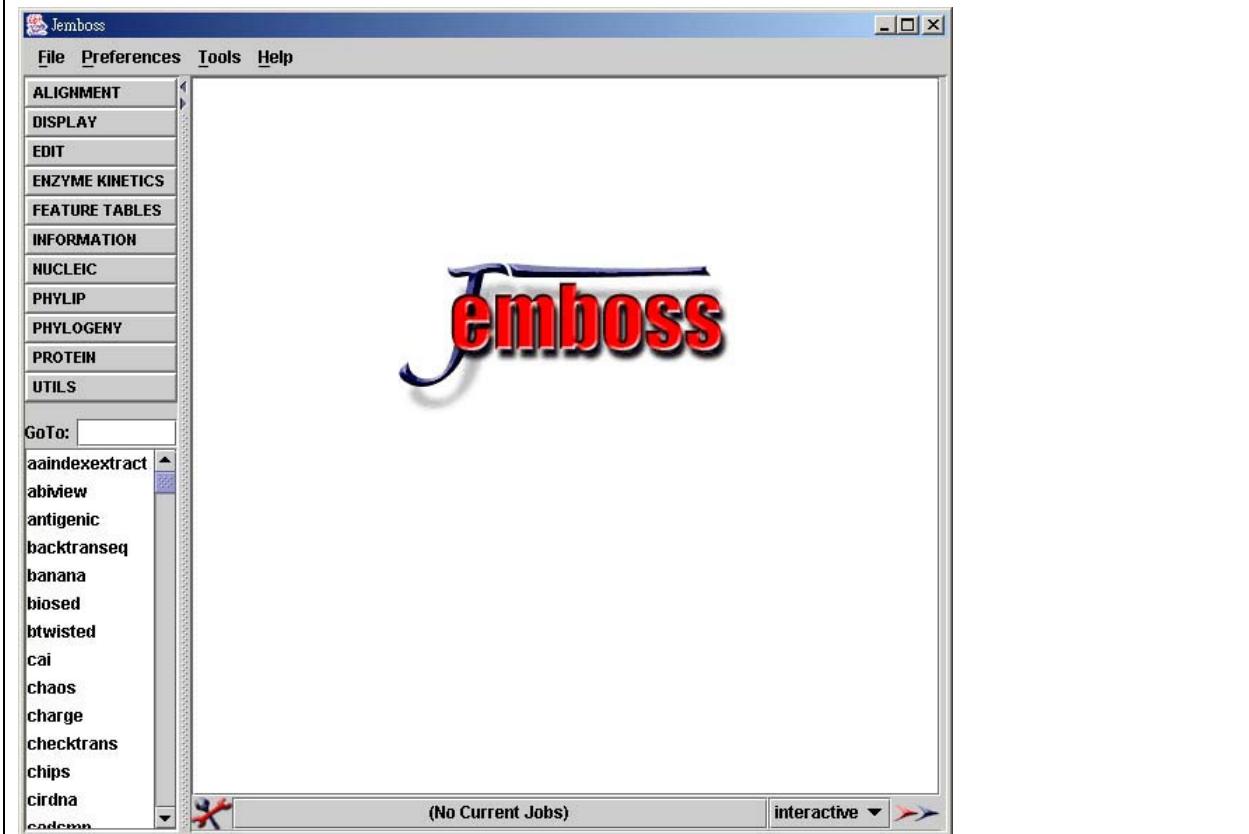
由於此版本的 Jemboss 使用的代理伺服器(Proxy)時會出現一些問題，所以我們必需將此設定表單中的 Proxy 選項中的 Use proxy settings to connect the server 方框中的「勾」拿掉(圖 2-9)，如此一來才能順利的連結到 Jemboss 伺服器。

圖 2-9 將 Proxy 表單中的打勾處移除，方能使用 Jemboss



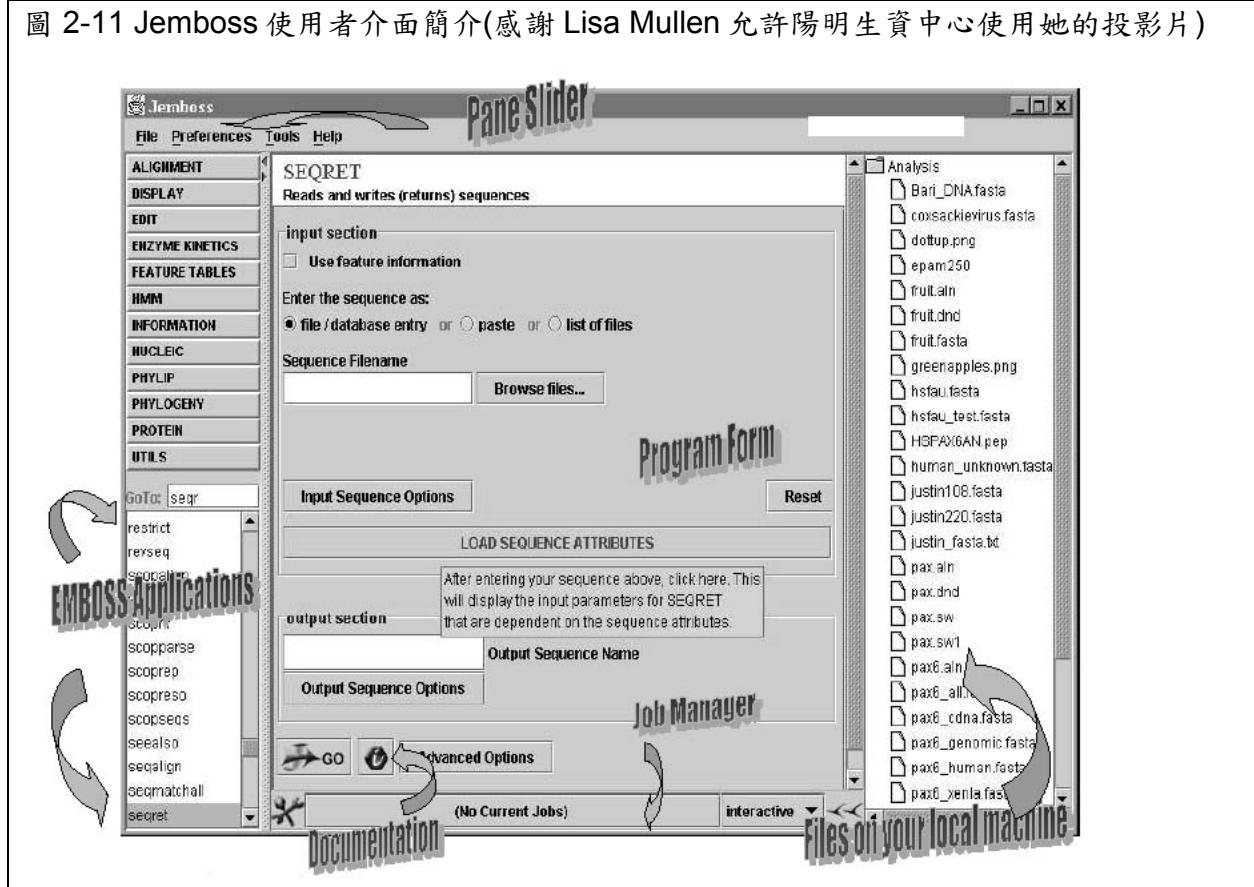
將此打勾處移除，並且按下確定按鈕，即可進入 Jemboss 的使用者介面(圖 2-10)。

圖 2-10 Jemboss 使用者介面



進入此介面後，可以按下右下角的向右箭頭，則會出現一個檔案管理員視窗(圖 2-11)。

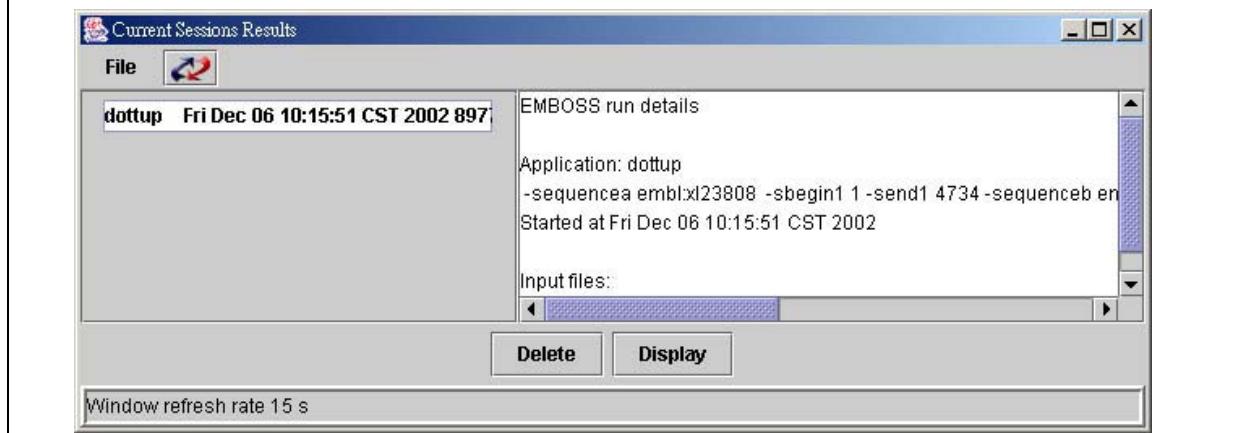
圖 2-11 Jemboss 使用者介面簡介(感謝 Lisa Mullen 允許陽明生資中心使用她的投影片)



Jemboss 使用者介面大致上可分為四區域，第一區域為左邊的 Jemboss 程式目錄，目錄分為兩種方式排列。上方的是依照程式的功能做為區分，下方是依照程式名稱的字母排序，其中還可以利用 GoTo 欄位直接輸入程式名稱，即可使用 Jemboss 中的分析工具。第二區域為程式表單，每一個程式的表單都不相同，其中 Advanced Options 按鈕按下後，即可進行每個程式的所有參數設定，按下 [i] 按鈕，可以讀取該程式的介紹。

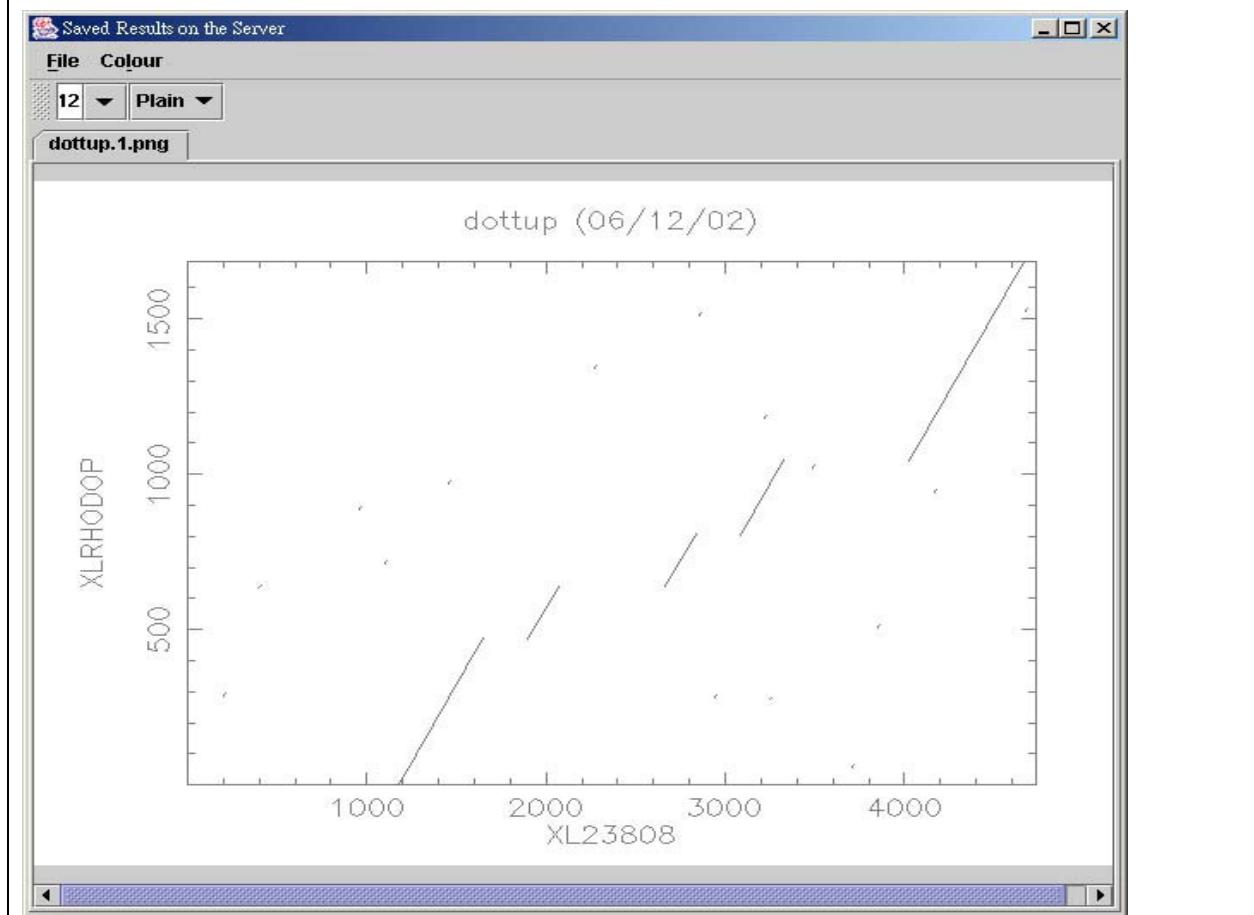
第三個區域為最下方的工作管理員，右邊的 `interactive` 選項也可改為 `batch`，這兩種方式差異在 `batch` 會將工作丟至伺服器背景執行，使用者可以進行另外的工作，而 `interactive` 則是即時等待伺服器將結果回傳到客戶端，若使用者欲進行的分析工作可以在短時間內完成，則選擇 `interactive` 選項，反之則選擇 `batch`，使用者選擇 `batch` 將程式丟至伺服器背景執行時，按下旁邊的 `Job` 按鈕，可檢視工作進度及結果，而此程式所有的指令模式參數設定也都顯示在右邊的視窗內(圖 2-12)。

圖 2-12 工作管理員視窗



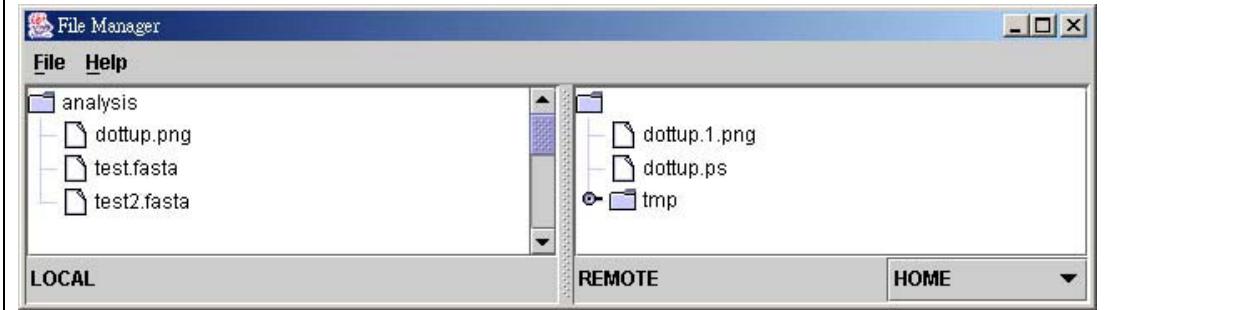
使用者也可將此指令直接在 UNIX 指令模式下執行，也會得到相同的結果，另外選取欲閱覽的工作後，按下 Display 按鈕則可檢視分析結果(圖 2-13)。

圖 2-13 檢視分析結果圖形



第四區域則是檔案管理員，此區域可顯示客戶端所有的檔案，使用者可以設定一專屬資料夾做為 EMBOSS 分析之用，另外按下左上角 File 內的 Local and Remote Files 則會同時出現客戶端與伺服器端的檔案列表(圖 2-14)。

圖 2-14 檔案管理員視窗



七. 結語

雖然在此簡介許多連接 EMBOSS 的方法，可是這些都是過渡性的方法。如第一章所述，在下一代的生物資訊學工具將採用工作流程的方式，串接 web 服務。目前陽明生資中心已在積極研發，相信在不久的未來，生物學者將可使用更簡單與直接的方式使用 EMBOSS，而且可能與資訊搜尋整合在一起。

在過去生物學者需要學習如何寫程式，才能做鉅量(high-throughput)分析。在有了下一代的生物資訊學工具後，生物學者將可將重點放在解決生物學的問題上，而不需要先學習程式寫作。這就像是 20 年前做基因選殖可以取得博士，可是現在連高中生都可以做基因選殖一樣。

參考網站

1. <http://biomed.ym.edu.tw/jemboss/>
2. <http://emboss.life.nthu.edu.tw/>
3. <https://engineering.purdue.edu/ECN/Resources/KnowledgeBase/Docs/20020202121609/>
4. <http://saturn.nchc.gov.tw:9091/Pise/>
5. <http://saw.ym.edu.tw/emboss/>

第三章 自學 EMBOSS 套組的基本技能

賴俊吉¹、楊永正²

¹ 陽明基因體研究中心、² 陽明大學生物資訊研究所

學習任何軟體都應先了解它的線上輔助系統，以便在練習時能隨時查閱使用法。其次要能掌握程式運作的原則，這樣才能舉一反三地試用未學過的程式。本章的目的在於提供使用者這些自學的基本技能，使讀者能由此書中體會出使用 EMBOSS 套組的方法與自學的要訣。

一. 線上輔助系統使用法

在 EMBOSS 的環境下提供兩種線上輔助系統，分別是 wossname 與 tfm。使用這兩種系統時，只要在提示符號(prompt)後鍵入「wossname」或「tfm」即可。wossname 提供以關鍵字方式尋找 EMBOSS 套組中的相關程式，而 tfm 則是 The Fine Manual 的縮寫字，會列出單一程式的詳細使用手冊。tfm 列出的程式使用手冊源自 EMBOSS 套組發源地 HGMP-RC 網站中的使用手冊 <http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/>，若你使用的是最新版的 EMBOSS 套組，那麼程式使用手冊內容是一樣的。EMBOSS 套組目前共有超過一百七十支程式，我們可以 wossname 找出要用的程式名稱，再以 tfm 來看詳細使用手冊。在此先介紹 wossname 如何用，其使用方法可參見方盒 3-1。

方盒 3-1 wossname 的使用法

%wossname

Finds programs by keywords in their one-line documentation

Keyword to search for, or blank to list all programs:

ALIGNMENT CONSENSUS

cons Creates a consensus from multiple alignments

megamerger Merge two large overlapping nucleic acid sequences

merger Merge two overlapping nucleic acid sequences

ALIGNMENT DIFFERENCES

diffseq Find differences (SNPs) between nearly identical sequences

ALIGNMENT DOT PLOTS

dotmatcher Displays a thresholded dotplot of two sequences

dotpath Displays a non-overlapping wordmatch dotplot of two sequences

dottup Displays a wordmatch dotplot of two sequences

polydot Displays all-against-all dotplots of a set of sequences

(以下省略)

在提示符號後鍵入 `wossname`，程式請使用者鍵入關鍵字搜尋或是直接按輸入鍵讓程式列出 EMBOSS 套組所有程式。在此 `wossname` 程式會分組列出程式，但它不會分頁列出，若能將其存入檔案中，我們才方便查詢。以下是將 `wossname` 的程式列表存入 `programs.list` 這個檔案的用法。

```
%wossname -search " -outfile programs.list  
Finds programs by keywords in their one-line documentation
```

顧名思義 `tfm` 所提供的是一個像個別「程式使用手冊」那樣的輔助。如範例 3-1 所示，`tfm` 分頁列出整個程式使用手冊。當你使用 `tfm` 時，也可以 `-outfile` 參數指定檔案，將程式使用手冊寫入檔案中。

範例 3-1 使用 `tfm` 查詢 `tfm` 本身的程式使用手冊

```
% tfm tfm  
Displays a program's help documentation manual  
          tfm  
          Program tfm  
Function  
  Displays a program's help documentation manual  
Description  
  This program displays the help documentation for an EMBOSS program.  
  The contributors of the EMBOSS programs do attempt to provide an  
  adequate description of the programs. This documentation is primarily  
  held as HTML pages at http://www.uk.embnet.org/Software/EMBOSS/Apps/  
  The documentation is also available however through the program tfm as normal text.  
--More--(12%)
```

初學 EMBOSS 套組者並不曉得個別程式的名子，因此先用 `wossname` 程式以類別列出所有程式或是以關鍵字查詢出相關程式，再去找到可應用的個別程式。在提示符號後鍵入「`tfm program_name`」直接顯示程式說明(例如想看 `textsearch` 程式的說明，可鍵入「`tfm textsearch`」)。

練習 3-1 請以 `tfm` 查詢 `wossname` 程式的使用手冊

```
Answer: tfm wossname  
Displays a program's help documentation manual  
          wossname  
Function  
  Finds programs by keywords in their one-line documentation  
Description  
  This allows a user to search for keywords or parts of words in the  
  brief documentation (as displayed by a program when it first starts).  
  The program name and the brief description is output. If no words to  
  ---  
  This program may find some use in automatically generating lists of  
--More--(4%)
```

第三章 自學 EMBOSS 套組的基本技能

在瞭解如何使用 wossname 與 tfm 後，就可以進一步解釋 EMBOSS 的線上輔助系統了。在 Unix 環境下，若直接在提示符號後鍵入「wossname translate」，可看見範例 3-2 的螢幕。此時螢幕上會出現三個和轉譯有關的程式。我們可以 tfm 查閱個別的程式使用手冊。我將以「backtranseq」為例說明閱讀程式線上輔助系統的原則。

範例 3-2 利用 wossname 查閱和「轉譯」相關的程式

```
%wossname translate
Finds programs by keywords in their one-line documentation
SEARCH FOR 'TRANSLATE'
backtranseq      Back translate a protein sequence
prettyseq        Output sequence with translated ranges
transeq          Translate nucleic acid sequences
```

在執行「tfm backtranseq」後，螢幕會列出「Function」、「Description」、等十多個副標題(範例 3-3)。因為每個程式的說明大概都有這十幾個選項，要想很快地找到想要的資訊，就需要知道這些副標題的意義。

範例 3-3 利用 tfm 查閱「反轉譯」程式的使用

```
Displays a program's help documentation manual
backtranseq
Function
    Back translate a protein sequence
Description
    backtranseq takes a protein sequence and makes a best estimate of the
    likely nucleic acid sequence it could have come from. It does this by
    codon frequency table. It is important to use a codon frequency table
--More--(11%)
```

「功能(Function)」是以一或兩句話簡述程式的功能，在每次執行程式時螢幕上出現的說明就是這簡述(範例 3-4 中未列出，請參閱範例 3-3)。「說明(Description)」則是比較深入的討論。每一個程式都有一個「用法(Usage)」的副標題，這是將程式執行的過程直接錄下來，使用者可由此範例中瞭解程式需要那些輸入檔，會產生怎樣的輸出檔案。因為整個程式執行的過程一目瞭然，讀「用法」要比讀「說明」容易抓到要點，初學者宜由此入手瞭解程式的運作。若有看不懂的，再查閱其他相關的副標題。

範例 3-4 backtranseq 程式的使用範例

```
Usage
Here is a sample session with backtranseq. Note that this is a human
protein and so the default (human) codon frequency file is used (i.e.
is not specified).
% backtranseq
Back translate a protein sequence
Input sequence: swissprot:opsd_human
Output sequence [opsd_human.fasta]:
%
```

第三章 自學 EMBOSS 套組的基本技能

如範例 3-4所示，執行「backtranseq」這個程式時，程式問使用者二個問題：第一個是要分析的序列；第二問題是輸出檔案的檔名。在 backtranseq 程式使用手冊中「命令列參數(Command line arguments)」副標題下，它簡略地指出有一個選擇性的參數，用來指定密碼使用頻率表(Codon usage table)，而預設的是人類的密碼使用頻率表。若想產生或修改密碼使用頻率表，可參考 cusp 程式的說明。可是對一個初學者而言，可能根本不知道密碼使用頻率表到底有什麼用。本書的另一個目的就是讓你在學習使用 EMBOSS 套組之外，能多熟悉這個領域的人所用的術語與思考方式。所以在方盒 3-2中介紹什麼是密碼使用頻率表，以便你能在第七章中應用它設計引子(primer)。

方盒 3-2 密碼使用頻率表

因為胺基酸的數目小於遺傳密碼的個數，有幾個遺傳密碼會對應到一個胺基酸，對同一個胺基酸而言，因為不同 tRNA 基因在不同生物的表現量不同所以喜歡使用的密碼並不完全相同。例如下表中顯示 Gly 有四個不同的密碼，大腸桿菌中大量表現的蛋白質使用 GGU 的頻率較高(59%)，而使用 GGA 的頻率卻很低(0%)。這些頻率表是根據 Genbank 中不同生物的序列而統計出來的，在 EMBOSS 環境下可用 embossdata 程式取得頻率使用表，以下是 Eeco_h.cut 檔案的一小部份：

Codon	AmAcid	Fraction	Number/1000	
GGA	G	0.020	1.390	118 ! 計算 Fraction 的方法:
GGG	G	0.040	3.140	267 ! Fraction = 1.39/(1.390+
GGC	G	0.430	35.110	2987 ! 3.140+35.110+
GGT	G	0.520	42.120	3583 ! 42.120)
-----				= 0.020
CCC	P	0.010	0.450	38

若與取自「<http://www.kazusa.or.jp/codon/>」的 *Xenopus laevis* 的頻率表比較，可注意到此表中所統計的密碼數目遠高於 GCG 提供的頻率表，其密碼安排的次序也不同，而且只列出出現的頻率而未加註使用的比例(Fraction)。在近期內，該網站將提供 GCG 格式的頻率表，屆時將可直接在 GCG 環境下使用各種不同生物的頻率表。使用網站上的資料的好處是它不斷地依 Genbank 現有的序列更新頻率表，因此所得的使用頻率比較正確。

Xenopus laevis [gbvrt]: 2108 CDS's (999077 codons) (節錄自網際網路上取得的頻率表)

fields: [triplet] [frequency: per thousand] ([number])

UUU 19.9(19900) UCU 19.3(19252) UAU 15.4(15376) UGU 11.0(11019)
---- ---- ---- ----
GUU 15.7(15698) GCU 20.8(20779) GAU 29.4(29400) GGU 13.2(13158)
GUC 12.0(11992) GCC 17.7(17695) GAC 22.9(22832) GGC 14.9(14871)
GUA 10.3(10328) GCA 19.9(19878) GAA 35.9(35873) GGA 21.4(21360)
GUG 21.1(21108) GCG 4.6(4614) GAG 33.9(33874) GGG 12.8(12783)

Coding GC 47.33% 1st letter GC 52.33% 2nd letter GC 40.92% 3rd letter GC 48.74%

在副標題「輸出檔案格式(Output file format)」中簡單解釋輸出檔案的特性。如何解讀此檔，這就需要參考「說明」。這又是一個初學者可能想不到之處，因為把解讀輸出檔案的方法寫在「說明」中，初學程式的人根本看不懂這「說明」；可是若先看「輸出檔案格式」，再看「說明」就恍然大悟了。事實上只要了解輔助系統的編輯原理，就知道該參照那一個副標題找到想要的資料，以後就可以舉一反三地應用到學習其他程式上。

對初學者而言，除了不會用輔助系統外，最困難的是不知什麼程式可以解決自己的問題。將程式分類固然有助於選擇程式，可是初學者通常不瞭解程式分類的原則，所以經常要花許多時間摸索。副標題「(See also)」會列出和這程式相似或相關的程式。例如和 backtranseq 相關的程式有 charge 與 checktrans 等等，初學者可透過這些程式了解 backtranseq 的應用。

現在應已知道如何瀏覽輔助系統。所以要試著去練習看輔助系統的內容與描述的方式。

練習 3-2 請查閱 backtranseq 中「數據檔案(Data Files)」這個副標題，以瞭解如何指定反轉譯所需的數據檔案。

Answer :

The codon usage table is read by default from "Ehum.cut" in the 'data/CODONS' directory of the EMBOSS distribution. If the name of a codon usage file is specified on the command line, then this file will first be searched for in the current directory and then in the 'data/CODONS' directory of the EMBOSS distribution.

To see the available EMBOSS codon usage files, run:

% embossdata -showall

To fetch one of the codon usage tables (for example 'Emus.cut') into your current directory for you to inspect or modify, run:

% embossdata -fetch -file Emus.cut

Not every file but most are described in the README file from <ftp://ftp.ebi.ac.uk/pub/> databases/codonusage

You can use the EMBOSS program 'cutgextract' on the CUTG database to get files with more meaningful (long) names.

在這練習中可發現程式的運作，要求密碼使用頻率表外，可用 Fetch 程式將其取回修改。只要此檔案檔名未變，而且存在於你的子目錄下，程式就會使用這修改過的檔案。如果你希望給這檔案一個新名字，例如「mycode.txt」，則必須利用指令行的選項來指定檔名。「backtranseq -help」出現在每一個程式的「Data files」這副標題下，若參照「命令列參數」(圖 3-1)即知「-cfile Codon_usage_table_name」的用法（在 -cfile 與 Codon_usage_table_name 間為空白字元）。

圖 3-1「命令列參數」顯示「-cfile Codon_usage_table_name」使用密碼使用頻率表的寫法

Command line arguments

Mandatory qualifiers:

[-sequence]	sequence	Sequence USA
[-outfile]	seqout	Output sequence USA

第三章 自學 EMBOSS 套組的基本技能

Optional qualifiers:		
-cfile	codon	Codon usage table name
Advanced qualifiers: (none)		
General qualifiers:		
-help	boolean	Report command line options. More Information on associated and general qualifiers can be found with -help -verbose

在此處的目的是在教怎麼交互參照線上輔助系統的不同部份，之後才真正地練習使用數據檔案。雖然使用互動式(interactive)的程式對初學者比較容易，可是對較熟練的人來說卻是一種折磨，因為必須一一回答這些問題。一種折衷的辦法是讓使用者能以指令的方式將程式所要的資訊寫在一個檔案中，或是寫在程式名稱之後，這些寫在指令行的資料能控制程式執行的方式，有助於將序列分析自動化。因為不同的程式需要提供不同的資料，所以使用一個不熟悉的程式時，可以參考「命令列參數」來學習怎樣將條件寫在指令行執行。

在使用程式時如果有不易解釋的結果應查閱「範例 3-3」。以反轉譯程式為例，若程式無指定密碼使用頻率表，則密碼使用頻率表會使用預設之人類密碼使用頻率表檔案 Ehum.cut。因此若你不指定密碼使用頻率表而來反轉譯其他物種之序列，將產生錯誤的結果。此外，在「說明」這副標題下會說明使用這程式該注意的事。

範例 3-5 使用backTranseq 程式的「說明」，以瞭解使用程式時的注意事項。

Description

backtranseq takes a protein sequence and makes a best estimate of the likely nucleic acid sequence it could have come from. It does this by using a codon frequency table. For each amino acid, the corresponding most frequently occurring codon is used in the construction of the nucleic acid sequence.

Codon usage table name

backtranseq reads in a data file containing the codon frequency tables. The default codon frequency table is 'Ehum.cut' - the human codon frequency table. It is important to use a codon frequency table that is appropriate for the species that your protein comes from. See the Data Files section below for more details on these files.

對初學者而言，這是一個很重要的綜論，敘述使用 EMBOSS 套組的一些原則性問題，即使不能應用到所有的程式上，也可以應用到某類的程式上。若無實做的經驗，讀起來很枯燥。可是若因實做而產生問題，再來查閱此部份，就有如魚得水的感覺。

二. 程式的執行

在上面的例子中，以 backtranseq 的例子學習輔助系統的使用方法，可是尚未練習怎樣執行一個程式。現在要利用所學的技巧來摸索另一個程式。請你自己查閱輔助系統，以完成下面的練習，並說明如何解讀。

練習 3-3 如何利用 showdb 程式，自 EMBOSS 套組中搜尋已安裝的資料庫。

Answer:

%**showdb**

```
Displays information on the currently available databases
# Name      Type ID  Qry All Comment
# ====
swissprot   P      OK   OK   Swissprot sequences
embl        N      OK   OK   EMBL sequences
```

三． 資料庫的簡介

在 EMBOSS 的環境下的資料庫主要是由 EMBOSS 套組安裝管理者所安裝設定。資料庫分為核酸、蛋白質與模組樣式(pattern)等三種資料庫，每一類的資料庫又可能有幾種。例如核酸資料庫有 Genbank (簡稱為 gb)、EMBL (簡稱為 EMBL)等兩種；蛋白質資料庫也有 Swiss-Prot (簡稱為 sw)與 Protein Information Resources (簡稱為 PIR)等兩種；模組樣式資料庫則有 ProSite、TRANSFAC、PRINTS 等三種。各種資料庫的資料格式稍有不同。在 EMBOSS 套組中提供一系列程式，可以將以上的資料庫編索引並納入 EMBOSS 環境中。

四． 資料庫的搜尋

搜尋資料庫有三種策略，最容易了解的是「字串搜尋(string search)」。在圖書館中常用的「Medline」系統就是利用這種方式及像「醫學主題(Medical Subject Headings, MESH)」這樣的控制語彙(control word)來做檢索。使用控制語彙，使我們不必在意作者使用單複數名詞、形容詞或同義字。可是序列資料庫的管理者沒有那麼多人力去編寫控制語彙，所以採用其他的運算法，以便於辨識模組樣式(pattern)或是同源(homologous)的序列。像 fuzznuc 這樣的程式，可以讓使用者輸入比較含糊的描述方式，例如 EcoRII 酵素的辨識序列「CCAGG」與「CCTGG」兩種序列可表示成「CCWGG」。可是有些時候使用者寫不出那麼明確的模組樣式，只能給一些條件來規定什麼叫做相像。「textsearch」是 EMBOSS 套組中搜尋資料庫序列名稱或說明的程式，但在它的程式使用手冊「功能」中，建議我們使用 SRS 或是 Entrez 會比較快。所以我們最方便快速的作法是由查詢 SRS 或是 Entrez，取得我們需要序列的 accession number，再在 EMBOSS 環境中指定需要的序列來進行分析較為方便。

在 EMBOSS 套組中欲取得序列進行分析，是以一套標準的命名方式稱為 USA(Uniform Sequence Addresses)，如下列範例 3-6。最常用的是 dbname:entry，dbname 指的是 EMBOSS 已安裝提供之資料庫名稱，請參照練習 3-3。

範例 3-6 USA(Uniform Sequence Addresses)表示法

The USA syntax is basically one of:

```
"file"
"file:entry"
"format::file"
```

```
"format::file:entry"
"dbname:entry"
"dbname"
"@listfile" (a file of file-names)
Example:
format::file
  embl::myfile.seq
  gcg::myresults.seq
dbname:entry
swissprot:tf3a_xenla
```

若要查詢特定 ID 或 accession number 來自那個資料庫可用「whichdb」程式。搜尋資料庫時，雖然沒有必要將序列取回自己的子目錄使用，在實用上卻可利用這個方法給序列一個好記的名字，範例 3-7 中以程式 seqret 來達到這目的。

範例 3-7 請取回*Xenopus laevis*的TFIIIA蛋白質序列，並將其命名為tf3a.pro

Answer:
%**seqret swissprot:tf3a_xenla**
Reads and writes (returns) sequences
Output sequence [tf3a_xenla.fasta]: tf3a.pro

五. 結語

在這一章中，首先以 backtranseq 為例，解釋如何使用線上輔助系統，再由讀者自己摸索出使用及解讀 showdb 程式的方法。在會用這種程式後介紹了各式資料庫，及如何取出資料庫中特定序列來進行後續分析。

參考網站

<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/>

第四章 SRS 資料庫查詢

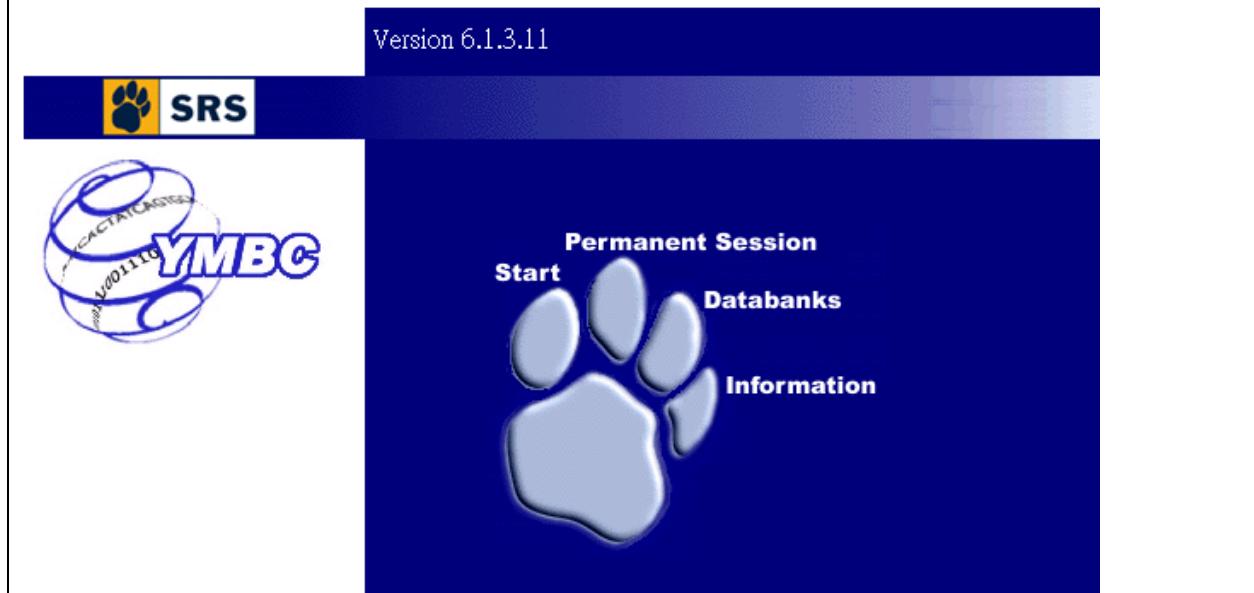
王聿泰¹、楊永正²

¹陽明生物化學所、²陽明大學生物資訊研究所

一. 簡單查詢

SRS 是一個整合、分析、與呈現生物資料庫資訊的工具。它可以整合將近 600 個生物資訊資料庫的搜索系統，它可以同時搜尋不同類別的資料庫，例如：序列、模組、轉錄因子....等，而每一類之下的資料庫又可複選，然後一起搜尋。其介面如下：

圖 4-1 The Start page



1. 按下 上的 Start，你將會到達 Top page(亦即 Library Select page) 而可以開始使用 SRS 來查詢。

圖 4-2 The Top page



2. 選擇你想搜尋的資料庫，例如：若想搜尋 embl 資料庫就在 embl 左邊的方框中打勾。

圖 4-3 選擇 embl 資料庫



3. 按下 **Standard**，此時會顯現標準查詢表格。

第四章 SRS 資料庫查詢

圖 4-4 embl 的標準查詢表格

4. 輸入查詢字串，例如：若想查詢 embl 中 Description 為 kinase，因此我利用下拉式選單將欄位選為 Description，並在方格內輸入 kinase。

圖 4-5 查詢特定欄位的資料

5. 按下 **Quick Search**，此時會顯現查詢到的結果，若你想看整筆資料，只要點超連結即可。

圖 4-6 查詢到的結果

Query "[embl-Description:kinase*]" found 67220 entries				
Perform operation		Description	SeqLength	
<input type="radio"/> on all but selected	<input type="checkbox"/> EMBL:AA283218	AA283218 RTH316 HTCDL1 Homo sapiens cDNA 5'/3' similar to Humna Kinase(TTK), mRNA sequence.	157	
<input type="radio"/> on selected	<input type="checkbox"/> EMBL:AA283466	AA283466 RTH244 HTCDL1 Homo sapiens cDNA 5'/3' similar to Protein Kinase, mRNA sequence.	234	
<input type="radio"/>	<input type="checkbox"/> EMBL:AA506772	AA506772 EST010 Human MCF7 cDNA subtracted with MDA-MB-231 cDNA Homo sapiens cDNA clone DEME-16 A similar to p55Phosphatidyl Inositol Kinase, mRNA sequence.	272	
<input type="radio"/>	<input type="checkbox"/> EMBL:AA506773	AA506773 EST011 Human MCF7 cDNA subtracted with MDA-MB-231 cDNA Homo sapiens cDNA clone DEME-16 B similar to p55Phosphatidyl Inositol Kinase, mRNA sequence.	265	
<input type="radio"/>	<input type="checkbox"/> EMBL:AA543024	AA543024 ni55h08.s1 NCL_CGAP_Ov2 Homo sapiens cDNA clone IMAGE:980799 similar to gb:M86699 DUAL SPECIFICITY PROTEIN KINASE TTK (HUMAN);, mRNA sequence.	232	
<input type="radio"/>	<input type="checkbox"/> EMBL:AA543099	AA543099 nf96h07.s1 NCL_CGAP_Co3 Homo sapiens cDNA clone IMAGE:927805 3' similar to gb:X15334_m1 CREATINE KINASE, B CHAIN (HUMAN);, mRNA sequence.	526	

二. 連結其他資料庫

6. 若想知道查詢到的 embl 結果中有哪些資料可以連結到 swissprot 資料庫，只要按下 **Link**，就會顯現 LINK page。

圖 4-7 LINK page

The screenshot shows the SRS interface with the following details:

- Top Navigation Bar:** TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS, DATABASES, HELP.
- Search Query:** Current query: "[embl-Description:kinase*]"
- Set Db Options:**
 - Find all Entries:** Radio button selected.
 - in the selected databanks which are linked to the current query
 - in the current query which are linked to all selected databanks
 - in the current query which are not linked to any of the selected databanks
- Buttons:** Submit Link, show all, collapse all.
- Sequence libraries - complete:**
 - To Parent Library
 - Sequence libraries - complete
- Databank Selection:**
 - all
 - EMBL
 - TREMBL
 - SWALL
 - SWISSPROT
 - ENSEMBL_MOUSE
- Page Configuration:** Number of entries to display per page: 30.

7. 選擇「SetDb」的三個選項中的第一項，表示要將圖 4-7 中的各筆資料連接到圖 4-8 中所選擇的資料庫。在這個範例中若要連結 swissprot 資料庫，則在 swissprot 左邊方格中打勾。

圖 4-8 選擇連結的資料庫

The screenshot shows the SRS interface with the following details:

- Top Navigation Bar:** TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS, DATABASES, HELP.
- Search Query:** Current query: "[embl-Description:kinase*]"
- Set Db Options:**
 - Find all Entries:** Radio button selected.
 - in the selected databanks which are linked to the current query
 - in the current query which are linked to all selected databanks
 - in the current query which are not linked to any of the selected databanks
- Buttons:** Submit Link, show all, collapse all.
- Sequence libraries - complete:**
 - To Parent Library
 - Sequence libraries - complete
- Databank Selection:**
 - all
 - EMBL
 - TREMBL
 - SWALL
 - SWISSPROT
 - ENSEMBL_MOUSE
- Page Configuration:** Number of entries to display per page: 30.

8. 按下 **Submit Link**，會得到所有與你剛查詢的 embl 結果中與 swissprot 相關的結果。

圖 4-9 顯示查詢的結果

The screenshot shows a search results page from the SRS system. The query was "([embl-Description:kinase*] > SWISSPROT)". The results table has columns: SWISSPROT, Accession, Description, and SeqLength. There are 2326 entries found. The first few entries are:

SWISSPROT	Accession	Description	SeqLength
SWISSPROT:1433_OENHQ	P29307	14-3-3-like protein.	260
SWISSPROT:1433_SPIOL	P29308	14-3-3-like protein (Fragment).	220
SWISSPROT:AAIP_WHEAT	Q02066	Abcisic acid-inducible protein kinase (EC 2.7.1.-) (Fragment).	332
SWISSPROT:AAK2_PIG	Q28948	5'-AMP-activated protein kinase, catalytic alpha-2 chain (EC 2.7.1.-) (AMPK alpha-2 chain) (Fragment).	129
SWISSPROT:AAK2_RAT	Q09137	5'-AMP-activated protein kinase, catalytic alpha-2 chain (EC 2.7.1.-) (AMPK alpha-2 chain).	552
SWISSPROT:AAKC_RAT	Q9QZH4	5'-AMP-activated protein kinase, beta-2 subunit (AMPK beta-2 chain).	271
SWISSPROT:AAKG_BOVIN	P58108	5'-AMP-activated protein kinase, gamma-1 subunit (AMPK gamma-1 chain) (AMPK _g).	330
SWISSPROT:AAKG_RAT	P80385	5'-AMP-activated protein kinase, gamma-1 subunit (AMPK gamma-1 chain) (AMPK _g).	330
SWISSPROT:AAKL_HUMAN	Q9UGI9	5'-AMP-activated protein kinase, gamma-3 subunit (AMPK gamma-3 chain) (AMPK gamma3).	464

三. View 的設定

9. 在圖 4-10 中所呈現的表格是根據一個預設的「SeqSimple View」設定的。可是使用者也可以自己決定表格的呈現方式，這項功能可以從“View”欄來設定。

The screenshot shows the 'Create New View' configuration page. It includes fields for 'View name:' (with a yellow highlight), 'Select fields from' (with radio buttons for 'all fields in libraries' and 'just common fields'), and two lists of databases for defining the view and linking it to displayed entries. The 'Views' tab is selected in the top navigation bar.

8. 在 “View name” 下方的文字方塊輸入自己的名字，例如 “myTestView” 。
9. 在 “Select databanks to define a view for” 中點選 “EMBL” 。
10. 在 “Select databanks to be linked to displayed entry” 中點選 “SWISSPROT” 。
11. 點選 “Create New View” 按鈕。

此時進入資料呈現管理頁第二頁中選擇你想看的資料欄位。

圖 4-11 勾選資料庫的資料欄

The screenshot shows the 'Create New View' page. At the top, there's a navigation bar with links: TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS (which is highlighted in red), DATABANKS, and HELP. Below the navigation bar, there are two main sections for defining the view:

- EMBL**: A list of fields with checkboxes:
 - ID
 - Division
 - AccNumber
 - SeqVersion
 - Molecule
 - Description
 - Keywords
 - Organism
 - Taxon
 - Organelle
 - Comment
 - DateCreated
 - LastUpdated
 - SeqLength
 - Link
 - Sequence (with a dropdown menu showing 'fasta')
- link to SWISSPROT**: A list of fields with checkboxes:
 - ID
 - AccNumber
 - Description
 - GeneName
 - Keywords
 - DateCreated
 - LastUpdated
 - Organism
 - Taxon
 - NCBI_TaxId
 - Organelle
 - ProteinID

12. 選擇資料欄位，在兩個資料庫當中均可點選 “Description” 。
13. 在點選 “Save View” 鍵，接下來可以用我們剛編輯好的 “View” 來呈現資料。
14. 點選 “Result”，將會出現查詢管理網頁。

圖 4-12 利用自定選單觀看查詢結果

The screenshot shows the SRS software interface with the following details:

- Top Navigation Bar:** Includes links for TOP PAGE, QUERY, RESULTS (highlighted in blue), PROJECTS, VIEWS, and DATA.
- Left Sidebar:** Contains buttons for Reset, Save, Delete, Link, and View. A dropdown menu under View is open, showing options like * Complete entries *, default view, * Names only *, * Complete entries *, SeqSimpleView, FastaSeqs, EMBL, and myTestView. The option "myTestView" is highlighted with a green background.
- Center Content Area:** Titled "Successful Queries". It displays a table with the following data:

Name	Type	N Total	From Library	N
Q1	query	2405069	EMBL	2405069

15. 點選確認盒以勾選擬要看的查詢(例如 Q1, Q2)。
16. 在下拉式選單當中勾選擬剛剛我們所儲存的“myTestView”。
17. 再點選“View”鍵。

圖 4-13 在自定表單下呈現的資料

The screenshot shows a search results page with the following details:

- Top Navigation:** TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS, DATABASES, HELP.
- Search Query:** Query "[embl-AllText:cancer*]" found 2405069 entries.
- Perform operation:**
 - Radio buttons: "on all but selected" (selected) and "on selected".
 - Buttons: Link, Save, View.
 - Dropdown: myTestView.
- Number of entries to display per page:** 30.
- Printer Friendly:**
- Table Headers:** EMBL, ID, Description, SWISSPROT, Description.
- Table Data:** A list of 6 EMBL entries, each with a checkbox, ID, and a detailed description. The descriptions mention various genes and mRNA sequences related to cancer.

18. 回到查詢管理頁，此時可以點選“Result”欄。
19. 點選查詢 Q2。
20. 再點選“View”鍵。
21. 在此後我們將可以利用系統所提供的生物資訊分析程式來分析。

圖 4-14 勾選一個分析工具

The screenshot shows a search results page with the following details:

- Top Navigation:** TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS, DATABASES, HELP.
- Search Query:** Query "[libs=(embl swall) -AllText:cancer*]" found 2449831 entries.
- Perform operation:**
 - Radio buttons: "on all but selected" (selected) and "on selected".
 - Buttons: Link, Save, View.
 - Dropdown: SeqSimpleView.
 - Button: Launch.
 - Dropdown: BlastN (selected), with options: BlastN, NFastA, FastX, FastY, NCustalW, Restrictionmap, BlastP, FastA, ClustalW, HHMPfam, PPSEARCH.
- Table Headers:** EMBL SWALL (SPTR), Accession.
- Table Data:** A list of 13 EMBL entries, each with a checkbox, Accession number, and a detailed description. The descriptions mention various genes and mRNA sequences, with some entries specifically mentioning BlastN analysis.

22. 在“Launch”下方有一下拉式選單選擇想要用的分析工具

圖 4-15 可使用預設值送交執行

SRS 6.1.3.11 | [feedback](#)

The screenshot shows the BlastP search interface. At the top, there are fields for 'Name of job:' (temp) and 'Database to search:' (SWISSPROT). Below this is the 'Launch' configuration panel. It contains several sections:

- Note:** This application is executed by LSF batch queueing system. Name of the queue is extsrv_interactive -R blast -L /bin/sh (batch).
- Output Options:** Number of alignments to show: 250; Number of best hits from a region to keep: 100; Number of one-line descriptions: 500.
- Search Parameters:**
 - Filter query sequence:
 - Scoring matrix: BLOSUM62
 - The E value: 10.000000
 - word size: Default
 - Perform gapped alignment:
 - Cost to open a gap: Default
 - Cost to extend a gap: Default

At the bottom left of the configuration panel are buttons for 'Reset' and 'Launch'.

23. 點選“Launch”將會進入應用程式參數頁，如果第一次使用，可以試著使用參數的內定值
 24. 點選“Launch”鍵，系統將會接受指令執行程式

圖 4-16 執行中的工作狀態

The screenshot shows the execution status page. At the top, there is a navigation bar with links: Top Page, Query, Results, Projects, Views, Databanks. Below the navigation bar, a message states: "Application was submitted to Queue: extsrv_interactive -R blast -L /bin/sh(batch)." Underneath this, it says "Application command:" followed by a code block containing the command: "/ebi/extserv/bin/ncbi-blast/blastall -p blastn -d \$IDATA_CURRENT/blastdb/embl -". At the bottom of the page, there is a link: "Use Batch job status page to view the results".

第四章 SRS 資料庫查詢

25. SRS 也提供多種格式可以看結果，在“View”按鍵下的下拉式選單選擇模式。

圖 4-17 執行結果

The screenshot shows the SRS search results for a BLASTN query. The top bar displays "Query [BlastN-JobName.temp_job1]" found 250 entries and a "next" button. On the left, there's a "Perform operation" section with radio buttons for "on all but selected" or "on selected", and buttons for "Link", "Save", and "View". Below that is a dropdown menu set to "BlastN" and a "Launch" button. A "Printer Friendly" link is also present. The main area shows the first few hits. The first hit is highlighted with a checkbox and the identifier >EM_INV:AC006681 AC006681.1 Caenorhabditis elegans cosmid R13H9, complete sequence. It includes statistics: Score = 2.885e+04 bits (14551), Expect = 0.0, Identities = 14551/14551 (100%), and Strand = Plus / Plus. Below this, sequence alignments for Query 1, Sbjct: 1, Query 61, Sbjct: 61, and Query 121, Sbjct: 121 are shown with their respective sequence lines.

圖 4-18 存檔格式設定

This screenshot shows the same SRS search results interface as Figure 4-17, but with a different "View" selection. The "View" dropdown is now set to "Blast_View", which changes the presentation of the results into a tabular format. The table has columns for BLASTIN (link to query page), Query (link to query page), Search Database (link to database page), Hit (link to hit page), Description (hit details), Top Score (top score), E value (E-value), and Percentage Identity (percentage identity). The hits listed are the same as in Figure 4-17, with the first hit >EM_INV:AC006681 being the top result.

BLASTIN	Query	Search Database	Hit	Description	Top Score	E value	Percentage Identity
BLASTN:temp_job1.embl 1 AC006681			>EM_INV:AC006681	Caenorhabditis elegans cosmid R13H9, complete sequence.			100
BLASTN:temp_job1.embl 2 AF078790			>EM_INV:AF078790	Caenorhabditis elegans cosmid F36H12, complete sequence.			99
BLASTN:temp_job1.embl 3 CEC09B9			>EM_INV:CEC09B9				100
BLASTN:temp_job1.embl 4 AF038605			>EM_INV:AF038605	Caenorhabditis elegans cosmid C02B10, complete sequence.			100
BLASTN:temp_job1.embl 5 CEUZK354			>EM_INV:CEUZK354				97
BLASTN:temp_job1.embl 6 AC024839			>EM_INV:AC024839	Caenorhabditis elegans cosmid Y59E9AR, complete			95

26. 點選“Save”進入下載結果網頁，可以選擇所下載的資料格式也可以選擇純文字檔或網頁文件格式亦或是 pdb 格式

圖 4-19 存檔格式設定

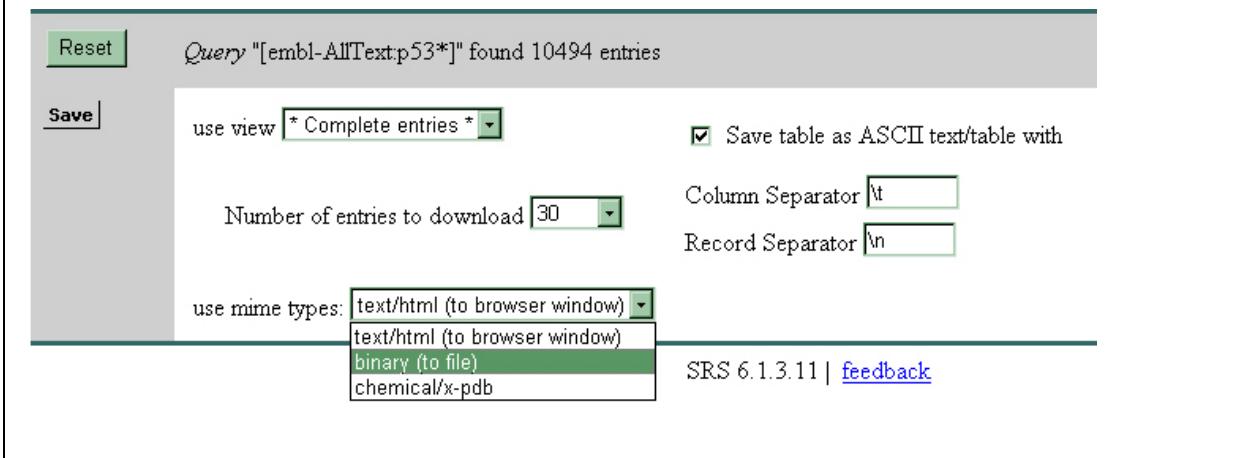
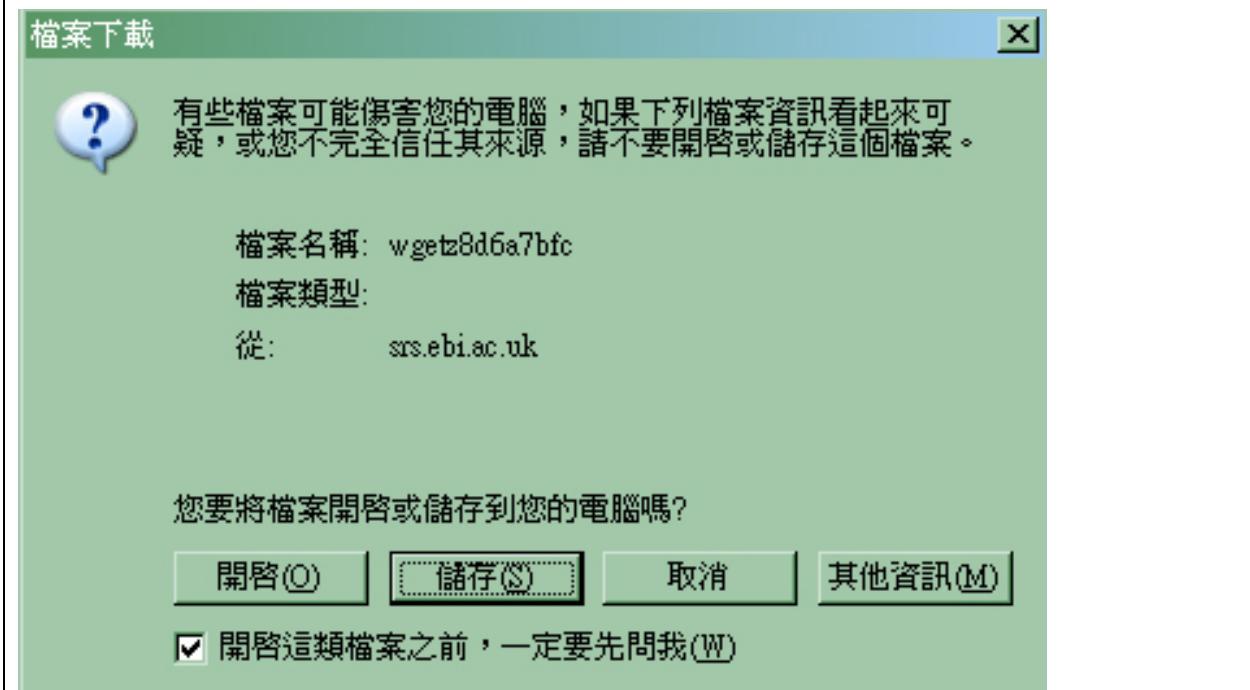


圖 4-20 下載視窗



27. 在點選“Save”後，就會出現下載對話視窗，將檔案放置在你想放的位置

四. SRS 分案查詢

SRS 可以允許你將工作類型或目標一致的工作建立一個專案，並且將查詢結果，之前設定的資料呈現方式等存在一起，做日後的追蹤或者延續之前的工作，這個章節將會讓你學下列事項：

- 使用暫存專案與永久專案的好處
- 如何開啟一個專案或回到既有專案中
- 永久專案中所可以使用的內容
- 如何使用暫存專案並將其資料與其他專案使用
- 如何將暫存專案的工作移至永久專案當中

1. 專案介紹

在 SRS 中有兩種專案可以使用，以下將詳述

(a) 暫時性的專案

當你使用暫存性的專案時，你的查詢及資料呈現的格式將存在系統暫時的位置，雖然在你結束查詢之後，他會存在一陣子，但是你不應該依賴這種方式，儘管你在你的網頁伺服器當中留下書籤做紀錄，但是這些資料也只能使用到系統管理員清除暫存區資料以前。

暫存區的專案設計只是為了做簡單的查詢，或只是想很快的看一下結果。我們稍後會再詳述暫存性專案。

(b) 永久性專案

當你使用永久性專案，你所做的查詢及資料呈現格式將會紀錄在專案裡面，這意味著這些資料在未來將永遠存在系統裡面，在這項專案中系統還提供密碼保護，允許你做限制項的存取。

如果你有以下需求，你應該選擇永久性的專案來做查詢

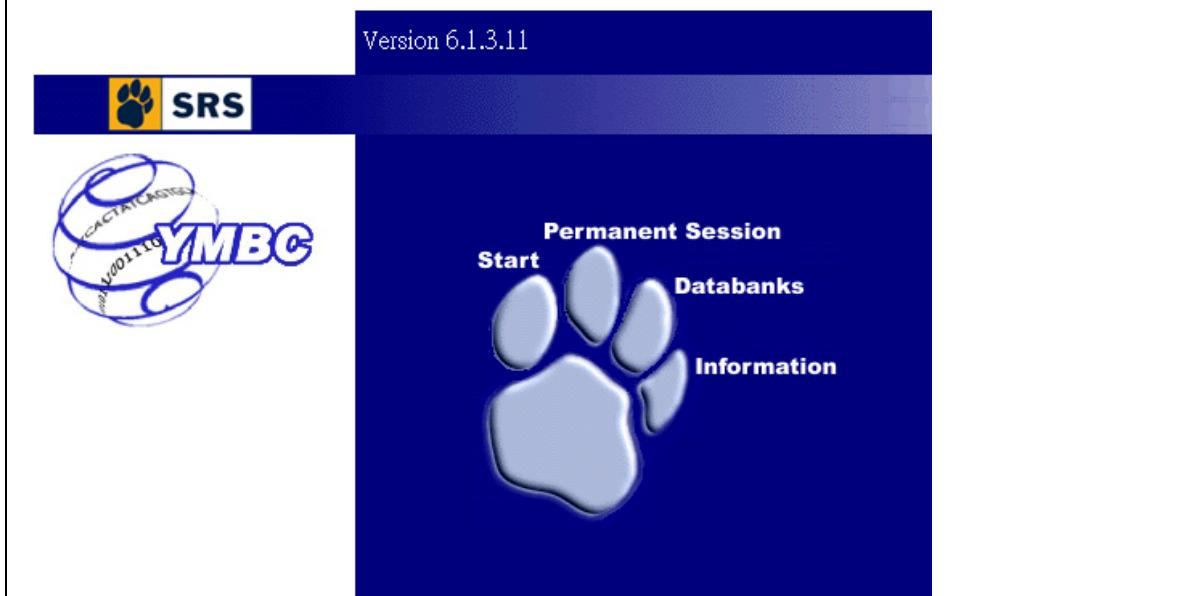
- 你和你的工作夥伴稍後想繼續工作
- 你希望能將一個專案當中的工作轉移到另一個專案當中
- 你想你的專案工作能受到保護
- 你想透過上傳的方式回復到之前的專案中繼續工作

我們稍後會詳述永久性專案。

(c) 開啟專案

不論你決定做暫時性的專案或者永久性的專案都是從開始頁開始。

圖 4-21開始頁面



在開始頁，有四個超連結位置，分別是

- Start
起始一個新的暫存專案做查詢
- Permanent Session
開始進入永久性專案查詢或回到之前的永久性專案查詢
- Databanks
你可以檢視系統中資料庫的狀態
- Information
在這裡，你可以查詢所有的文件，包含網頁版或 PDF 版的手冊均在此，其中還包含系統及 Icarus 語言手冊及 SRS 常問問題及回答

2. 暫時性的專案

暫時性的專案只適用於暫時性的查詢，其查詢紀錄會隨著系統例行性的清除動作而消失。

(a) 開啟一個暫時性的專案

暫時性的專案是從開始頁點選開始後進入
進入後就進入起始頁，開始暫時性專案工作
注意：你的暫時性工作也可以從 Databanks 進入

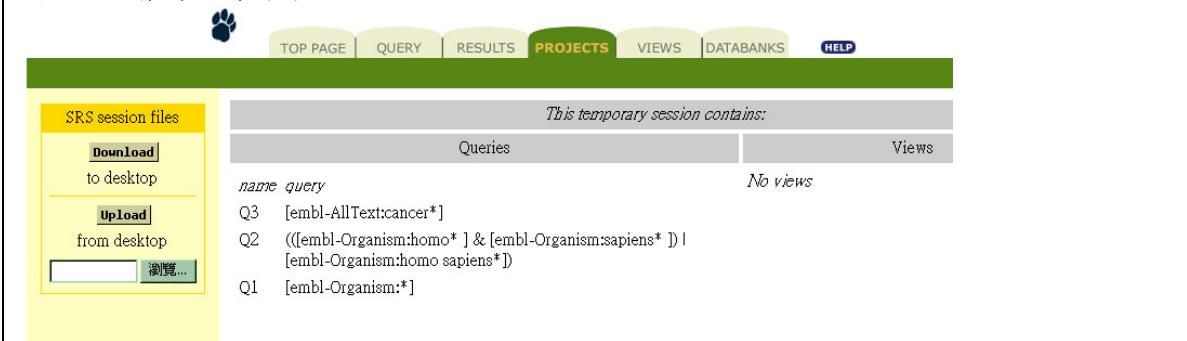
圖 4-22 資料庫選擇頁



(b) 管理暫時性專案資料

點選 Projects 欄，可以看到所有工作，進行管理

圖 4-23 暫時性專案管理



專案管理頁當中提供必要的管理工具包含已有的專案，在這裡你可以上傳和下載專案，當你的工作完成之後，可以下載至磁碟機中，除此之外還可以利用這項功能將專案工作轉移至其他專案，但是他不像永久性專案可以改變名稱，刪除，切換暫時性專案也不可以在暫時性專案中彼此分享查詢，及資料呈現的方式。

稍後我們將繼續介紹專案管理頁。

3. 永久性專案

永久性專案有兩種，一為安全性永久性專案，一為非安全性永久性專案

- 安全性永久性專案視需要透過密碼來授權使用，
- 非安全性永久性專案則是透過帳號來存取非安全性專案，也就是說只要知道帳號名稱，任何人都可以進去存取或修改資料內容

永久性專案是不管你是何種使用者，他將會儲存你的所有工作在同一個位置，你可以任意將他再呼叫出來，你也可以將工作在兩個專案間彼此交換資料。

注意：請洽詢你的管理者系統的政策，也許你需要管理者人工的方式幫你開帳號。

(a) 開啟一個永久性專案

開始一個新的永久性專案或回到原來的永久性專案中

- 在開始頁點選 Permanent Session 連到永久性專案區，系統將會回應出現登入對話。

注意：此出對話視窗圖片是擷取自美國微軟公司網際探索家第五版，其視窗樣式會隨著使用不同的瀏覽器及其版本而有所不同

圖 4-24 SRS 安全型登入視窗



圖 4-25 SRS 非安全型登入視窗

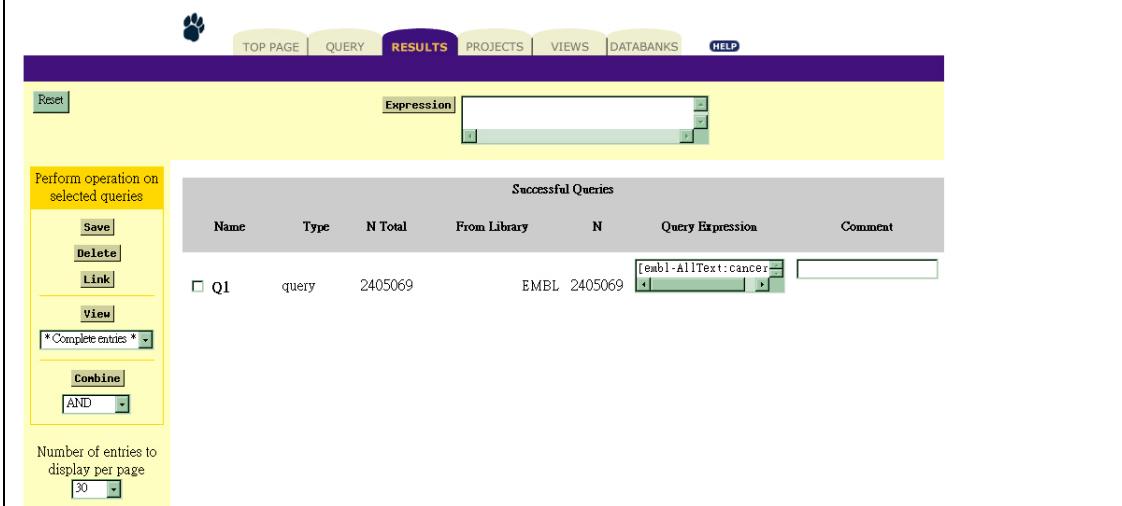


- 在安全型永久性專案登入時所使用的帳號密碼必須探詢系統管理者，其號密碼與你所使用的電腦中的帳號密碼未必相同
- 點選“OK”

(b) 永久性專案的管理

專案管理頁提供必要的工具管理你的工作、查詢結果、及資料呈現的方式，除此之外你還可以刪除系統中或將系統中的資料儲存到你的電腦上，也可以將資料上傳至系統中，請至開始頁點選永久性專案連結。

圖 4-26 永久性專案管理頁



4. 專案管理

已存在的專案

在專案管理頁中允許你在兩個專案間轉移資料

回到已經存在的專案中繼續工作

1. 點選在“Change to”按鈕下的下拉式選單
2. 選擇你想恢復的專案
3. 按下 按鈕

你所點選的專案將會出現在你的專案管理頁中，你可以繼續工作

注意：這個頁面在暫存性專案中是看不到的

開始一個新的專案

1. 點選“New Project”

這將會開始一個新的專案，並且在開始頁會顯示出來

注意：這個在暫存性的專案中並不會出現

(a) 專案管理操作

有四種方式可以達到專案管理操作

- Rename
更改專案名字
 - Delete
移除專案
 - Upload
上傳一個已經存好的專案
 - Download
下載一個專案
- 它們的功能將在下面詳述

注意：在暫存性專案中更改名字及刪除

Rename

內建專案名稱 (project1, project2, projectN) 能夠個人化並可以追蹤
執行專案工作時可以達到有效率的管理

更改專案的名稱

1. 輸入新的名稱在標示 “This is ProjectN” 左側文字方塊中
2. 點選 “Rename” 按鈕

圖 4-27 個人化你的專案名稱



注意：在暫存性專案中並無此項功能

Delete

當你不再需要一個專案時，你可以將他刪除，這項功能是必要的，每個帳號中所能記載的專案數最多為 99 個

1. 你必須要將專案調出來呈現在網頁上才能按刪除鍵
2. 點選 Delete 鍵

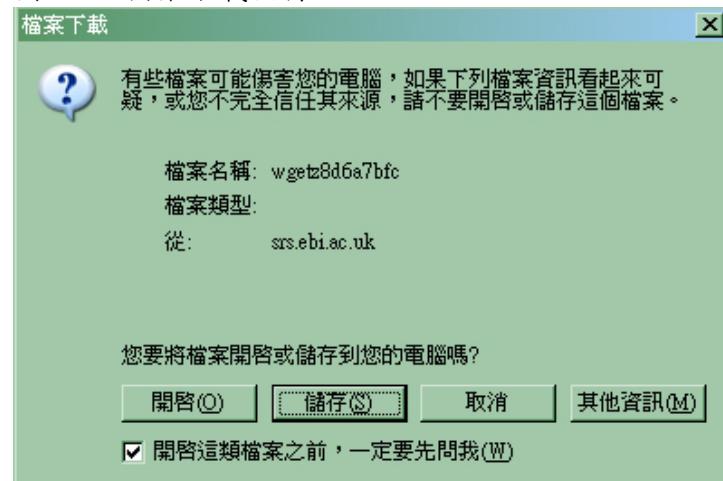
注意：在暫存性專案中並無此項功能

Download

如果你想跟別人分享一個專案，或者你想移到暫時性專案做處理，你必須先將專案下載到你的電腦上

1. 將想下載的專案呼叫出來
2. 點選 “Download” 按鈕
3. 點選儲存到電腦

圖 4-28 檔案下載視窗



4. 點選儲存檔案
5. 請輸入一個檔案名稱

圖 4-29 儲存對話視窗



6. 請點選儲存
此時下載的檔案就可以再度上傳並且上傳後繼續工作
- 上傳
在專案管理頁，你可以上傳已經存檔的專案繼續工作

從本機硬碟上傳專案

1. 在“Upload”鍵旁邊文字方塊，輸入想要上傳的檔案名稱，
2. 按下“Upload”

你也可以透過瀏覽的方式上傳

- 按下“Browse...”
- 上傳對話視窗就會出現，可以利用它找尋想要上傳的檔案

圖 4-30 瀏覽專案檔案



3. 點選“開啟”

這將會將檔案包含存在的路徑名稱自動放進文字方塊當中

4. 點選“Upload”

這將會把檔案放進系統中並且開始工作

五. SRS 資料庫查尋

使用 SRS 的最大優點是其搜尋資料庫的能力，而此單元是讓你知道有哪些查詢 SRS 的方式，在此章你將學習到：

- 何謂搜尋名稱(Search Term)
- 如何執行快速查詢(Quick Search)
- 如何使用標準搜尋表格(Standard Query Form)及延伸搜尋表格(Extended Query Form)
- 如何使用表示法來搜尋資料庫(Expression Query)
- 如何瀏覽索引

1. 搜尋名稱 (Search Terms)

不論你利用何種查詢方式，你都要使用一些搜尋名稱，而 SRS 使用的搜尋名稱可以分成五大類：

- 單一字搜尋(Single-word search)
- 多字句(Multiple-word phrases)
- 數字和日期(Numbers and dates)
- 正規表示法(Regular expressions)
- 萬用字元(Wildcards)

(a) 單一字(Single Words)

當你使用單一字於單一個欄位，例如：還原酵素(reductase)來搜尋資料庫時，你所得到的結果會是在這個欄位中包含這個字的一串項目。

(b) 多字句(Multiple-word phrases)

你可以使用多於一個字的詞句，例如：醛還原酵素(aldehyde reductase)來搜尋資料庫。若你將這個詞句用引號括起來，例如：“aldehyde reductase”，你所搜尋到的詞句是要跟 aldehyde reductase 完全吻合才會顯示其結果；若這個詞句不用引號括起來，SRS 會將詞句拆開後將個別的字分開查詢([aldehyde* & reductase*] |[aldehyde reductase])再將結果合併在一起表示；你也可以利用運算子在字串中來顯示個個字間的關係，此類的運算子有“和(&)”、“或(|)”及“但不是(!)”

(c) 數字和日期(Numbers and dates)

SRS 利用數值項目讓使用者可以搜尋日期或者是特定長度的項目(例如：序列長度)，而數值項目可以利用不同的算符再跟表示法合併來搜尋。這些算符包括：小於、小於或等於、大於及大於或等於，我們將這些算符和另兩種算符(冒號(:)及驚嘆號(!))合併在一起。我們來看幾個例子：

12:15 表示其大於或等於 12 但是小於或等於 15

12: 表示其大於或等於 12

!12: 表示其大於但不等於 12

:12 表示其小於或等於 12

:!12 表示其小於但不等於 12

因此冒號視其數字的位置可代表大於或小於，若冒號在數字右邊代表大於，若冒號在數字左邊代表小於；而驚嘆號(!)可表示為不(not)或不等於(not equal to)

(d) 正規表示法(Regular expressions)

當你想要搜尋有不同拼法或有相同字根但字尾不同的字時，利用正規表示法來搜尋資料庫是很有用的。你可以合併一些字元和正規表示字元來達到你要的目的，在正規表示字串的前後要加上反斜線(/)，例如：你可以利用/^phos/來找到所有字首為 phos 的字(例如：phosphate，phosphorylase)，而/ase\$/ 會找到字尾為 ase 的字(例如：kinase，phosphate)

(e) 萬用字元(Wildcards)

SRS 查詢語言包含了"*"和"?"的萬用字元，例如："cell*ase"會找到字首為"cell"及字尾為"ase"的字(例如：cellobiase，cellulase)

2. 簡單的資料庫搜尋

(a) 快速查詢(Quick Search)

快速查詢讓你可以利用最少的步驟從選擇資料庫到觀察結果。

1. 先到 TOP PAGE

圖 4-31 開始頁面

TOP PAGE QUERY RESULTS PROJECTS VIEWS DATABANKS HELP

Reset Quick Search All Entries

Query forms

Standard Extended

show all collapse all

Sequence libraries - complete

all EMBL TREMBL SWALL SWISSPROT
 ENSEMBL_MOUSE

2. 選擇你要的資料庫，在你所要選的資料庫左邊打勾

圖 4-32 選擇資料庫

TOP PAGE QUERY RESULTS PROJECTS VIEWS DATABANKS HELP

Reset Quick Search All Entries

Query forms

Standard Extended

show all collapse all

Sequence libraries - complete

all EMBL TREMBL SWALL SWISSPROT
 ENSEMBL_MOUSE

3. 在 Quick Search 旁的方框中填入你想查詢的字(例如：death)

圖 4-33 填入關鍵字

The screenshot shows the SRS TOP PAGE interface. At the top, there is a navigation bar with links for TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS, DATABANKS, and HELP. Below the navigation bar, there is a search bar labeled "Quick Search" containing the word "death". To the right of the search bar is a button labeled "All Entries". On the left, there is a section titled "Query forms" with two options: "Standard" and "Extended". Underneath this, there is a section titled "Sequence libraries - complete" with several checkboxes. The checkboxes for "EMBL", "TREMBL", "SWALL", and "SWISSPROT" are checked, while "all", "ENSEMBL", and "MOUSE" are not checked. There are also "show all" and "collapse all" buttons.

4. 按下 **Quick Search**

(b) 查詢表格(Query Form)

在查詢資料庫時，我們可以加入一些資訊來限制獲得的筆數，而獲得較準確的結果。而 SRS 提供兩種查詢表格，標準查詢表格(Standard Query Form)和延伸查詢表格(Extended Query Form)

(c) 利用標準查詢表格

標準查詢方式最多可以輸入四個查詢字串並且最多可同時查詢四個欄位

1. 先到 TOP PAGE 並選擇你要的資料庫，在你所要選的資料庫左邊打勾

圖 4-34 勾選查詢資料庫

This screenshot is identical to Figure 4-33, showing the SRS TOP PAGE with the "Standard" query form selected. The "Sequence libraries - complete" section shows checkboxes for EMBL, TREMBL, SWALL, and SWISSPROT, all of which are checked. The "all", ENSEMBL, and MOUSE checkboxes are unselected. The "show all" and "collapse all" buttons are also present.

2. 按下 Query forms 下的 **Standard**，則標準查詢表格將會顯示出來

圖 4-35 標準查詢表格

The screenshot shows the SRS (Swiss-Prot) search interface. At the top, there's a navigation bar with links for TOP PAGE, QUERY (which is highlighted in green), RESULTS, PROJECTS, VIEWS, DATABANKS, and HELP. Below the navigation bar, there's a search bar with the text "search SWISSPROT". To the right of the search bar is an "Info" button and a dropdown menu set to "about field AllText".

On the left side, there's a yellow sidebar with several configuration options:

- "Submit Query" button.
- "append wildcards to words" checkbox (checked).
- "combine searches with" dropdown menu set to "AND".
- "Number of entries to display per page" dropdown menu set to "30".
- "Extended query form" button.

The main search area contains four input fields, each with a dropdown menu set to "AllText". Below these fields is a dropdown menu for "retrieve entries of type" set to "Entry".

Underneath the search area, there's a section for "Create your own view" with a dropdown menu set to "SeqSimpleView".

At the bottom, there's a list of fields to select for display, with "ID" checked, and a dropdown menu for "sequence format" set to "swiss".

3. 先選欄位名稱再填入搜尋字串，可以利用 combine searches with 來設定在不同欄位間所需要的結合方式(例如：AND，OR 或 BUTNOT)

圖 4-36 組合查詢方式

The screenshot shows the SWISSPROT search interface. At the top, there is a navigation bar with links for TOP PAGE, QUERY (which is highlighted in green), RESULTS, PROJECTS, VIEWS, and DATABANKS, along with a HELP link. Below the navigation bar, there is a search bar with the text "search SWISSPROT" and a "Reset" button. To the right of the search bar is an "Info" field set to "about field AllText". On the left side, there is a sidebar with several options: "append wildcards to words" (with a checked checkbox), "combine searches with AND" (with a dropdown menu showing "AND"), and "Number of entries to display per page" (set to 30). Below these is a button labeled "Extended query form". The main search area contains several search fields: "Description" (set to "death domain"), "DateCreated" (set to "16-OCT-2001"), and two "AllText" fields. Below these fields is a button to "retrieve entries of type Entry". There is also a section titled "Use view" with a dropdown menu set to "SeqSimpleView". Underneath this, there is a "Create your own view" section with a "Select fields to display:" list containing "ID", "AccNumber", "Description", "GeneName", "Keywords", "DateCreated", and "LastUpdated". To the right of this list is a "sequence format" dropdown menu set to "swiss".

4. 按下 **Submit Query**

(d) 進階查詢模式

進階查詢模式可以列出資料庫中所有的資料項目並且可以依照所知道的資料，作多項資料欄位聯合查詢。

圖 4-37 進階查詢模式

1. 從最前面一頁(Top page)，選擇所想要查詢的資料庫，在資料庫前方方塊中點選。
2. 從左方 Query forms 點選“Extended”按鈕，將會進入進階查詢模式。
3. 進入進階查詢模式之後，將可看到一連串欄位，雖然你不需要填寫所有欄位，但是如果提供的越詳盡，則越可能查到你要的資料。
4. 你可以選擇聯合查詢的模式，如交集(AND)，聯集(OR)或排除可能性(BUTNOT)
5. 指定在資料呈現時，每一頁所呈現的資料筆數。
6. 如果想看精簡或詳細資料，可以使用“view”功能所提供的下拉式選單，選擇呈現的模式。
7. 按下“Submit Query”按鈕。

圖 4-38 進階查詢模式

在 EMBL 資料庫進階模式查詢中，以 Description，Keyword，Organism 和 DataCreated 欄位中下達 histamine，histamine，human 以及限制必須在 Jan. 1st 2002 以後所記載的資料。

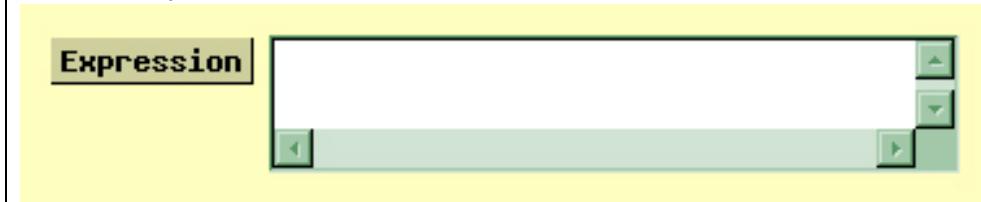
Description	histamine
Keywords	histamine
Organism	human
Taxon	
Organelle	
Comment	
DateCreated	<input type="button" value="≥"/> <input type="text" value="1-Jan-2000"/> <input type="button" value="≤"/> <input type="text"/>

3. 直接下指令方式查詢

(a) 關於指令查詢

你可以在 Query Manager 頁面中的文字方塊中下指令，並按下“Expression”進行查詢，你還可以結合已做過的查詢紀錄中以不同的關鍵字作再查詢

圖 4-39 Expression 按鍵與文字方塊



(b) 如何使用指令查詢

1. 在 Query Manager 頁中，文字方塊中鍵入你的查詢指令
2. 按 Expression 按鈕

所舉的例子，在查詢結果 Q1 及 Q2 中同時存在的資料，你可以用下列指令輸入

圖 4-40 輸入 Q1&Q2

The screenshot shows a yellow header bar with the text 'Expression' in bold. Below it is a white input field containing the text 'Q1 & Q2'. To the right of the input field are four small green square buttons with arrows pointing up, down, left, and right respectively. At the bottom of the input field is a green horizontal scroll bar with a small black arrow pointing to the right.

此時會產生 Q3 其中的資料就是同時存在於 Q1 及 Q2 的資料

圖 4-41 Q3 所有資料為輸入值，再到 swissprot 查詢

The screenshot is similar to Figure 4-40, showing the 'Expression' field with the text 'Q3 < swissprot'. The same set of four green square buttons and a green scroll bar are visible to the right of the input field.

你還可以以 Q3 當中所有的紀錄當成查詢的輸入值，做進一步的查詢，例如想在 swissprot 資料庫中取得跟 Q3 相關的所有資料，他的指令如下。

4. 瀏覽索引

(a) 關於瀏覽索引

在其他的查詢方法，你可以從索引中來做查詢

圖 4-42 欄位資料頁中可以提供“Description”欄位的查詢

The screenshot shows a web-based SRS interface. At the top is a navigation bar with a logo, followed by links for TOP PAGE, QUERY, RESULTS, PROJECTS, VIEWS, DATABANKS, and HELP. The main area has a red sidebar on the left with menu items: Field Name, Description, Data-fields in SRS, and Browse Index. The 'Description' item is currently selected and highlighted in red. The main content area displays the 'Description' field's definition: 'This is probably the best data field for searching an entry you don't know very much about; however, you can't easily search for entries of a class, since often different conventions are used for naming enzymes, organisms, genes, etc..'. Below this is a table titled 'Data-fields in SRS' with one row for 'Description'. The table has columns: Databank, Name, Short Name, Type, No of Keys, No of Entry References, Indexing Date, and Status. The 'Description' row shows: EMBL, Description, des, index, 0, 0, and see member library. At the bottom is a search form with the text 'List Values' and a search criteria: 'that match *' and 'and occur in at least 1 entries'.

(b) 瀏覽索引

你可以直接瀏覽欄位資料，從資料欄位中選擇你所有興趣的資料欄位種類，如“Description, SeqLength”。

1. 在欄位資料頁中選擇資料庫，並且輸入你想看的欄位值注意，這裡的查詢方式是不同於關鍵字查詢，系統不會做查詢字串的延伸，因此所輸入的資料必須是完全正確，才會有結果
2. 按下“List Values”按鍵，你將會看到瀏覽索引頁

圖 4-43 濱覽 Description 資料欄的索引



The screenshot shows a table titled "Values in EMBLRELEASE" with two columns: "Value" and "No of Entries". The "Value" column contains various entries like "!!!!", "#kung", "#", "#.", "#00206.", "#00236.", "#00245.", "#00289.", "#007", "#0087-1d13", "#01", "#010", and "#0101_3). The "No of Entries" column shows the count for each entry. A "More Values" link is available at the top left, and a "Make Query" button is at the bottom left.

Value	No of Entries
<i>Values in EMBLRELEASE</i>	
<input type="checkbox"/> !!!!	463
<input type="checkbox"/> #kung	2
<input type="checkbox"/> #	19039
<input type="checkbox"/> #.	31
<input type="checkbox"/> #.	1
<input type="checkbox"/> #00206.	1
<input type="checkbox"/> #00236.	1
<input type="checkbox"/> #00245.	1
<input type="checkbox"/> #00289.	1
<input type="checkbox"/> #007	84
<input type="checkbox"/> #0087-1d13	1
<input type="checkbox"/> #01	1
<input type="checkbox"/> #010	1
<input type="checkbox"/> #0101_3)	1

3. 點選想看的欄位值再按下“Make Query”按鍵，將會看到所有相符的資料

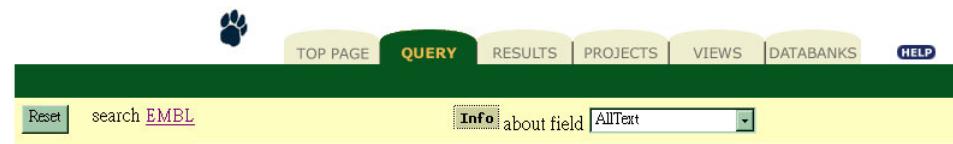
(c) 逐步進入欄位資料頁

有許多方式可以進入欄位資料頁，這裡將詳述最普通的方式

從查詢格式進入

從標準查詢模式及近接查詢模式中上方的下拉式選單進入

圖 4-44 進階查詢模式中點選“Info”鍵將可進入下拉式選單中所顯示的欄位



從下拉式選單中選擇欄位資訊，並點選“Info”

在進階查詢中，所有欄位名稱都是以超連結連結至欄位資料頁

第四章 SRS 資料庫查詢

圖 4-45 在進階查詢中資料欄都是以超連結方式呈現

Field Name	Query	Include in View
AllText	<input type="text"/>	
ID	<input type="text"/>	<input type="checkbox"/>
Division	<input type="checkbox"/> est <input type="checkbox"/> fun <input type="checkbox"/> gss <input type="checkbox"/> htc <input type="checkbox"/> htg <input type="checkbox"/> hum <input type="checkbox"/> inv <input type="checkbox"/> mam <input type="checkbox"/> mus <input type="checkbox"/> org <input type="checkbox"/> phg <input type="checkbox"/> pln <input type="checkbox"/> pro <input type="checkbox"/> rod <input type="checkbox"/> sts <input type="checkbox"/> syn <input type="checkbox"/> unc <input type="checkbox"/> vrl <input type="checkbox"/> vrt	<input type="checkbox"/>
AccNumber	<input type="text"/>	<input type="checkbox"/>
SeqVersion	<input type="button" value="≥"/> <input type="text"/> <input type="button" value="≤"/> <input type="text"/>	<input type="checkbox"/>
Molecule	<input type="checkbox"/> circular dna <input type="checkbox"/> circular rna <input type="checkbox"/> dna <input type="checkbox"/> rna <input type="checkbox"/> xxx	<input type="checkbox"/>
Description	<input type="text"/>	<input type="checkbox"/>
Keywords	<input type="text"/>	<input type="checkbox"/>

你可以從超連結進入你想要去的欄位資料頁，從資料庫資訊頁進入。

在首頁當中“Databank”欄將可以讓你進入資料庫資訊頁，在這裡你可以看到所有可以使用的資料庫，點選資料庫的超連結，將可進入資料庫的資訊頁，並且包含資料庫中所有的資料欄。

圖 4-46 在資料庫資訊頁中，資料欄將以超連結方式呈現

Data-fields in SRS	Name	Short Name	Type	No of Keys	No of Entry References	Indexing Date	Status
	AllText	all	group				see member library
	ID	id	id				see member library
	Division	div	index				see member library
	AccNumber	acc	index				see member library
	SeqVersion	sv	num				see member library
	Molecule	mol	index				see member library
	Description	des	index				see member library
	Keywords	key	index				see member library
	Organism	org	index				see member library
	Taxon	tax	index				see member library
	Organelle	ogn	index				see member library
	Comment	cc	index				see member library
	DateCreated	crd	num				see member library
	LastUpdated	crlu	num				see member library
	SeqLength	sl	num				see member library
	Link	lnk	show				see member library
	Sequence	seq	show				see member library

點選其中一個資料欄之超連結，將會引導你進入欄位資料頁

第五章 資料庫搜尋與多序列排比

范廷佳¹、楊永正²

¹ 陽明生物資訊研究中心、² 陽明大學生物資訊研究所

一、 簡介

由於遺傳工程技術的進步，決定 DNA 序列比決定蛋白質序列要容易許多，使得資料累積十分迅速；因此，目前有許多序列雖已被決定出來，但到底具有何種功能則不清楚。在這樣的情形下，兩個序列間的相似性往往用以協助鑑定蛋白質或核酸的功能。過去曾有人在純化蛋白質時，錯誤地純化出血清中的主成份 BSA 或是親合力管柱中的 ConA 的例子；若當時研究者能先定出產物之部分序列，並做搜尋資料庫，就不致產生此種錯誤。因此在決定蛋白質序列或選殖基因的過程中，只要有部份序列資料產生，就應搜尋資料庫，以確定所找到的序列是否值得繼續研究下去。資料庫的搜尋雖是一個非常重要的工具，若不小心使用，極易被結果誤導，必須充份瞭解程式的運作才知道如何解釋結果。

一般評估資料庫搜尋程式是根據靈敏度(sensitivity)，選擇性(selectivity)與速度等三方面來討論的。靈敏度不夠，就會誤將有親緣關係的序列判斷為雜訊，而選擇性太低則會誤將不相干的特性，例如序列組成的相似性或是疏水性等特性判斷為有親緣關係。目前各種基因體研究計劃正在進行，在資料庫不斷膨脹的情形下，我們被迫在準確性(包含了靈敏度與選擇性)與速度間求取一平衡點。在此比較不同的程式，讓使用者能根據自己的需要來選用程式。不同的程式各有一套方法來判斷其真偽，初學者沒有必要去學所有的方法，而應集中力量弄清楚分析的原理，一旦徹底弄懂了原理，以後再學不同的方法就容易多了。

資料庫搜尋相當於將查詢的序列，與資料庫中所有的序列一一做區域性的序列排比，然後找出可能具有生物意義的序列。雖然資料庫搜尋程式會應用特定的運算法加速分析，可是還是必須先瞭解搜尋字串的基本策略(第十章第三節)，才能進一步瞭解較複雜的序列排比。

不論是廣域性或是區域性的序列排比，都會耗用很大儲存空間，例如兩個 1Kb 的序列互相比對，至少要佔用 10^6 個位置存得分。雖然資訊學者有方法可以降低對記憶體之需求量，這種類型的分析還是較慢的。除了上述因子外，計算次數也是另一個限制速度的因子。在做雙序列排比時或許不覺，可是若要以一個序列搜尋資料庫序列時，計算所花的時間就很可觀，因此才有人設計不同的運算法來解決運算速度的問題。在此將先討論 FastA，作為討論目前使用最廣的 Blast 的基礎。為求實用，在下面的介紹中，將先討論如何分析數據，在有必要時才去討論運算法。

二、FastA

1. 如何解讀 FastA 的輸出結果

一個典型的 FastA 的輸出，包含三個部份。第一部份是一張柱狀圖(histogram)(圖 5-1)，第二部份是一張相似序列的清單(圖 5-2)，第三部份是查詢(query)序列與資料庫中序列的序列排比結果(圖 5-3)。其中又列出四種不同的分數 z-score、initl、initn 與 opt 及一個期望值 E()，以利分析輸出結果的生物意義。想要解讀結果就必須追問這些分數是怎麼出來的。

若一個查詢序列在整個資料庫中都找不到任何一個和它相似的序列，此時資料庫中的序列，對此查詢序列而言都是雜訊。以生物學家的語言來說，如果兩序列中沒有守舊的區域，可能會因為結構單元(胺基酸或核苷酸)的種類不同，所以產生的雜訊也不同。胺基酸的種類有 20 種，所以隨機產生的配對機會較少也就是雜訊較低，核酸序列則因核苷酸種類較少，比對時的雜訊較高。若有一個估計相似性的計分方法，資料庫只每一序列與查詢序列的相似性都可被算出，以同一得分數的序列總數，對得分做柱狀圖(圖 5-1)時，理應類似於常態分佈(normal distribution)，也就是得分越靠近平均值的序列越多，而得分非常高或非常小的序列較少。當然這只是一個理想的狀況，事實上較長的序列比較有機會得到高分，若不加修正，則會稍偏離理想的曲線。在輸出的柱狀圖中顯示兩種符號，「=」代表根據計分結果所繪出的分佈圖，「*」代表得分經過長度修正後所繪出的分佈圖。若以統計的術語來說，在距離隨機配對的得分平均值數個標準差(standard deviation)之外所看到的序列，比較不可能是因為隨機配對而得高分的。換言之，得分較高的序列中，較有可能找到真正的訊號。在得分高的地方，如果看到所找到的序列數(「=」)高於理論上可觀察到的序列數目(「*」)，就代表在資料庫中有一些和查詢序列相似的序列(即「訊號」)，可是在這些得分高的序列中可能也混雜著一些不相干的雜訊，此時，就需要用統計的方法來檢視訊號的真實性。在圖 5-1 中每一符號表示 88 個序列，而圖下部(z-score 在 92 以上)，右側有一放大圖，每一符號表示七個序列，這樣才看的出「*」與「=」的相對大小。

圖 5-1 FastA 的柱狀圖

Histogram Key:

Each histogram symbol represents 88 search set sequences

Each inset symbol represents 7 search set sequences

z-scores computed from opt scores

z-score obs exp

	(=)	(*)
< 20	178	0 :*===
22	0	0 :*
24	1	0 :*
26	1	1 :*
28	6	13 :*
30	76	81 :*
32	287	312 :====*
34	919	846 :=====*=
36	1963	1737 :=====*===
38	3319	2870 :=====*======
40	4353	4004 :=====*======
42	4923	4894 :=====*======
44	5258	5399 :=====*======
46	5147	5499 :=====*======
48	4735	5264 :=====*= * *
50	4623	4804 :=====*= *
52	3929	4223 :=====*= *
54	3494	3607 :=====*=
56	2941	3013 :=====*=
58	2447	2474 :=====*=
60	2007	2004 :=====*=
62	1582	1607 :=====*=
64	1302	1278 :=====*=
66	1104	1010 :=====*=
68	835	794 :=====*=
70	710	622 :=====*=
72	602	486 :=====*=
74	449	379 :=====*=
76	330	295 :====*
78	269	229 :==*=
80	218	178 :==*
82	163	136 :=*
84	110	108 :=*
86	94	84 :*=
88	66	65 :*
90	62	50 :*
92	50	39 :* :=====*===

```

94     38     30 :*      :====*=
96     26     23 :*      :====*
98     14     18 :*      :==*=
100    15     14 :*      :=*=-
102    11     11 :*      :=*-
104    7      8 :*      :=*-
106    14     6 :*      :=*=-
108    4      5 :*      :=*-
110    5      4 :*      :=*-
112    9      3 :*      :=*=-
114    6      2 :*      :=*-
116    7      2 :*      :=*-
118    6      1 :*      :=*-
>120   301    1 :*==== :*=====
Results sorted and z-values calculated from opt score
1555 scores saved that exceeded 77
49745 optimizations performed
Joining threshold: 37, optimization threshold: 25, opt. width: 16

```

在圖 5-2 中所列出的期望值代表一個序列會因隨機之方式而得到 z-score 或更高分的次數。在期望值大於「1」時，你所看到的結果可能就沒有統計上的意義，也就是說，它對查尋序列的相似可能只是隨機的。可是沒有統計上的意義不代表沒有生物意義，事實上在大的基因族(gene family)中，一些親緣關係較遠的序列，期望值可能落在 1~10 之間，甚至更高，使用者應根據自己的需求來判斷你要將期望值的門檻設在何處，因為列出一堆沒參考價值的搜尋結果，只有徒增分析上的困擾。例如你若對基因族的分析不感興趣，只希望尋找基因的可能功能，或是尋找可能存在的模組樣式，那就沒有必要去分析那些期望值很大的序列。可是如果你的序列在資料庫中找不到很相似的序列，或是你想分析某一基因的基因族，那就必須進一步瞭解計分的方式。

圖 5-2 不同ktup值時，FastA程式所找到的相似序列清單。

A. ktup = 2

```
The best scores are:           init1 initn   opt    z-sc E(58538)..
SW:TF3A_XENLA      Begin: 1  End:  344
! P03001 xenopus laevis (african claw... 2467  2467  2467  2843.4      0
SW:TF3A_XENBO      Begin: 1  End:  339
! P17842 xenopus borealis (kenyan cla... 2002  2101  2120  2444.5      0
SW:TF3A_RANPI      Begin: 1  End:  320
! P34695 rana pipiens (northern leopa... 1522  1522  1605  1852.5      0
SW:TF3A_BUFA       Begin: 1  End:  339
! P34694 bufo americanus (american to... 955   1465  1549  1788.0      0
SW:P43_XENLA       Begin: 17 End:  307
! P25456 xenopus laevis (african claw... 281   432   665   771.2  2.2e-36
SW:P43_XENBO       Begin: 9  End:  307
! P25066 xenopus borealis (kenyan cla... 266   383   641   743.6  7.6e-35
SW:ZN41_HUMAN      Begin: 262 End:  517
! P51814 homo sapiens (human). zinc f... 115   115   511   591.7  2.2e-26
SW:ZN83_HUMAN      Begin: 38  End:  292
! P51522 homo sapiens (human). zinc f... 91    149   494   573.6  2.2e-25
SW:MLZ4_MOUSE      Begin: 93 End:  388
! Q03309 mus musculus (mouse). zinc f... 109   109   482   560.3  1.2e-24
SW:Z135_HUMAN      Begin: 197 End:  452
! P52742 homo sapiens (human). zinc f... 106   106   482   559.3  1.4e-24
SW:ZO22_XENLA      Begin: 183 End:  430
! P18745 xenopus laevis (african claw... 128   194   478   555.1  2.4e-24
SW:Z136_HUMAN      Begin: 201 End:  462
! P52737 homo sapiens (human). zinc f... 110   110   452   523.9  1.3e-22
SW:ZO71_XENLA      Begin: 624 End:  894
! P18751 xenopus laevis (african claw... 103   103   450   518.5  2.6e-22
```

B. ktup = 1

```
The best scores are:           init1 initn   opt    z-sc E(58538)..
SW:TF3A_XENLA      Begin: 1  End:  344
! P03001 xenopus laevis (african claw... 2467  2467  2467  3023.9      0
```

SW:TF3A_XENBO	Begin:	1	End:	339				
!	P17842	xenopus borealis	(kenyan cla...	2008	2117	2120	2598.3	0
SW:TF3A_RANPI	Begin:	1	End:	335				
!	P34695	rana pipiens	(northern leopa...	1522	1522	1607	1969.1	0
SW:TF3A_BUFA	Begin:	1	End:	339				
!	P34694	bufo americanus	(american to...	966	1514	1549	1897.8	0
SW:P43_XENLA	Begin:	17	End:	307				
!	P25456	xenopus laevis	(african claw...	324	634	665	812.9	1.1e-38
SW:P43_XENBO	Begin:	9	End:	307				
!	P25066	xenopus borealis	(kenyan cla...	310	614	641	783.4	4.6e-37
SW:ZN81_HUMAN	Begin:	4	End:	270				
!	P51508	homo sapiens	(human). zinc f...	115	205	531	649.1	1.4e-29
SW:ZN43_HUMAN	Begin:	418	End:	698				
!	P28160	homo sapiens	(human). zinc f...	114	212	529	640.3	4.3e-29
SW:ZO6_XENLA	Begin:	177	End:	438				
!	P18749	xenopus laevis	(african claw...	108	108	511	622.4	4.3e-28
SW:ZN85_HUMAN	Begin:	340	End:	595				
!	Q03923	homo sapiens	(human). zinc f...	121	121	500	606.9	3.1e-27
SW:ZG57_XENLA	Begin:	34	End:	331				
!	P18729	xenopus laevis	(african claw...	117	117	495	604.9	4e-27
SW:ZN41_HUMAN	Begin:	144	End:	405				
!	P51814	homo sapiens	(human). zinc f...	131	218	495	601.5	6.3e-27
SW:Z135_HUMAN	Begin:	135	End:	415				
!	P52742	homo sapiens	(human). zinc f...	109	203	488	593.9	1.7e-26

2. Fast A 計分的方式

根據 Pearson 的建議，在期望值低於 0.02 以下的序列，可認為是與查詢序列同源的(homologous)。若所找到的序列的期望值都高於 0.02，就必須增加程式的靈敏度，也就是降低 ktup 值後再搜尋一次。要瞭解 ktup 如何影響靈敏度，就必須由搜尋速度與靈敏度的關係談起。

如果要以查詢序列和資料庫中的每一個序列做序列排比會花許多時間，要想加快搜尋的速度就必須能很快挑出值得分析的序列，然後才對這些序列做序列排比。在 1983 年 Wilbur 與 Lipman⁽¹⁾首先引入排出值得分析的序列的快速方法，他們先利用混亂編碼(hash coding)的方式建立對照表，以尋找兩個序列相似之處，可是這樣所找到的相似區都是很小的片段，其長度為給定之「字」的長度。因此需要利用點矩陣法來延伸所找到的區域，最後再針對上述區域插入空隙，求取最佳之並列方式。在混亂編碼的階段所尋

找的是連續配對的區域，若使用的「字」較長，雖能很快找到相似性高的區域，卻可能會找不到較短的相似區。例如在圖 10-4 中，雖可找到兩個 TCG，可是找不到任何一 TCGA。換言之，若字的長度為 4 個鹼基，就無法在圖 5-1 中找到相似區。如果使用較長的字做混亂編碼，會使程式的靈敏度(sensitivity)降低，找不到已知存在的相似性。那為什麼不直接採用較短的字長呢？當然，最理想的狀況是字長為「1」，也就是直接比對，不做混亂編碼，可這樣就無法加快運算的速度，因此使用者必須根據自己的需求，在速度與靈敏度之間求取一折衷點。

這一個策略在 1985 年經過 Lipman 與 Pearson⁽²⁾的修正，到 1988 年 Pearson 與 Lipman 再修正寫成了現在大家所熟知的 FastA 程式⁽³⁾。它與最初發展出來的策略有兩項重要的修正。一是在計分時，程式會先存下得分最高的十個，而不是一個點矩陣中的對角線（即區域性的相似），這樣才不會誤失相似性較低的區域。其次，程式會根據規則將數個對角線連成一個較大的區域，在此允許加入空隙(gap)，以增加靈敏度。這一個經過連接的區域才會利用 Needleman-Wunsch algorithm 來做序列排比。

綜合上述 FastA 發展之過程，此程式的運作可分成三個步驟。第一步是利用給定之字長，製作對照表，尋找連續配對的小片段，並找出配對密度最高的十個片段的位置。其次是利用相似矩陣(例如 PAM250 或 Blosum50)針對這十個片段重新計分，並除去每個區域兩側得分較低的部份。此時所得的每個區域都是沒有空隙，但可能有誤配的並列序列，此時相似的區域的分數會被存下，稱之為「initl」。然後 FastA 會試著將相鄰的相似區域(即對角線)連接起，這個過程允許插入空隙，唯每插入一個空隙就要扣一定的分數。在經過連接後，最相似的區域的得分稱之為「initn」。最後的一步的序列排比是採用趙坤茂博士所設計的方法，它是 Smith -Waterman algorithm 的一種變異，在性質上還是區域性的序列排比，但是速度較快。在得到最佳並列方式後之得分，稱為「opt」，程式會根據序列長度校正此值，而計算出 z-score，前述之柱狀圖就是根據 z-score 所繪的。

圖 5-3 FastA的序列排比和區域性序列排比之輸出完全相同。

```
ID  P43_XENLA      STANDARD;      PRT;    365 AA.
AC  P25456;
DT  01-MAY-1992 (REL. 22, CREATED)
DT  01-MAY-1992 (REL. 22, LAST SEQUENCE UPDATE)
DT  01-FEB-1994 (REL. 28, LAST ANNOTATION UPDATE)
DE  P43 5S RNA BINDING PROTEIN (42S P43) (THESAURIN B). . . .

SCORES      Initl:  281  Initn:  432  Opt:  665  z-score: 771.2 E(): 2.2e-36
Smith-Waterman score: 665;   33.7% identity in 294 aa overlap
                           10       20       30       40       50
```

在將 `ktup` 調到「1」之後，若可以看到期望值小於 0.02 的序列，而且又無低複雜性的或重覆的序列存在，這些序列就很可能與查詢序列同源。事實上「連接」這個步驟的效果很容易由比較各種不同的得分看出來，在連接後所得的「`initn`」通常都會比連接前的「`init1`」大；可是經過做最佳化之後的「`opt`」，則有可能小於「`init1`」。「`opt`」

小於「init1」代表連接過程有不妥當之處。Pearson 的建議是在發現完全不相干的序列有小於 0.2 的期望值時，就考慮增大「Gapweight」，這樣可使期望值小於一的不相干序列的數目下降數倍。因為罰分不夠重時，可能會將不相干的序列和同源的區域接(join)在一起，反而降低了「opt」與 z-score。

表 5-1 參數 ktup 的設定對 FastA 輸出的影響

Statistics	ktup=2	ktup=1
scores saved that exceeded 77	1555	1843
optimizations performed	49745	50877
Joining threshold	37	43
optimization threshold	25	31
opt. width	16	32
p43 init1	281	324
initn	432	634
opt	665	665
z-score	771.2	812.9
E(58538)	2.2e-36	1.1e-38
% identity	33.7	33.7
aa overlap	294	294

在將 ktup 值降為「1」後，可發現在整體的統計數字上，會存較多序列的得分，也會對較多之對角線作最佳序列排比，因此在做連接及最佳序列排比時由程式自動調整之門檻值也都較高。這些都表示程式的靈敏度提高了。若針對某一找到的序列，例如 *Xenopus* 的 p43 蛋白質，來做分析，會發現 init1 與 initn 都大幅度提高，這表示這一蛋白質比在用 ktup = 2 時更容易通過 cutoff 值，唯有先通過 cutoff 才有機會進行序列排比，求出 opt。第一階段利用混亂編碼的方法找尋與延伸相似片段時，如果 ktup=2，則可找到一個範圍從 79 到 86 的片段(圖 5-4A)，它的兩側都是連續兩個相同的核苷酸。若 ktup=1 則邊緣部份只需一個核苷酸，所以相似片段的範圍較大(圖 5-4B，從 67 到 93)，得分較高，也就是比較容易通過門檻而被留下來。

圖 5-4 不同 ktup 值時，第一階段所找到的片段長度不同

A.**ktup=2**

	60	70	80	90	100	110
tf3a_xenla	HLTRHSLTHTGEKNFTCDSD GCDILRFTTKANMKKHFNRFHN IKICVYVCHFENCGKAFKK					
	:: : : ::: : : : :::: : : : : :: :					
P43_XENLA	QILKHVKRHLALKLSCPTA GCKMTFSTKKSLSRH KLYKHGEAVPLK-CFVPGCKRSFRK					

B.**ktup=1**

	60	70	80	90	100	110
tf3a_xenla	HLTRHSLT HTGEKNFTCDSDGCDILRFTTKANMKKHFNRFHN IKICVYVCHFENCGKAFKK					
	:: : : ::: : : : :::: : : : : :: :					
P43_XENLA	QILKHVKR HLALKLSCPTAGCKMTFSTKKSLSRH KLYKHGEAVPLK-CFVPGCKRSFRK					

事實上只要一旦片段超過門檻，不論 ktup 是多少，最佳的序列並列是完全一樣的(亦可由% identity 等數值看出)。問題是同樣的 opt，為何會有不同的 z-score 呢？(比較圖 5-2 A 與 B)這是因為在做統計分析之前必須先剔除得分高的序列，這樣才能將資料庫中的序列當做不相干的序列。因為參考的序列數目受 ktup 影響，連帶的也影響 z-score 的計算。

在此要特別強調，在序列相似時，不論選用哪一種常用的計分系統(PAM250, Blosum50,⋯⋯等)或是不同的插入空隙罰分，所得的期望值變異都不大，不太可能誤判，所以統計的數據可作為支持同源性的證據。可是對於親緣關係較遠的序列而言，參數的改變影響很大，所以統計數字遠不如實驗數據可靠。一般而言，只能用統計數字證明兩序列的同源性，而能不用統計數字證推斷兩個序列沒有同源性。

在利用插入空隙罰分(gap insertion penalty)，降低不相干序列之得分後，若仍有一些相似與否不很確定的序列時，可用隨機重排(random shuffling)的序列來測試而找到的序列是否有「統計」上的相關性。其目的是要排除序列組成所可能有的些許效應，因此將查詢序列隨機重排，產生數百種隨機序列，再將這些序列當成資料庫做比對，計算各相似序列的期望值。在此隨機資料庫中最高分的相似序列之期望值，大於資料庫查詢時所算出的期望值，則表示此相似序列有統計上的意義。對大部份的應用而言，沒有必要做此分析。

在序列排比的部份(圖 5-5)還列出相同(identity)胺基酸或鹼基的百分比，這是判斷同源性的另一個重要指標。對一個給定的序列而言，假設任何一個位置上的胺基酸或核

苷酸皆可獨立突變，經過一段很長的時間後，突變序列和原序列的相似性會達平衡而不再改變。在統計上來說，突變後的序列會與原序列分別有 5% (1/20) 與 25% (1/4) 的相同序列。當然 5% 與 25% 只是平均值，若考慮到得分的分佈情形，一個長 50 個胺基酸的序列在 95% 的情形下(兩個標準差)，會與一個隨機序列會有 5-11% 的相同序列。序列越長，則標準偏差越小，例如 200 個胺基酸長的蛋白質序列，則有 95% 的機會和一隨機序列有 5-9% 的相同序列。如果考慮插入空隙，則序列比對時的百分比相同之平均值會隨著加入空隙的量增加而上升。若允許隨意加入空隙，則在沒有空隙的區域可能達到 100% 的相同。綜合考慮插入空隙的影響，與統計上分佈的問題，一般認為兩個蛋白序列有 20% 的相同序列時是不能確定是否有同源性的。若是相同序列小於 15% 或大於 25%，則分別代表無，或有同源性。

3. FastA 的應用與限制

在表 5-1 中，前四項的期望值都是零，事實上也們所代表的是不同生物的 TFIIIA 的蛋白質。可是自第五個項目開始則 Z-score 降低甚多。若觀察序列排比的結果(圖 5-6)，則發現相同的序列僅 37%，可是它們都發生在重要的位置。如粗體字部份所示，許多相同的胺基酸都是鋅指中的守舊胺基酸。這個結果顯示 FastA 除了有能力找到廣域的相似性外，也有能力找到模組樣式，事實上這是尋找新的模組樣式的一個很好的起點，這在高級課程中將有進一步的解說。

FastA 的作者已將程式的性能調到很好的狀態，因此也沒有給使用者太多的調整空間，一般的使用者事實上也沒有必要多作調整。對使用者而言，真正重要的是解其性能，不要做超出它能力之外的事。例如 FastA 在做對照表時，不會將 B、Z、X 等具有疑慮的符號轉換為多個可能的胺基酸，它只把這些符號當作多出的胺基酸。換言之，在對照表比較中，X 只會與 X 配對，而不會與其他胺基酸配對。使用者若不瞭解此特性，就可能誤失可能有的相似序列。

三、Blast

1. 如何解讀 Blast 的輸出結果

動態線性規劃(Dynamic programming)能有系統地尋找同源的序列，可是它的速度太慢。在資料庫搜尋上通常是利用經驗法則(rules of thumb)先挑出值得仔細分析的片段，再做 Needleman-Wunsch 的最佳化排列。像 FastA 這樣利用法則加速搜尋的方法，稱之為「heuristic algorithm」。雖然 FastA 已將搜尋資料庫所需的時間減少了許多，可是在資料庫不斷擴增的情形下，若有更快的方法能找到資料庫中的相似序列，將有助於瞭解 DNA 語言。Blast (Basic Local Alignment Search Tool)是根據統計理論所設計的，

雖然它也是一種建築在經驗法則上的演算法，其計分方式卻能模擬找出突變最少的狀況的過程。這種的計算速度約比 FastA 快一個數量級以上。

因為是利用數學導衍出序列排比時得分的統計分佈，所以可以利用公式直接計算統計上之係數，其典型的輸出結果並不需要柱狀圖但仍包括相似序列的清單(表 5-2)與序列排比的結果(圖 5-5)。在導衍公式時 Blast 並不試圖連接各相似的片段，換言之它不允許空隙的存在，所以它會計算每一個相似區的得分，並將此序列中得分最高的片段的分數列出。HSP (High-scoring Segment Pair) 僅代表一些分較高的片段，它們單獨存在時或許無法通過統計上的測驗，可是數個連在一起，則通過測驗，在清單中的「Smallest sum probability」就代表將數 HSP 連在一起之後之統計資料。其中 N 代表所參與的 HSP 的個數，P(N) 代表在給定條件的搜尋中，找到與 high score 得分相同或更高分的片段的機率。每一個蛋白質的長度不同，即使得分相同，其機率也不相同。在相似序列的排列順序上，並不是根據得分排序，而是根據機率排序，所以有些得分較高的序列反而被排在後面，對蛋白質的比較來說，機率小於 0.02 就被認為是同源的。因為程式預設保留 250 個序列，若被認為有意義的序列超過此數字，程式會自動警告你(表 5-2 底部)。如果被認為有意義的序列總數超過 1000，可能是在序列中有一些重覆序列，必須將其濾掉，以免干擾搜尋的結果(請參閱第 5-15 頁，「Blast 的應用與限制」)。

表 5-2 BLAST典型輸出結果形式

```
WARNING: -hspmax 100 was exceeded with 23 of the database sequences, with as
many as 230 HSPs being found at one time.
```

Sequences producing High-scoring Segment Pairs:	Smallest		
	Sum		
	High	Probability	
Sequences producing High-scoring Segment Pairs:	Score	P(N)	N
..			
SW:TF3A_XENLA ! P03001 xenopus laevis (african clawed fro...	1930	1.2e-268	1
SW:TF3A_XENBO ! P17842 xenopus borealis (kenyan clawed fr...	1564	2.4e-231	2
SW:TF3A_RANPI ! P34695 rana pipiens (northern leopard fro...	1173	2.7e-172	2
SW:TF3A_BUFA ! P34694 bufo americanus (american toad). t...	741	8.2e-161	3
SP_HUM:Q13097 ! Q13097 homo sapiens (human). dna/rna-bind...	544	4.8e-149	3
SP_HUM:Q92664 ! Q92664 homo sapiens (human). xenopus tran...	537	2.2e-141	3
SP_HUM:Q12963 ! Q12963 homo sapiens (human). transcriptio...	454	5.5e-134	3
SP_OV:P79797 ! P79797 ictalurus punctatus (channel catfis...	510	5.4e-86	2
SW:P43_XENLA ! P25456 xenopus laevis (african clawed frog...	231	2.8e-55	4
SW:P43_XENBO ! P25066 xenopus borealis (kenyan clawed fro...	224	2.5e-52	4
SP_HUM:Q14590 ! Q14590 homo sapiens (human). zinc finger ...	103	1.2e-41	6

SW:ZG3_XENLA ! P18718 xenopus laevis (african clawed frog...)	88	1.5e-40	7
SW:ZG8_XENLA ! P18737 xenopus laevis (african clawed frog...)	95	3.1e-40	7
SW:ZF64_HUMAN ! P15622 homo sapiens (human). zinc finger ...	81	5.6e-40	7
SW:ZG17_XENLA ! P18713 xenopus laevis (african clawed fro...)	81	1.7e-39	7
SW:ZG52_XENLA ! P18727 xenopus laevis (african clawed fro...)	74	4.3e-39	7
SW:HKR1_HUMAN ! P10072 homo sapiens (human). kruppel-rela...	91	4.7e-38	7
SW:Z143_HUMAN ! P52747 homo sapiens (human). zinc finger ...	194	2.4e-37	3
SW:ZO26_XENLA ! P18746 xenopus laevis (african clawed fro...)	80	3.0e-37	7
SP_RO:Q61776 ! Q61776 mus musculus (mouse). zinc finger p...	86	3.0e-37	7
SP_HUM:Q15914 ! Q15914 homo sapiens (human). zfoc1 (fragm...	101	1.5e-36	6
SW:ZO61_XENLA ! P18750 xenopus laevis (african clawed fro...)	90	2.0e-36	7
SW:ZG28_XENLA ! P18716 xenopus laevis (african clawed fro...)	95	2.4e-36	7
SW:HF12_HUMAN ! P13683 homo sapiens (human). zinc finger ...	78	4.0e-36	6
SP_HUM:Q99676 ! Q99676 homo sapiens (human). kruppel-rela...	91	5.0e-36	8
SP_RO:Q61898 ! Q61898 mus musculus (mouse). zinc finger p...	86	5.8e-36	8
SP_OV:Q91853 ! Q91853 xenopus laevis (african clawed frog...)	186	6.2e-36	3
SP_HUM:Q14584 ! Q14584 homo sapiens (human). zinc finger ...	87	6.3e-36	7

...

SW:MTF1_MOUSE ! Q07243 mus musculus (mouse). transcriptio...	204	4.0e-34	3
SW:ZG7_XENLA ! P18735 xenopus laevis (african clawed frog...)	80	5.1e-34	6
SP_RO:P97365 ! P97365 mus musculus (mouse). zfp64. 5/97	75	7.1e-18	6
SP_HUM:Q13106 ! Q13106 homo sapiens (human). zinc finger ...	77	7.5e-18	4
SW:REX1_MOUSE ! P22227 mus musculus (mouse). rex-1 protei...	186	8.5e-18	1

\End of List

WARNING: Descriptions of 464 database sequences were not reported due to the limiting value of parameter V = 250.

在序列排比結果(圖 5-5)的部份，Blast 比 FastA 的輸出(圖 5-3)多很多，因為 Blast 會列出多個相似的區域，若以尋找模組樣式的目的而言，這可能讓我們看到一些在 FastA 中不會出現的相似區。可是若目的是在尋找親緣關係較遠的序列，Blast 就無法產生具有生物意義的並列結果，因為親緣關係較遠的序列可能有許多插入或刪除的序列，若不加入空隙，在序列排比時就會看到許多相似的片段，而看不出整體的相似性。在 FastA 的輸出結果中，可見到 TFIIIA 與 p43 都具有九個鋅指結構，整個結果不到一頁即可印出，可是若看 Blast 輸出中相同的部份則無法一眼看出 TFIIIA 與 p43 間的關係，所看到的是印了近三頁半的相似片段，使用者必須花時間分析，才有可能看到全貌。因此若初步判斷這序列可能有意義，就需另做區域性序列排比。Blast 在蛋白質分析時，其靈敏

度與 FastA 相若，在做核酸分析時則較差。為何 Blast 的速度比 FastA 快，而又不影響靈敏度呢？這必須由其計分方式說起。

圖 5-5 Blast 程式所列出的序列排比結果。

```
>SW:P43_XENLA P25456 Xenopus laevis (african clawed frog). p43 5s rna binding
protein (42s p43) (thesaurin b). 2/94
Length = 365

Score = 231 (108.4 bits), Expect = 2.8e-55, Sum P(4) = 2.8e-55
Identities = 45/115 (39%), Positives = 64/115 (55%)

Query: 194 CDVCNRKFRHKDYLRDHQKTHEKERTVYLCPRDGCDRSYTTAFNLRSHIQSFHEEQRPFV 253
        C C + F+      LR H+ TH K+      CPR     CD+++++ FNL   H++   H   +
Sbjct:  193 CAACKPKFKASALRRHKATHAKPLQLPCPRQDCDKTFSSVFNLTHHVRKLHLCLQTHR 252

Query: 254 CEHAGCGKCFAMKKSLERHSVVHDPEKRKLKECPRPKRSLASRLTGYIPPKSKE 308
        C H+GC + FAM++SL RH VVHDPE++KLK K R           R T     P +E
Sbjct:  253 CPHSGCTRSFAMRESLLRHLVVHDPERKKLKLKFVRGSKFLGRGTRCRTPVVEE 307

Score = 156 (73.2 bits), Expect = 2.8e-55, Sum P(4) = 2.8e-55
Identities = 27/79 (34%), Positives = 42/79 (53%)

Query: 15 CSFADCAGAYNKNWKLQAHLCHTGEKPFPCKEEGCEKGFTSLHHLTRHSLTHTGEKNFT 74
        C A C A Y K   KLQ H+ H+ +KP+ C + C+K F       + +H   H   K +
Sbjct:  17 CPAAGCKAFYRKEGKLQDHMAGHSEQKPKWCGIKDCDKVFARKRQILKHVKRHLALKLS 76

Query: 75 CDSDGCDLRFTTKANMKHH 93
        C + GC + F+TK ++ +H
Sbjct:  77 CPTAGCKMTFSTKKSLSRH 95

Score = 101 (47.4 bits), Expect = 9.8e-14, Sum P(2) = 9.8e-14
Identities = 21/57 (36%), Positives = 28/57 (49%)

Query: 223 CPRDGCDRSYTTAFNLRSHIQSFHEEQRPFVCEHAGCGKCFAMKKSLERHSVVHDPE 279
        CP  GC    +++T  +L   H     H E   P   C     GC + F   K++L RH   VH   E
Sbjct:  77 CPTAGCKMTFSTKKSLSRHKLKYHGEAVPLKCFVPGCKRSFRKKRALRRHLSVHSNE 133

Score = 95 (44.6 bits), Expect = 3.5e-34, Sum P(3) = 3.5e-34
```

```

Identities = 21/74 (28%), Positives = 29/74 (39%)

Query: 34 LCKHTGEKPFPCKEEGCEKGFTSLHHLTRHS LTHTGEKNFTCDSDGCDLRFTTKANMKKH 93
        L KH      P C    GC++   F      L RH     H+E      CD  GC + ++ A +   H
Sbjct: 97 LYKHGEAVPLKCFVPGCKRSFRKKRALRRHLSVHSNEPLSVCDVPGCSWKSSSVAKLVAH 156

Query: 94 FNRFHNKICVYVC 107
        R      +      C
Sbjct: 157 QKRHRGYRCSYEGC 170

```

2. Blast 的計分方式

雖然 FastA 與 Blast 同為 heuristic algorithm，前者在計算的過程中並無統計理論的基礎，而是在找到相似區之後再做統計意義的評估(利用最佳化之後的得分算期望值)。Blast 則是以嚴謹的統計理論，直接選出有意義的區域再做序列排比。不過目前的理論無法處理空隙的問題，這是 Blast 只能輸出相似片段的原因。

Blast 利用混亂編碼的方式，能夠很快地找到相似性高的區，並向兩側延伸，因此每一對序列都有許多相似的片段。在這些片段中，得分最高的片段稱之為 MSP(Maximal Segment Pair)。根據統計理論，可直接由隨機序列的模型中，算出在給定的查詢序列長度，資料庫大小之下，讓此片段的相似性具統計意義的門檻值(cutoff score) S。在比較的過程中，會先選擇值得分析的「字」，也就是其得分必須高於某一可調整的門檻值 t，這個字才會被納入計分，進一步分析其延伸後的總分是否超過 S。在 Blast 中所使用的設定是電腦模擬所求出的最佳「字長」(word length) 門檻值 T，能平衡速度與靈敏度的需求。整個的篩選過程中不需做序列排比，即可確定各片段在統計上的意義，所以速度很快。若在找到有意義的區域後，自動做區域性為序列互序列排比，則速度就與 FastA 相似，而不再具有特色。因此 Blast 是以搜尋速度為主要的訴求，讓使用者很快找到相似的序列，再決定該怎樣做下一步的分析。

3. Blast 的使用與限制

為了達到增加計算速度的目的，Blast 所使用的資料庫是經過壓縮，可能載入記憶體的。在 NCB1 的 Blast 伺服器甚至讓資料庫常駐在記憶體中讓大家共用，因此可服務全世界的需求。也因為它的資料庫需要壓縮，它不能使用資料庫的某一部份，也不接受檔名檔。如果有特別的需求，須先用 GCGToBlast，將 GCG 的序列變成 Blast 的資料庫後才能使用。

另一個使用上的技巧是將低複雜性的序列過濾掉，例如重覆序列或 Gln 較多的區域等。Blast 所根據的統計理論，假設胺基酸或核苷酸出現在給定位置的機率是正比於資料庫的胺基酸或核苷酸組成。這些低複雜性的序列，不能滿足此前提，所以必須去除後才能防止程式找到許多這種片段。Blast 提供 x-filter 可將短的重覆序列改為 X，也提供 s-filter 將低複雜性的序列改為 X，這樣在比對時需求指令行加入「-filter=xs」，即可同時啟動這兩個濾器。圖 5-6 中列出 GCG 線上輔助系統上的一個例子，是將一經過濾的序列與原序列並列在一起，其中 X 是被濾掉的區域，因此可看出哪一類型的序列會被除去。

圖 5-6 使用Blast程式中的過濾器除去低複雜性的或重覆性的序列的意義。

```

1 MAAKIFCLIMXXXXXXXXXXXXXIFPQCSQAPIASLLPPYLSPAMSSVCENPILLPYRIQQ 60
1 MAAKIFCLIMLLGLSASAATASIFPQCSQAPIASLLPPYLSPAMSSVCENPILLPYRIQQ 60

61 AIAAGIXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNIRXXXXXXXXXXXXYQQQQFLPFN 120
61 AIAAGILPLSPLFLQQSSALLQQQLPLVHLLAQNIRAQQLQQQLVLANLAAYSQQQQFLPFN 120

121 QXXXXXXXXXXXXXXXXXPFSQLAAAYPRQFLPFNQLAALNSHAYVXXXXXPFSQLAAVS 180
121 QLAALNSAAYLQQQQLLPFSQLAAAYPRQFLPFNQLAALNSHAYVQQQQLLPFSQLAAVS 180

181 PAAFLTQQQQLLPFYLHTAPNVGTXXXXXXXXTNPAAFYQQPIIGGALF 235
181 PAAFLTQQQQLLPFYLHTAPNVGTLLQLQQQLLPFDQLALTNPAAFYQQPIIGGALF 235

```

四、以蛋白質序列搜尋 DNA 資料庫

如果所測試的基因能轉錄產生 mRNA，再轉譯產生蛋白質，則比較蛋白質序列會比將 DNA 序列直接比較要好。因為遺傳密碼是多對一的，所以蛋白質序列相同時，DNA 的序列可能不同。而核苷酸又只有四種，所以隨機產生相配合的序列的機會遠大於蛋白質序列。一般來說，在經過適當排列後，蛋白質序列只要有 25% 以上完全相同就可判定是相關的序列，而兩個不相關的 DNA 序列很容易就有 40% 左右的相同。因此，比較蛋白質序列時訊號與雜訊之比例較大。此外，在未轉譯區中或在反向，或不同 reading frame 的 DNA 序列上也有可能找到與測試序列相似的蛋白質。目前 PIR 等蛋白質資料庫，雖將 Genbank 中已知的蛋白質序列轉譯出來，對於其他的區域卻未轉譯。若是以蛋白質序列搜尋蛋白質資料庫就不可能找到與這些未轉譯區相關的序列。此外，DNA 資料庫中有許多表現序列標幟(EST, expressed sequence tag)，因為它們只是部份的

cDNA 序列，多數沒有完整的蛋白質，所以也不會出現在 PIR 等蛋白質資料庫中。

基於上述理由，以蛋白質序列搜尋 DNA 資料庫是一件很有意義的事。在技術上，可將資料庫的六個 reading frame 全部轉譯出來比對。在 FastA 系列的程式組，必須使用另外一程式 TFastA 才能執行此功能。而 Blast 會根據給定之序列與資料庫，自動決定是否要轉譯序列，因此使用 Blast 是比較方便的。

五、多序列排比

在 EMBOSS 套組中的多序列排比程式叫做 Emma，其主要目的是做廣域性的序列排比。它適合用來尋找不同生物的同一蛋白質間之守舊區，不適合尋找不同蛋白質間的區域相似性。在瞭解此程式的特性後，有一些變通的辦法，可以讓你找到區域相似性。換言之，雖然 Emma 不是設計用來尋找新的模組樣式的，只要運用得當，亦可用它來找模組樣式。

因為嚴謹地序列排比會佔用非常大的記憶體空間，並且非常耗時。程式使用一種在建立親緣樹時最簡單的運算法 UPGMA(unweighted pair-group method using arithmetic averages)，這個簡單的運算法，在不同分支的演化速度相近時，可以用來建立親緣樹。因為在上述假設之下，核苷酸或胺基酸的置換速率與親緣遠近大約成正比，所以使用算術平均數還算合理。此法採用一系列漸進的雙序列排比來做。在程式啟動後，會先將各序列兩兩比對，以找出未來做進一步並列的順序。原則上是先將最相似的序列排列在一起，變為一群(cluster)，然後再將剩餘序列中與這兩個序列最相似的一個，與這兩個排好的序列群做序列排比。以圖 5-2A 為例，TF3A-XENLA 會先與 TF3A-XENBO 先做序列排比，其次 TF3A-BUFAM 會與 TF3A-RANPI 做序列排比，然後這兩組結果再做並列，其結果最後再與 TF3A-YEAST 做並列。

在做並列的過程中如果要插入空隙，則該群序列中的每個序列都會在同樣的位置插入空隙。像這樣漸進的，兩兩並列的方法所得的序列排比結果，會與哪一個序列先參與序列排比有關。目前所採用的方法，只有在演化速率相當時才能用，否則就無法得到滿意的結果。此外，若序列太不像，程式就會一直引入空隙，以至於所得到之並列結果沒有任何意義。因此程式預設最多插入 2000 個空隙，超過此值，程式就會自動停止，以免印出沒有意義的結果，由此可知，若將多個只有小部份區域相似的序列放在一起是無法找到共有序列(consensus)的，即使是相似的基因，若不妥善處理，也可能會有問題。

六、結語

一般而言，兩個排比好的蛋白質序列中，相同的胺基酸之比在 25% 以上可確定有同源性，在 15% 以下大概沒有同源性，若介於 15-25% 之間則只能當做參考，必須靠實驗做進一步的驗證。可是在仔細去看序列排比結果之前，可利用 FastA 的期望值，或

Blast 之機率小於 0.02 做為一個指標，來挑選可能有趣的序列。

在蛋白質分析上，FastA 與 Blast 之靈敏度相似，可是在核酸比對上，Blast 的靈敏度略遜一籌。即使如此，如果你的目的是要找非常相似的序列，用 FastA 與 Blast 其實無所謂。如果你在做完資料庫搜尋後要直接做多序列排比(例如設計引子，找協同變異等)，則 FastA 較容易得到好的序列排比結果。

參考文獻

1. Wilbur W. J. and Lipman D. J. (1983) Proc Natl Acad Sci U S A. 80(3):726-30.
2. Lipman D. J. and Pearson W R. (1985) Science. 227(4693):1435-41.
3. Pearson W. R. and Lipman D. J. (1988) Proc Natl Acad Sci U S A. 85(8):2444-8.

第六章 由應用例中學 EMBOSS 程式的功能

許玉璇¹、楊永正²

¹ 陽明生物資訊學程、² 陽明大學生物資訊研究所

遺傳工程的技術使基因選殖的工作成為現代生物學研究的一個重要步驟。雖然選殖的方法有很多，可是一旦選殖出一個基因，其序列分析步驟幾乎是相同的。在此將先介紹這幾種已經標準化的分析步驟，然後再回去介紹在選殖基因時可能用到的幾種序列分析工具。

而關於 EMBOSS 的分析工具，除第三章所介紹的線上輔助系統，在其原始網站 (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/index.html>) 或是 EMBOSS 主機上的說明文件內，有一說明文件將程式依其功能分類，例如序列排比(Alignment)，顯示(Display)，編輯(Edit)，核酸(Nucleic)，蛋白質(Protein)等等，總計有十大類三十三項，透過這樣的分類方式可以很快找到需要程式及說明文件。

一、 基因分析之標準步驟

序列片段的組合

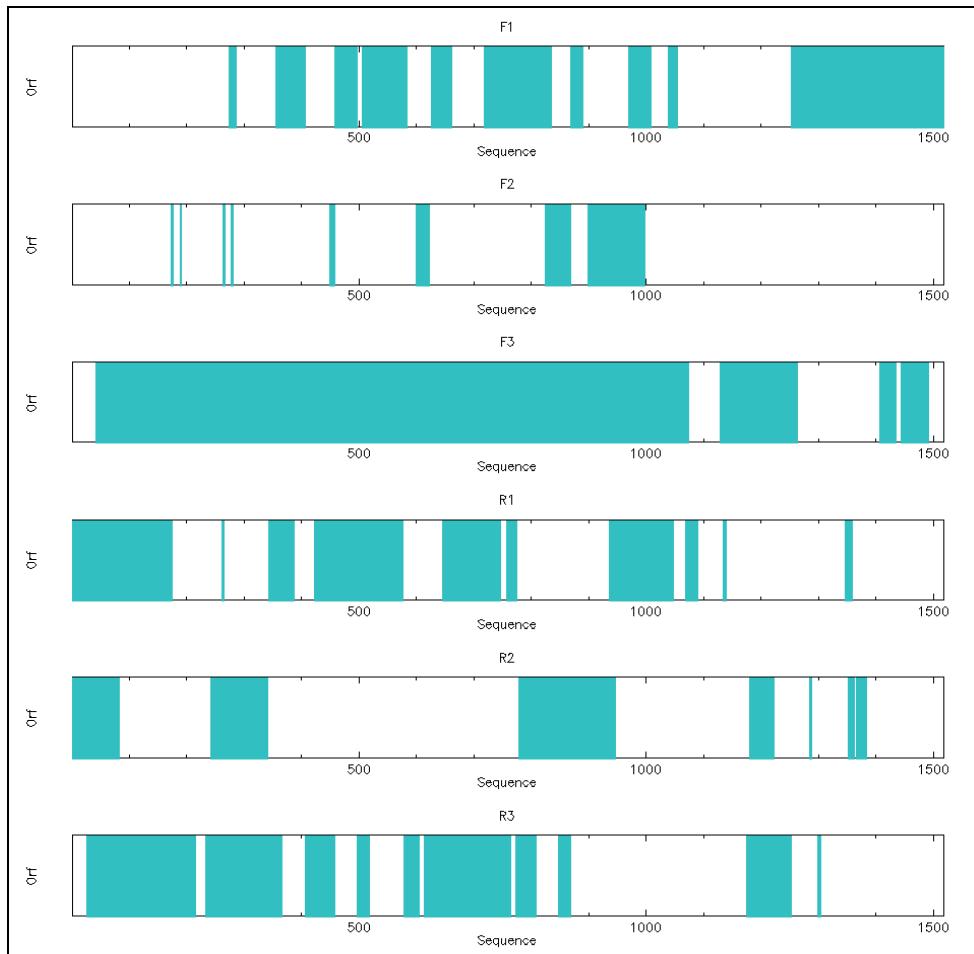
不論由基因體(genomic DNA)或 cDNA 基因庫中選殖出一段基因後，第一步工作就是決定其完整序列。但是在 EMBOSS 程式之中並沒有完整一套，用於核酸序列組合的程式，目前建議以使用 STADEN pregap4+gap4 (http://www.mrc-lmb.cam.ac.uk/pubseq/staden_home.html)、Phrap (<http://www.phrap.org/>) 或 CAP (<http://www.cs.sunysb.edu/~algorith/implement/cap/implement.shtml>)；但如果是只是欲將調部分重疊的片段合併，則可以使用 merger 這支程式，如果是要將多個序列合併成單一條序列，則可以使用 union 程式，但這兩支程式都需要進行事先分析。

1. 尋找開放讀架(open reading frame, orf)

如果是一個可轉譯為蛋白質的序列，一旦組合成完整完整序列後，就要尋找開放讀架(open reading frame)以確定蛋白質在基因中的位置。「尋找核酸基因(Nucleic Gene Finding)」下的 showorf 程式可以列出序列上六個方向所有蛋白質的序列，可是在序列很長時不易由文字檔找到最可能開放讀架，而「尋找核酸基因」下的 plotorf 程式可用圖形顯示開放讀架的大概位置 (參閱圖 6-1)，之後經由圖形判斷其轉譯方向和讀架，再配合 getorf 程式取出實際的位置和蛋白質的序列。因此利用 plotorf 及 getorf 程式是一個尋找開放讀架較好的方式。

第六章 由應用例中學 EMBOSS 程式的功能

圖 6-1 以 TFIIB 為例，顯示 plotorf 程式的輸出結果，第三個讀架上的 ORF 才是真正表現的蛋白質



2. 資料庫比對

在實驗上獲得了部份序列時，即可做資料庫比對，以資料庫中確定是否有與所研究的基因相似的序列。如果已有，或許可由別人的研究中知道這個基因的功能，這種比對有時也有助於確定蛋白質轉譯的讀架。在 EMBOSS 程式之中並沒有可以做序列資料庫比對的程式，所以在此建議使用美國生物技術資訊中心的 Blast 伺服器 (<http://www.ncbi.nih.gov/BLAST/>) 或是 Virginia University 的 FastA (<http://fasta.bioch.virginia.edu/>)。FastA 是一個比較靈敏、但是速度較慢的程式。可是在應用上，以蛋白質序列比對核酸資料庫其實比較有意義，因為核酸上可能有一些有意義的蛋白質序列是以前未被深入研究的，例如在反向、或不同的讀架上，可能會有一些在蛋白質資料庫中找不到蛋白質序列。此外，由於密碼的第三個位置的辨識不精確(wobble)，因此在演化的過程中，可能核酸的序列有變化而卻不會影響蛋白質的序列。若以蛋白質的序列比對核酸資料庫，則比較不易受到這些雜訊的干擾。TFastA 這程式即是以蛋白質序列比對核酸序列資料庫，

其性質與 FastA 相似，雖然速度較慢但是較為靈敏。Blast 則是一組程式，它會自動根據輸入的序列與所選擇的資料庫種類而執行適當的程式，它的特點是搜尋速度很快，缺點是在序列相似性較低時會有失誤，有可能會漏掉一些相關的序列。在時間充裕的情況下當然應以完備為先，應盡量使用 FastA 或 TFastA；若資料量大，需要快速分析，或是目的只在找相似性高的序列，那就可考慮使用 Blast。

3. 序列排比

如果在資料庫中可以找到相似的序列，就需要比較自己研究的序列與這些相似序列間的關係。如果資料庫中只有一個相似的序列，在做完資料庫比對後，不論是 FastA、TFastA 或 Blast 都會將相似之處列出。不過這些程式主要是做區域性的序列排比(local sequence alignment，參閱圖 6-2)。換言之，程式只會將兩個序列中最相似的部份切出做序列排比。例如 ABC transporter 與磷酸激酵素(kinase)都有 ATP 接合模組，卻是功能完全不同的蛋白質。其輸出的結果與執行程式 water 或 matcher 是相同的，這種分析的方法只會列出相似的區域，並不會告訴你這區域可能的功能。假設你的目的是在問這兩個序列是否屬於同一個基因族(gene family)，而在乎它們是否有共通的模組，則該用 needle 或 stretcher 程式做廣域序列排比(global sequence alignment，參閱圖 6-2)。這種分析的方式會在序列中插入空隙(gap)，使整段序列能儘可能地排在一起。例如 hemoglobin 與豆科植物中的 leghemoglobin 間的關係可用這種方法呈現出來。

4. 多序列排比

在資料庫中如果有多个序列與你所研究的序列相似，你或許希望能將這些序列並列在一起，以尋找代表重要功能的守舊(conserved)區域。在多序列分析中的 emma (emma 程式為 clustalw 的介面程式)程式是一個很好的程式，可是它並非萬能，所以在 emma 程式執行完後，最好再用人工做編輯，做一些細部的調整，使序列排比更合理。

圖 6-2 廣域與區域性序列排比的比較



在相似性低的區域，我們會希望一眼看出守舊的區域(參閱圖 6-3)，但是對於相似性高的區域，則反而會希望找出具有變異的區域(參閱圖 6-4)。emma 雖然可以做序列排比，但是輸出格式不易閱讀，我們可以利用多序列分析中的 prettyplot 或 showalign 程式，輔助找出共有序列(consensus sequence)或是突顯出想看的區域。

第六章 由應用例中學EMBOSS 程式的功能

圖 6-3 程式showalign的蛋白質序列排比輸出格式，突顯守舊的區域

```

      60       70       80       90       100
-----|-----|-----|-----|-----|-----|
Consensus xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxkxyic
TF3A_BUFA M-----K..IC
TF3A_RANPI -----K....K.YIC
TF3A_XENBO -----K....K.YIC
TF3A_XENLA -----K....K.YIC
TF3A_HUMAN .....IC
TF3A_ICTPU -----K...C
TF3A_YEAST .....
Consensus xxxxxxxxxxxxxxxxxxxxxxxxxkxyic

      110      120      130      140      150
-----|-----|-----|-----|-----|-----|
Consensus sfxDcXasyNkXwklqahlckhtgxrpfxcXxxxcxkgfxtxxxlTrhX1
TF3A_BUFA SF.DC.A.YNK..KLQAHLC KHTG.RPF.C....C.KGF.T...L.RH.L
TF3A_RANPI SF.DC.ASYNK.WKLQAHLC KHTG.RPF.C....C.KGF.T...LTRH..
TF3A_XENBO SF.DC.ASYNK.WKL.AHLC KHTG..PF.C....C.KGF....LTRH..
TF3A_XENLA SF.DC.A.YNK.WKLQAHLC KHTG..PF.C....C.KGF....LTRH.L
TF3A_HUMAN SF.DC.A.Y.K.WKL.AHLC KHTG.RPF.C....C.K.F....L.RH.L
TF3A_ICTPU SF..C.AS..K.WKL.AH.CKHTG.RPF.C...--C.K.F.T...LTRH.L
TF3A_YEAST ....C..S..K...L..HL...PF.C...--C.KG..T...L.RH..
Consensus sfxDcXasyNkXwklqahlckhtgxrpfxcXxxxcxkgfxtxxxlTrhX1

```

圖 6-4 程式showalign的蛋白質序列排比輸出格式，突顯有差異的區域

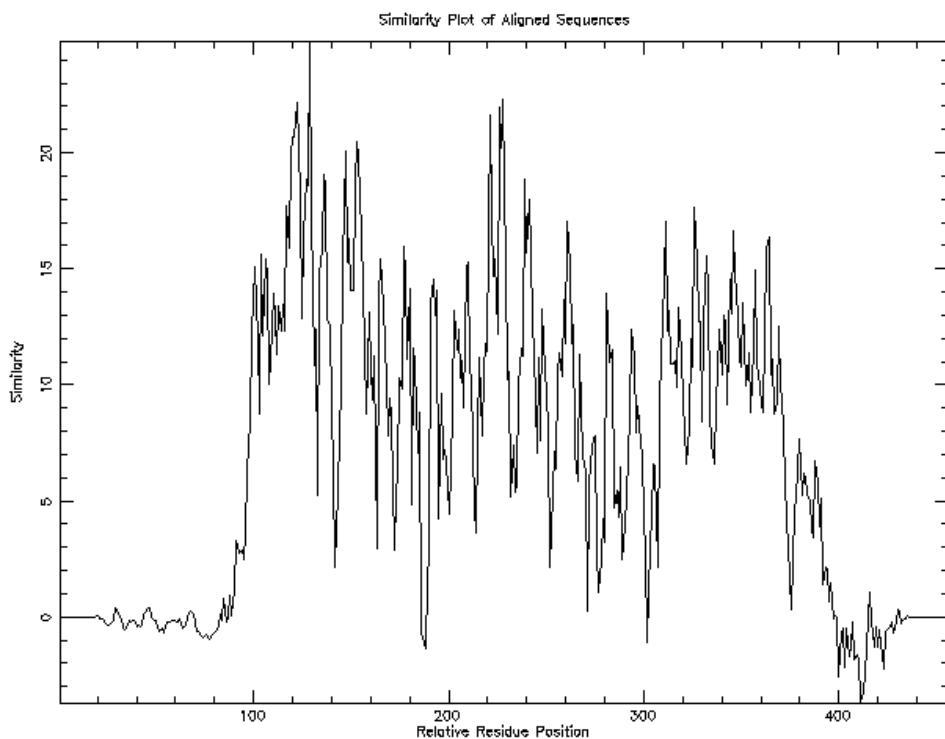
```

      110      120      130      140      150
-----|-----|-----|-----|-----|-----|
Consensus sfxDcXasyNkXwklqahlckhtgxrpfxcXxxxcxkgfxtxxxlTrhX1
TF3A_BUFA ..P..N.t...NR.....E...P.TYEG.E...V.LHH.N..V.
TF3A_RANPI ..A..S....N.....E...P.TVEG.G...V.LFH....Sm
TF3A_XENBO ..A..G....N...r.....Ek..P.KEEG.D...TSLHH....Si
TF3A_XENLA ..A..G.a...N.....Ek..P.KEEG.E...TSLHH....S.
TF3A_HUMAN ..P..S.n.s.A...D.....E...V.DYEG.G.A.IRDYH.s..I.
TF3A_ICTPU ..Ln.K..fs.A...e..Y....L...A.DR--.D.T.C.RCQ....N.
TF3A_YEAST QCDK.AK.fV.KSH.eR..YT.sDTk..Q.SY--.G..VT.RQQ.K..Ev
Consensus sfxDcXasyNkXwklqahlckhtgxrpfxcXxxxcxkgfxtxxxlTrhX1

      160      170      180      190      200
-----|-----|-----|-----|-----|-----|
Consensus htgekxfxcxsDGcDlXfxtxaXlkxhxxrxhXkkkxyvcHfxxcxkX
TF3A_BUFA s.....PCK.EteN.n.A.T.AsNmrl.FK.A.SSPAQV...y.AD.GqQ
TF3A_RANPI .....PCK.DaPD...S.T.MTN.rK.YQ.A.LSPSLI.E.y.AD.GqT
TF3A_XENBO .....N.K.D..K...T.T.K.Nm.K.FN.F.NLqLCV....EG.D.A
TF3A_XENLA .....N.T.D.....R.T.K.Nm.K.FN.F.NI.ICV.....EN.G.A
TF3A_HUMAN .....P.V.Aan...QK.N.KsN..K.FE.K.ENqQKQ.i.S.ED.K.T
TF3A_ICTPU s.s.k.PyQ.LE...SES.IsT.G..N.VE.V.QH.EKH...DyEG.A.E
TF3A_YEAST ....-.S.I.PEE..n.R.YKHPQ.rA.ILSV.LH.LTCPH.nKSFQRPY
Consensus htgekxfxcxsDGcDlXfxtxaXlkxhxxrxhXkkkxyvcHfxxcxkX

```

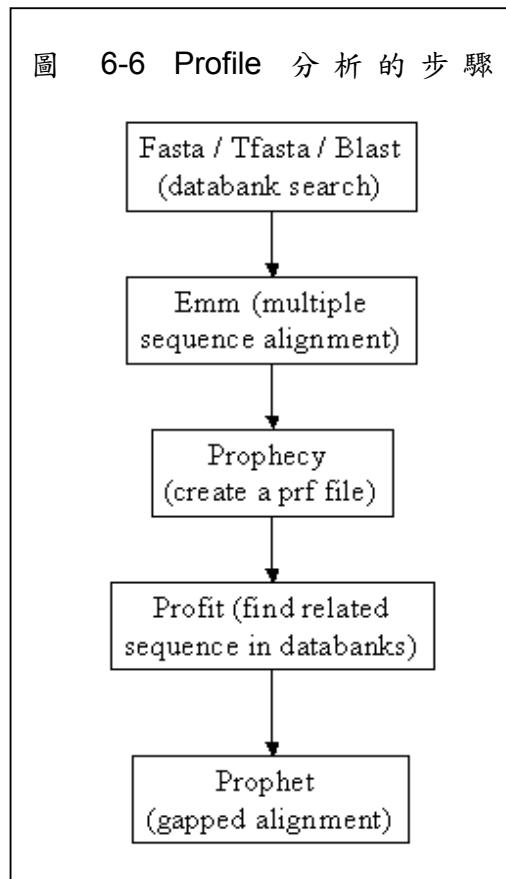
圖 6-5 使用 Plotcon 程式，以圖形顯示多序列排比中相似的區域



雖然 prettyplot 程式可呈現序列中哪一部份相似，可是序列很長時就不易得到一整體的概念。若能用「多序列分析」下的 plotcon 程式，以圖形方式繪出序列中相似的區域(圖 6-5)，則序列中相似的部份就可以一目了然。

5. Profile 分析

在多序列排比時，雖然能在 FastA 所找到的相似序列（參閱第三章結語）中找出共有序列（consensus sequence），可是 FastA 也找不到像 hemoglobin 與 leghemo-globin 這樣遠親。這是因為維持功能的特定胺基酸或活化區的結構不產生改變時，其他位置上中性突變的累積，就足以使兩個有親緣關係的蛋白質的序列變的很不相同。在 EMBOSS 環境下有一系列有關 Profile 分析的程式，就是為了尋找親緣關係很遠的序列，或是在三級結構上相似的序列而設計的。如圖 6-6 所示，其中 prophecy 能將 emma 的輸出檔變為 profile。而 profit 可利用所產生的 profile 來搜尋資料庫，這種搜尋方式比 FastA 靈敏，因為在 profile 中的共有序列，並不是根據在各序列中特定胺基酸是否出現在特定的位置而定的，而是根據胺基酸的相似性而定的，此外在比對資料庫序列時，profile 中比較守舊的區域在計分時所佔的比重較大；反之，變異較大的區域在計分時所佔的比重較小。若需將 profit 搜尋到的序列做序列排比，則可用 prophet 程式。

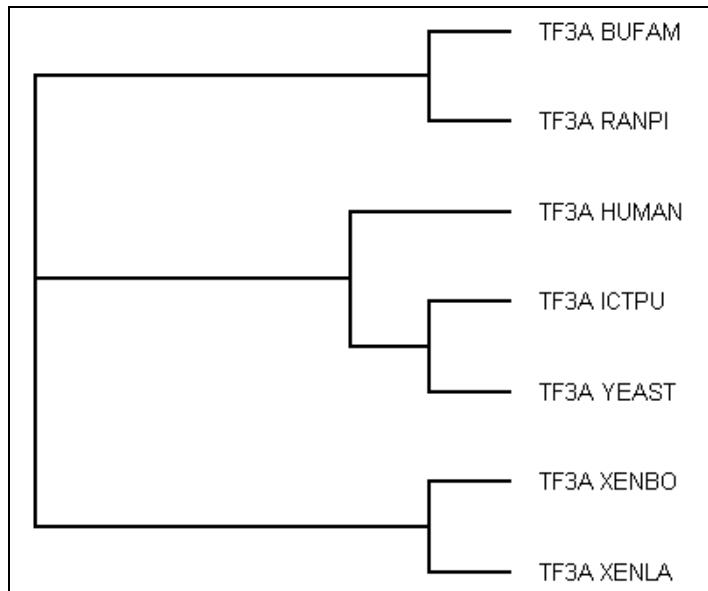


以模組樣式(Pattern)搜尋資料庫

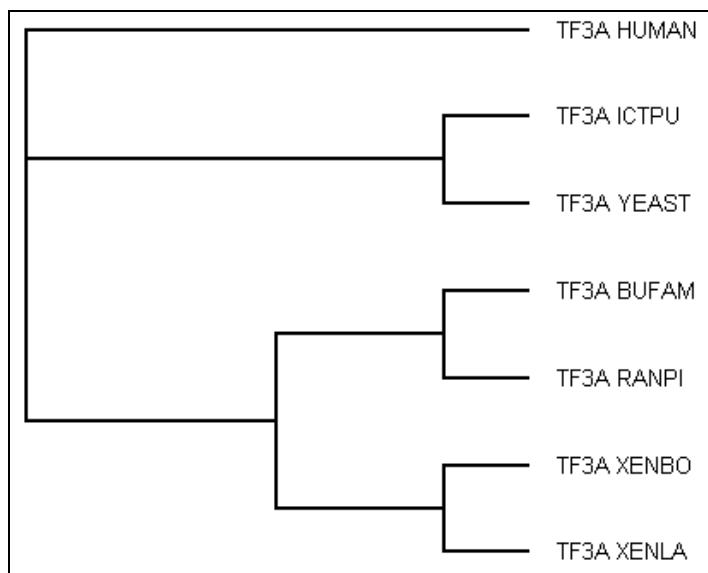
若在多序列排比或 Profile 分析時找到一個可能的模組，則有需要檢視序列資料庫，以確定是否有其他的序列也有這一個可能的模組樣式存在。要達到這個目的，首先必須將共有序列先寫成一個程式可以瞭解的模組樣式(參閱第九章)，然後再以 fuzzpro 程式來尋找資料庫中含有此模組樣式的序列。這將在第九章中討論鋅指模組時，有較仔細的介紹。

圖 6-7 emma 程式所產生的圖不是親緣分析樹

A. emma 所繪之圖



B. Phylip 中 eprotodist 計算 neighbor 所繪之圖



6. 親緣分析

在找出所有相關序列後，可進一步分析這些序列間的親緣關係。雖然在執行 emma 時，可利用 Figure 程式繪出一張類似親緣關係樹的圖(6-7A)，可是這並不是一個親緣分析，它

只繪出在序列排比時，序列比對的次序而已。在演化分析方面，EMBOSS 採用外掛 EMBASSY 套組 Phylip 程式，Phylip 程式是目前廣被使用演化分析程式，做完多序列排比後可直接做演化分析。其中 ednadist、eprotodist 是計算親緣遠近的主要工具，使用者必須瞭解親緣分析的理論，才知道如何選用 ednadist、eprotodist 中不同的計算方法。eneighbor 程式則可將 ednadist、eprotodist 的輸出結果繪成親緣樹(圖 6-7B)。

emma 程式是 clustalw 程式的介面程式，所以 clustalw 產生 treefile(.dnd 檔)，如果要輸出如圖 6-7 的圖，需要在安裝讀 treefile 的軟體，在此介紹這裡所用的 TreeView，是由 University of Glasgow 的 Dr. Rod Page 所發展 (<http://taxonomy.zoology.gla.ac.uk/rod/rod.html>)，可以免費下載安裝，有 PC，Mac 及 UNIX 等平台的版本。而 eneighbor 程式會產生文字模式的親緣樹，但如果要產生正確的親緣關係距離，仍然需要 TreeView 的軟體看圖。

7. 重覆序列的尋找

若在 FastA 搜尋資料庫時沒有找到相似的基因，這表示你所研究的基因很可能是一個新發現的基因。此時就需要做進一步的電腦分析，以便瞭解這個基因的各種性質。在核酸層次上，通常會希望找到具特徵的序列，如 Repeat，Terminator 等。在蛋白質序列上除了重覆序列外，也能夠找到一些已知的蛋白質模組，像鋅指(zinc finger)，加醣類的位置(glycosylation site)或加磷酸的位置(phosphorylation site)等。在實際作法上，可利用「比較」下的 dottup 與 dotmatcher 兩個程式，來尋找同向重覆序列(direct repeat)的大概位置。此法的好處是在尋找重覆序列時可考慮胺基酸間的相似性，而且圖形的顯示可呈現整個分析區域中各重覆序列的相對位置與大小。對核酸序列而言，核苷酸間並沒有相似性的問題，因此可直接利用 equicktandem、etandem 程式來搜尋重覆序列。它的優點是直接列印出重覆序列與其出現位置，不需要去解讀點矩陣；可是它只會根據你給的條件去搜尋。如果給定的條件太嚴，可能會找不到想要的重覆序列，條件太鬆，又找到太多重覆序列，很難分析。在尋找反向重覆序列方面則較困難，必須使用「核酸重複序列(Nucleic Repeats)」下的 palindrome、einverted 程式。palindrome、einverted 程式像 equicktandem、etandem 一樣，若給定的條件不當，可能會找不到序列中存在的反向重覆序列。

8. 檢視序列中是否有已知的模組樣式

若要尋找所研究的核酸序列或蛋白質序列中有哪些已知的模組，則必須有模組資料庫，在 EMBOSS 中提供的轉錄因子核酸位置(nucleic acid site)資料庫是 The German Research Centre for Biotechnology (<http://transfac.gbf.de/>) 所收集的轉錄因子資料庫（這部分的資料需要付費使用），其中列出被這些轉錄因子辨識的核酸序列；在蛋白質模組資料庫方面，有 Swiss Institute of Bioinformatics (SIB) 的 The ExPASy teams (<http://us.expasy.org/prosite/>) 收集的 ProSite 資料庫與 UMBER，the University of Manchester Specialist Node of EMBnet(<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>) 所維護的 PRINTS fingerprint database。查閱這三個不同的資料庫，必須分別使用不同的程式，轉錄因子資料庫使用 tfscan 程式，ProSite 資料庫使用 patmatmotifs 程式，PRINTS

資料庫使用 pscan 程式。所找到模組樣式會提供其特性與參考資料，它的格式與內容可參考第八章中的「鋅指」模組樣式節錄。

9. 蛋白質性質分析

不論在自己所研究的序列上是否能找到已知的模組，我們都希望多瞭解一下這蛋白質的性質或是這個基因的調控區。只是實驗曠日費時，若能在用過去的經驗做分析，而找出一些指導實驗設計的原則，則可省許多力。在 EMBOSS 環境中提供許多這種利用過去經驗來預測性質的工具。許多人覺得這些工具沒有用，因為他們認為預測的結果通常是錯的。可是在另一方面，我們也看到有些人成功地利用這些工具縮短了做實驗的時間。這可能和是否瞭解預測的原理與是否會解釋這些預測的數據有關。

有一些定義較不清楚的模組，只能用一些規則來描述它，而無法寫成像鋅指這樣的模組樣式。例如膜蛋白通常有 α -螺旋可穿過膜，而將蛋白質固著在膜上，而這部份的蛋白質通常具有疏水性，因此可利用蛋白質的結構預測與親水性質，來推測蛋白質是否可能有這樣穿過膜的區域(transmembrane region)。例如在「羅倫佐的油」這部電影中所描述的 ALD 症的致病基因已被找到，有人發現它與酵母菌 peroxisome 中的 pat1 與 pat2 基因很像，而這兩個基因是膜蛋白，卻不是一個脂肪酸代謝酵素，因此推測它與長鏈脂肪酸的輸送⁽¹⁾有關係。

對一個新發現的蛋白質而言，若能瞭解其結構，或許就能猜測它如何執行它的功能。因此可先用 pepinfo 瞭解其蛋白質序列胺基酸基本特性及預測親水性行為(hydropathy)(圖 6-8、圖 6-9)。garnier、helixturnhelix、pepcoil、pepnet 等程式是用來預測二級結構，供使用者自己判斷。而 antigenic 程式是預測蛋白質序列的抗原性(antigenic index)區域。

第六章 由應用例中學 EMBOSS 程式的功能

圖 6-8 以 TFIIIA 為例，顯示 pepinfo 程式的輸出結果

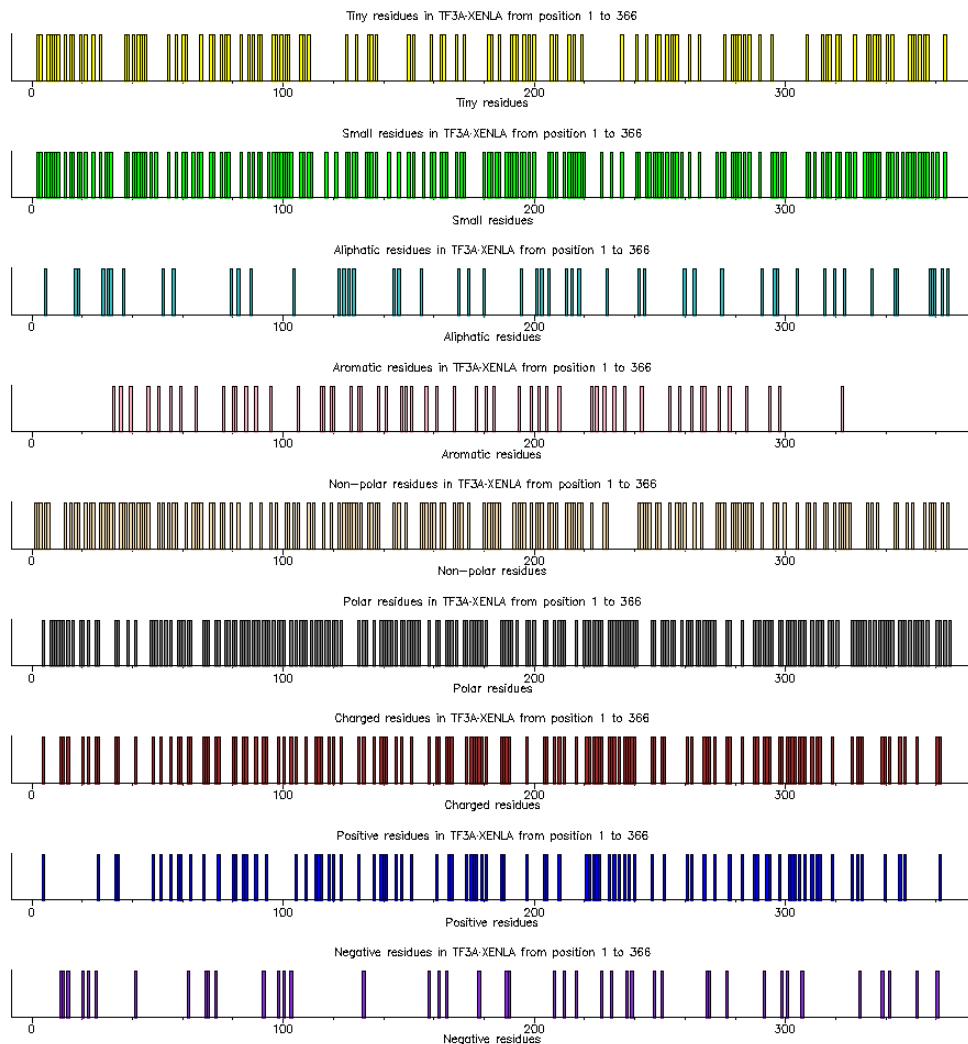
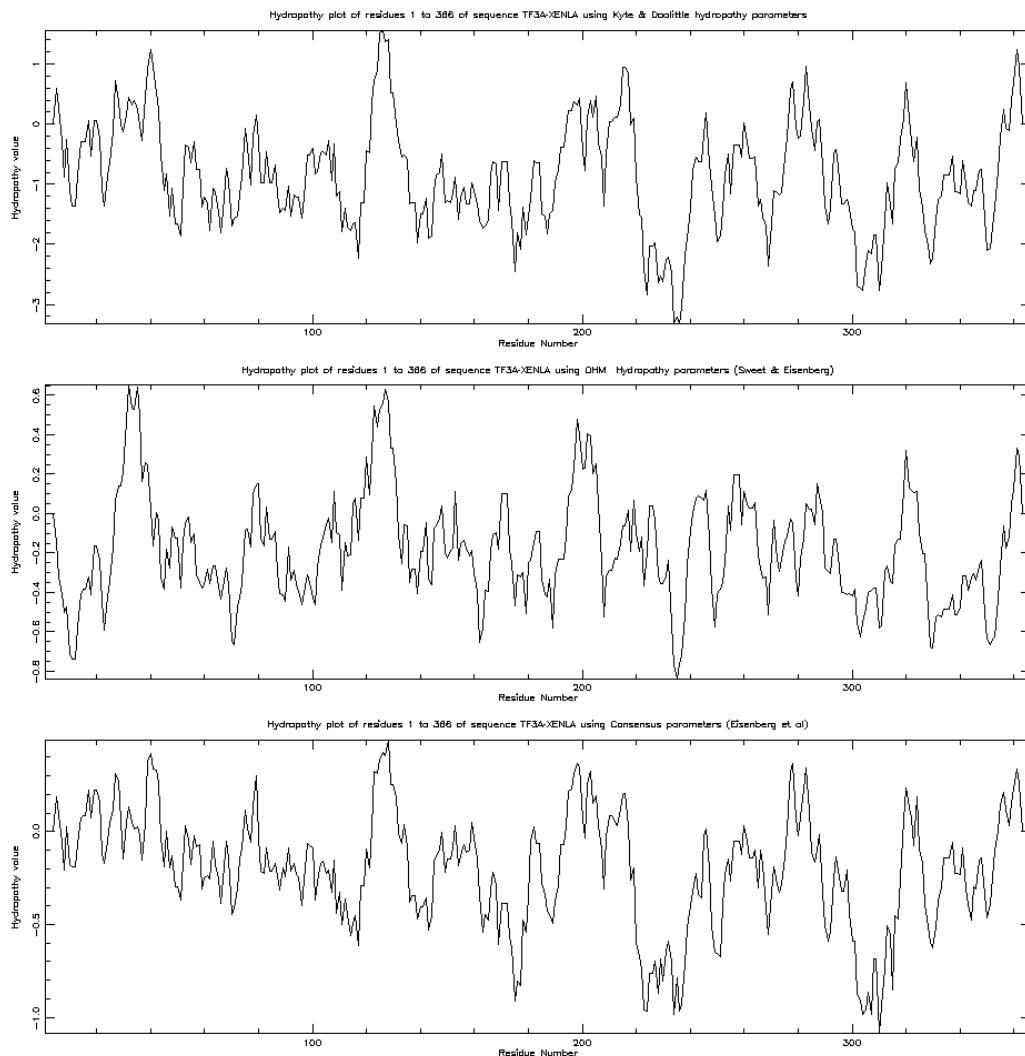


圖 6-9 以 TFIIIA 為例，顯示 pepinfo 程式的輸出結果



在描述蛋白質性質的書中（例如 Branden and Tooze, 1991）⁽²⁾，可以看到有些蛋白質的 α -螺旋在特定位置會出現特定性質的胺基酸。例如白胺酸拉鍊(leucine zipper)剛好在 α -螺旋的一側，有一連串的白胺基酸出現，因為 α -螺旋每一轉是 3.6 個胺基酸，所以大約每隔七個胺基酸(720°)就會出現在螺旋的同一側。這種排列的方式可用像 pepwheel 這樣的程式表現出來(圖 6-10)。又如在 intermediate filament 的蛋白質中，兩股 α -螺旋會相絞而形成一個超螺旋(super-helix)在兩個螺旋交互作用之處會出現疏水性胺基酸，因此這些胺基酸也會出現在螺旋的一側。另有一些兩性的(amphiphilic) α -螺旋一側是親水性的，另一側是疏水性，是否有這些規律性在 pepwheel 的輸出圖形中均可看的很清楚。不過問題是在於蛋白質中可能有多個 α -螺旋，怎麼知道那一段可能具有上述的特定胺基酸分佈呢？hmoment 這程式的輸出(圖 6-11)可以幫助使用者判定那一個區域可能具有兩性的 α -螺旋(參閱第八章)，值得用 pepwheel 做進一步的分析。

圖 6-10 檢視 TFIIIA 蛋白質第四號鋅指中的 α -螺旋是否具有兩性的特性

PEPWHEEL of TF3A-XENLA from 137 to 151

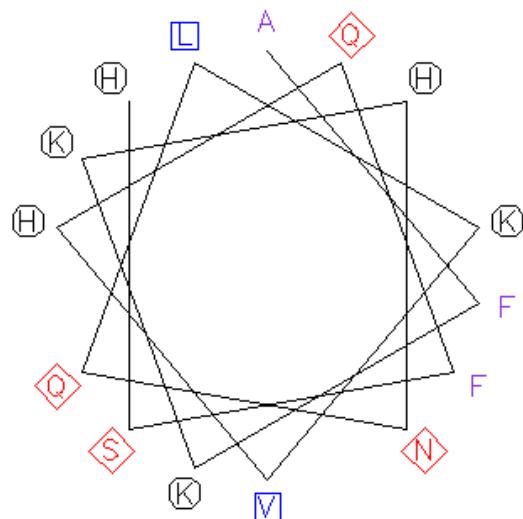
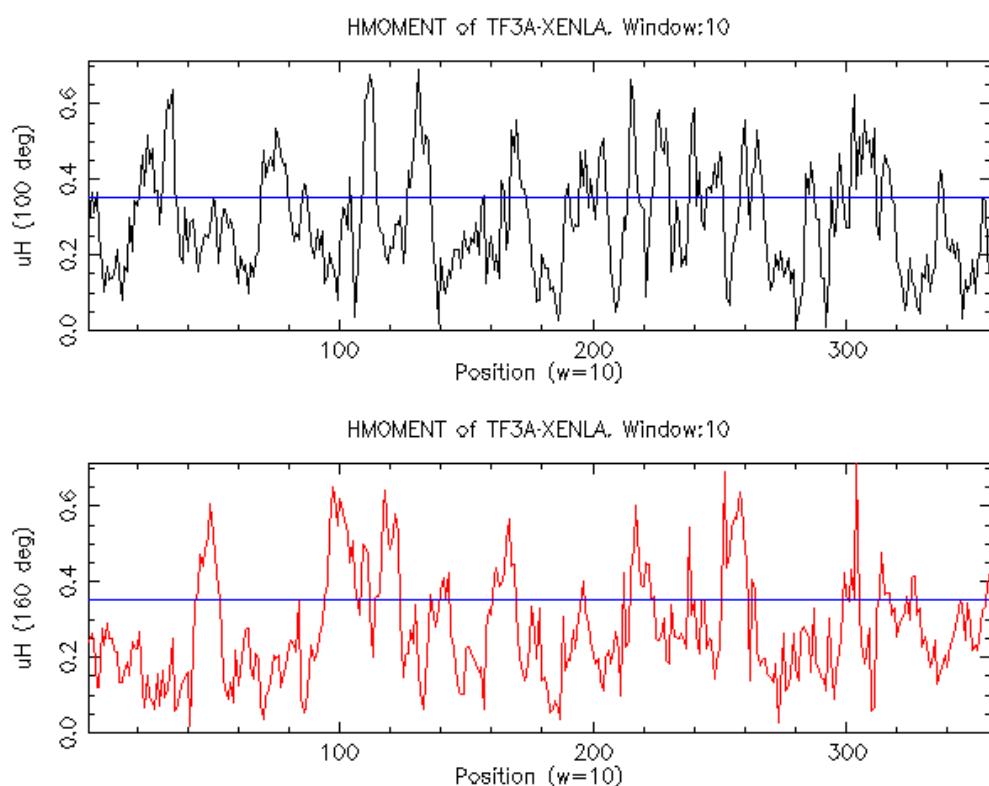


圖 6-11 蛋白質 TFIIIA 的 hmoment 分析

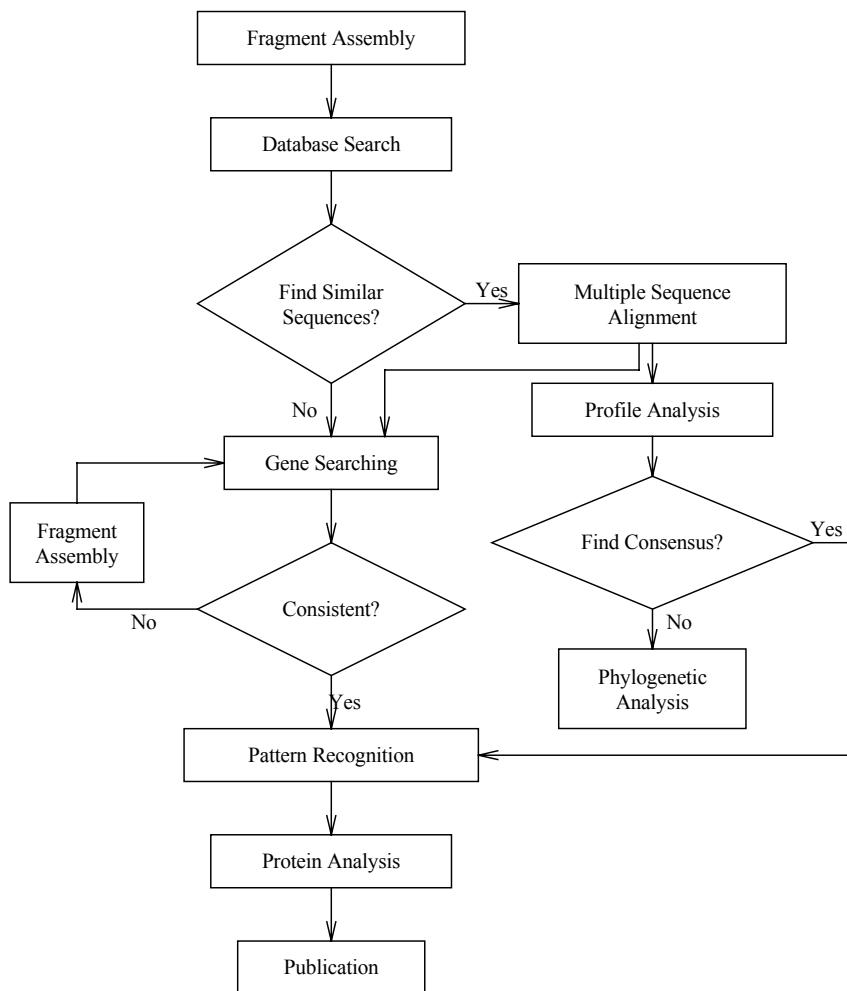


序列的發表

若想發表序列的部份，可用「編輯(Edit)」下的一系列程式來重新組合序列資料。多序列排比的結果則可用 prettyplot 或 showalign 程式來做。至於要將序列整理成適合送到 Genbank 的格式，則需利用 Authorin，Bankit 或 Sequin 等公用程式。

上述各節敘述的是基因分析的標準步驟，圖 6-12 顯示其間的關係：

圖 6-12 基因分析標準步驟間的關係



二、選殖基因時可能用的序列分析步驟

1. 電腦上的基因剪接

在做選殖工作時，常需要限制酵素切割圖譜輔助實驗設計。若載體與要研究的基因的序列都被決定出來了，則可先用 `textsreach` 程式由資料庫中找到序列，但是在 EMBOSS 的說明文件中不建議如此做法，主要原因是這程式的效率不佳，建議以 EBI 的 SRS 或是 NCBI 的 Entrez，再用 `Edit` 相關程式在電腦上剪接出新的序列。一旦有了序列就可利用「核酸限制(Nucleic Restriction)」下的 `remap` 程式，在核酸序列上顯示切割位置與蛋白質序列。若需知道切割後每個片段的大小可用 `restrict` 程式，可用 `cirdna` 程式，繪出質體的圓形圖譜或用 `lindna` 程式繪出 DNA 直線圖譜。這一系列的工作在學會使用 EMBOSS 的技巧後應可自行操作，此外，在習題組解答中有操作步驟，可在陽明的 FTP 伺服器上取得。

2. 模組序列的選殖

想要瞭解一個新發現的基因的功能時，最常用的方式就是檢查其序列中是否有已知功能的模組。有一類型的研究策略是依模組中比較守舊的區域的序列設計 PCR 引子，然後以這些引子由 cDNA 中擴增出這個模組。在決定這些會被表現的模組的序列後，有可能找到這家族的序列中的新成員。例如早在 1986 年，Jackle⁽³⁾的實驗室就利用這樣的策略找出一群含鋅指模組的蛋白質，並發現這一群的蛋白質多是與生物發育有關的轉錄因子。因此這一類型的研究不但可找到新的基因，也能由找到的這一群基因的性質推測模組的功能，最近亦有用這種方法用來比較正常細胞與癌細胞所表現的基因是否有定性上的差異。

在設計引子時首先要找出守舊的區域，因此必須先到資料庫中找出所有相關的序列。在用 `FastA`、`TFastA` 或 `Blast` 找齊了相關序列後，則要用 `emma` 做多序列排比，找出守舊的蛋白質序列，然後再將此序列用 `backtranseq` 反轉錄為核酸的序列。根據此序列所合成出的 `degenerate` 引子，可用來擴增 cDNA 中相對於蛋白質模組部份的 DNA 序列。這些短的、不完整的 cDNA 片段稱之為表現序列標幟(EST，Expressed Sequence Tag)。因為使用的是 `degenerate` 引子，所以一次可能可以找到多種 EST。其中有一部份是當初用來設計引子時所用的已知基因序列，若發現新的 EST，在實驗上可篩選 cDNA 基因庫以釣出全長的 cDNA，並用片段組合的程式組合其序列，再用 `FastA` 等程式搜尋資料庫以確定它是否是一個新發現的基因；在另一方面，可利用基因體分析計劃(genome project)所產生的大量序列資訊來輔助分析。

三、結語

在第三章中已學會如何執行程式與如何使用線上輔助系統。這一章的目的是讓初學者瞭解在研究的過程中會需要哪些工具，這些工具的輸出檔案或圖形像什麼樣子？在未來的幾個章節中，將讓使用者由應用例中摸索使用程式的方法。

參考文獻

1. Hettema E. H., van Roermund C. W., Distel B., van den Berg M., Vilela C., Rodrigues-Pousada C., Wanders R. J., Tabak H. F. (1996). EMBO J. 15(15):3813-22.
2. Branden C. and Tooze J. (1991). Introduction to protein structure. Garland Press, New York.
3. Jackle H., Seifert E., Preiss A., Rosenberg U. B. (1986). J Embryol Exp Morphol. 97:157-68.

參考網站

1. <http://fasta.bioch.virginia.edu/>
2. <http://taxonomy.zoology.gla.ac.uk/rod/rod.html>
3. <http://us.expasy.org/prosite/>
4. <http://www.bioinf.man.ac.uk/dbrowser/PRINTS/>
5. <http://www.cs.sunysb.edu/~algorith/implement/cap/implement.shtml>
6. <http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/index.html>
7. http://www.mrc-lmb.cam.ac.uk/pubseq/staden_home.html
8. <http://www.ncbi.nih.gov/BLAST/>
9. <http://www.phrap.org/>

第七章 問題導向學習(I): 轉錄因子 TFIIIA 的選殖

汪詩海¹、楊永正²

¹ 國家衛生研究院、²陽明大學生物資訊研究所

核糖體 RNA 中的 5S RNA 很早就被選殖出來，因此可在試管內研究其轉錄作用。利用試管內的轉錄系統，發現這個由 RNA 聚合酵素 III 所轉錄的基因的啟動子是位於基因內。另需三個轉錄因子輔助其啟始作用，其中，轉錄因子 TFIIIA 會與 5S 基因的啟動子直接接合。

在卵母細胞中會儲存大量的 5S RNA 和 TFIIIA，作為早期胚胎發育之用。這兩物種會接合而形成 7S 複合物，可以很容易地由卵母細胞中分離出來，所以 TFIIIA 成為第一個被純化出來的轉錄因子。Ginsberg 等人在 1984 年純化出 TFIIIA 後，決定了部份蛋白質的序列，其中序列較清楚的兩段分別是「FHNIKI」與「CHFENCG」。於是他們將這兩段蛋白質序列反轉譯為核酸序列並設計探針去篩選基因庫，現在讓我們利用 EMBOSS 套組來模擬他們分析的過程。

一. 序列的輸入

首先要解決的是取得這兩小段序列的電腦檔以便做反轉譯。在第二章時，我們曾用 vi 編輯器輸入「FHNIKI」的序列，並將其叫做「pep1」。在 EMBOSS 程式之中，對於序列的格式並沒有嚴謹的要求。一般而言，EMBOSS 由 local 的資料庫中取得的序列多為 FastA 的序列格式，除此之外，程式也可以將序列轉換為 genbank，或者是 GCG 的序列格式。但是對於以 vi 編輯器直接輸入的純文字模式的序列，EMBOSS 程式仍舊可以認得，並且能夠直接分析。

二. 反轉譯

在有了 EMBOSS 程式可接受的蛋白質序列檔後，接下來要將其反轉譯為核酸序列。在 EMBOSS 中可使用 backtranseq 程式做反轉譯，它可根據指定的密碼使用表(codon usage table)，對一條輸入的蛋白質序列進行反轉譯的動作。一般而言，預設值使用的是 Ehum.cut 的密碼使用表 (*.cut 代表著 codon usage table 的意思)，但如果今日讀者要使用其他的密碼使用表作為反轉譯的依據時，可以利用：embossdata –showall 的指令，找出適合的密碼使用表檔案，以此檔案作為參考依據即可。

範例 7-1 以 embossdata –showall 顯示合適的密碼使用表

```
% embossdata –showall  
Finds or fetches the data files read in by the EMBOSS programs
```

```
EGC.10  
EGC.11
```

```
EGC.12
EGC.13
EGC.14
:
:
:
Ehuman.cut
Eratsp.cut
Esta.cut
Echmp.cut
Ecristp.cut
Efish.cut
Esty.cut
Esus.cut
:
:
:
```

由此開始是不同生物的密碼使用表，讀者可依所需而自行選擇

由 embossdata –showall 的結果中，我們可以發現程式有提供 *Xenopus* 的密碼使用表，且名稱為 Exenopus.cut，因此我們在進行反轉譯蛋白質序列上所搭配的密碼使用表，就可以輸入 Exenopus.cut 並且得到結果。

圖 7-1 pep1.pro 序列的反轉譯

```
%backtranseq -opt
Back translate a protein sequence
Input sequence: pep1.pro
Codon usage file [Ehum.cut]: Exenopus.cut
Output sequence [pep1.fasta]: pep1a.dna
```

在圖 7-2 中，檔案「pepla.dna」是根據 *Xenopus* 的密碼使用表所寫出的核酸序列。此外，也可以請讀者使用程式預設的密碼使用表作看看，並將輸出結果檔名取為 pep1b.dna，同時比較一下兩者反轉譯出來的結果是否相同？

圖 7-2 backtranseq 的輸出結果

```
% more pep1a.dna
>pep1 protein sequence
TTCCACAAACATCAAGATC
%more pep1b.dna
> pep1 protein sequence
TTCCACAAACATCAAGATC%
```

在顯示的結果中，我們可以發現對於 pep1.pro 的蛋白質序列而言，不論是以 Exenopus.cut 或是 Ehum.cut 的密碼使用表，所得到的結果是一樣的。

三. 開放讀架的搜尋

在選殖出你要的菌株後，通常會抽出 DNA 做序列分析。在本課程中不討論序列的組合(若你需要，在利用輔助系統自學之外，亦可參考習題組的解答)，而直接假設序列已被決定，接下來的問題是選殖出的序列是否確實是 TFIIIA 的基因。回答這問題最簡單的方法是將核酸序列轉譯為蛋白質序列，然後看預測出的序列中是否有 pep1 和 pep2 的序列。要得到 TFIIIA 的正確序列就需先決定開放讀架，這可利用 EMBOSS 下的 plotorf 程式來做。因為 plotorf 程式是一個會輸出圖形的程式，因此可用下列兩個方法在監視器上看見圖形：

(1) 使用 X-Windows 的介面：參閱第二章

(2) 在程式執行之輸入 **Graph type [x11]** 的步驟，選擇要輸出的檔案類型。

於第二個選項中，程式會詢問使用者欲將 plotorf 的分析結果，以何種檔案型態輸出，如：postscript、ps、hpgl、hp7470、hp7580、meta、colourps、cps、xwindows、x11、tektronics、tekt、tek4107t、tek、none、null、text、data、xterm、png，總共有 20 種選項。建議使用者將結果以 png 的圖形檔案輸出，由於 png 圖檔格式的開發主要是用來取代 gif 圖檔，因此 gif 圖檔的優點 png 都有，像是透明背景、交錯顯示、跨平台等，除此之外，它的色彩支援到 48bits，又採取非失真的壓縮方式，所以選取 png 圖檔格式的分析結果不但容易觀看，其圖檔解析度也很好，適合列印或發表。

另一方面，postscript 的檔案格式，也是另外一種相當不錯的選擇，由於 postscript 為向量圖形檔，因此以 plotorf 分析結果的圖檔即使經放大之後，也不容易失真，也是相當適合列印與發表。Postscript 的檔案，可以利用 ghostview 軟體開啟，並加以調整。讀者特別需注意的是：由於 png 與 postscript 檔案無法以 more 指令察看內容，因此必須利用 ftp 將檔案下載回個人電腦中儲存，才能將結果開啟。

以下範例為使用 plotorf 尋找 TFIIIA 核酸序列之 open reading frames 的用法，包括以 png 檔及 postscript 的檔案格式輸出。

範例 7-2 請找出 TFIIIA 在核酸序列上的開放讀架。

```
%plotorf -opt  
Plot potential open reading frames  
Input sequence: xltfiiia.fasta  
Graph type [x11]: png  
Created plotorf.1.png
```

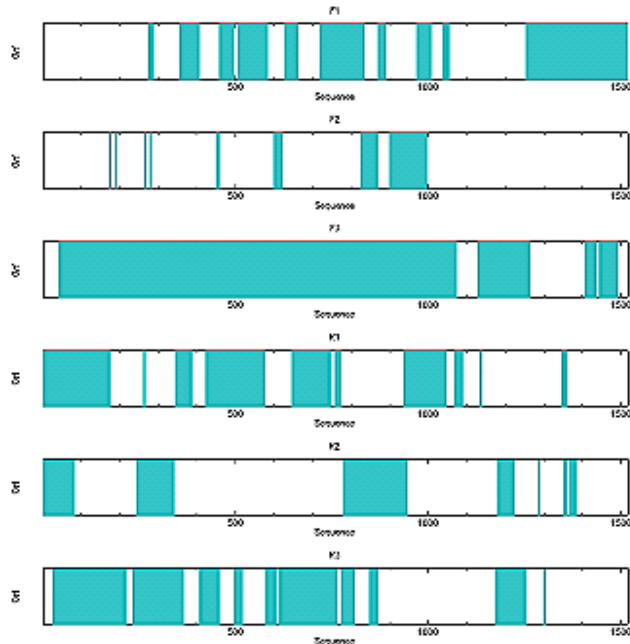
選擇以 png 格式
輸出結果

```
=====
```

```
%plotorf -opt  
Plot potential open reading frames  
Input sequence: xltfiiia.fasta  
Graph type [x11]: postscript  
Created plotorf.ps
```

選擇以 postscript
格式輸出結果

圖 7-3 TFIIIA 在核酸序列上的開放讀架的結果(在個人電腦中觀察的結果)。



結果一：以 png 格式

輸出之檔案

由上而下分別為：

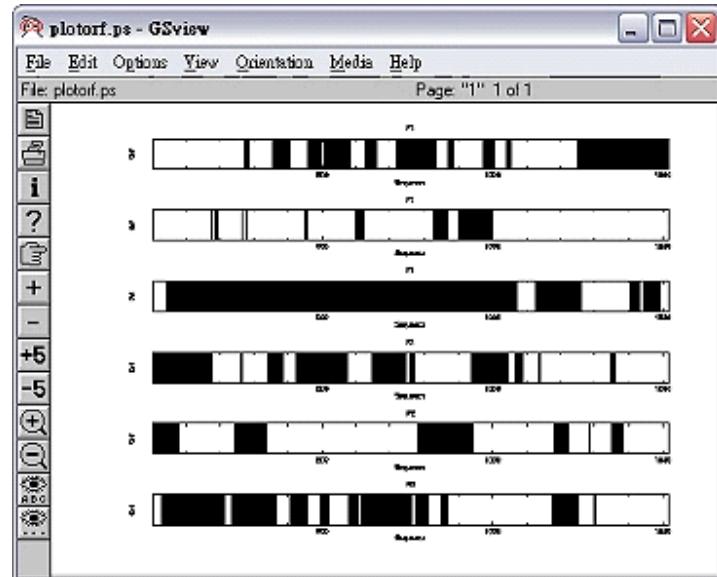
frame1, frame2,

frame3,

frame4, frame5,

frame6

著色部分代表可能為
開放讀架的區域



結果二：以 postscript

格式輸出之檔案，並

以 ghostview 開啟

由上而下分別為：

frame1, frame2,

frame3,

frame4, frame5,

frame6

著色部分代表可能為
開放讀架的區域

第七章 問題導向學習(I): 轉錄因子 TFIIIA 的選殖

練習 7-1 請重覆上述步驟(圖 7-2與圖 7-3)，並從圖中估計最可能的開放讀架的開始和結束位置。

Answer: 依第6-2頁，圖6-1的正向第三個讀架上，從序列起始，到約第1080個鹼基對的區域是最可能的，因為它不但很長，而且極少出現許多未著色的區域。

用圖形的方式雖然很容易找到最長的 ORF，可是要讀出精確的起始與結束位置卻很困難。若用人工檢視 `tfiiia.dna` 的序列，因為要顧及讀架的正確性，相當耗費精力。在下個範例中將說明怎樣用電腦找出起始與結束密碼的精確位置。這可以利用 EMBOSS 中另一個叫做 `getorf` 的程式來進行檢視，這程式可精確地找出啟始密碼 (AUG)或終止密碼(UAG, UAA, UGA) 等字串的位置，進而預測並列出核酸序列上的 ORF。

`Getorf` 程式除了直接標示開放讀架區域的核酸序列之外，亦提供轉譯的功能，還能設定尋找大於使用者給定長度的開放讀架。因為練習 7-1 中所找到的開放讀架約有 300 個胺基酸，所以讀者可以選擇只顯示長於 200 個胺基酸的開放讀架 (程式預設值為只顯示長於 30 個胺基酸的開放讀架)，以免列出太多不須要的結果。另一方面，若需要 `getorf` 程式將 ORF 結果轉譯為蛋白質序列，使用者可以選擇合適的遺傳密碼，以轉譯出正確的蛋白質序列。

範例 7-3 請找出 TFIIIA 核酸序列中最可能的開放讀架之起始密碼(AUG)與結束密碼的位置。

```
% getorf -opt
Finds and extracts open reading frames (ORFs)
Input sequence(s): xltfiiia.fasta
Genetic codes
 0 : Standard
  1 : Standard (with alternative initiation codons)
  2 : Vertebrate Mitochondrial
  3 : Yeast Mitochondrial
  4 : Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma /Spiroplasma
  5 : Invertebrate Mitochondrial
  6 : Ciliate Macronuclear and Dasycladacean
  9 : Echinoderm Mitochondrial
 10 : Euplotid Nuclear
 11 : Bacterial
 12 : Alternative Yeast Nuclear
 13 : Ascidian Mitochondrial
 14 : Flatworm Mitochondrial
 15 : Blepharisma Macronuclear
 16 : Chlorophycean Mitochondrial
 21 : Trematode Mitochondrial
 22 : Scenedesmus obliquus
 23 : Thraustochytrium Mitochondrial
Code to use [0]:
Minimum nucleotide size of ORF to report [30]: 200
Type of sequence to output
  0 : Translation of regions between STOP codons
  1 : Translation of regions between START and STOP codons
  2 : Nucleic sequences between STOP codons
  3 : Nucleic sequences between START and STOP codons
  4 : Nucleotides flanking START codons
  5 : Nucleotides flanking initial STOP codons
  6 : Nucleotides flanking ending STOP codons
Type of output [0]: 3
Output sequence [xltfiiia.orf]: tfiiia.orf
```

第七章 問題導向學習(I): 轉錄因子 TFIIB 的選殖

下圖中顯示 `getorf` 程式的部份輸出結果，程式會將符合條件的開放讀架一一列出，讀者可由列出的開放讀架之起始與終止位置，以及序列長度，同時搭配 `plotorf` 的圖形結果，來正確判斷出最可能的開放讀架及區域。因此我們可以知道，在第三個讀架 (`frame 3`)，由第 42 個鹼基對開始，至第 1073 個鹼基對，即為我們想要的結果。

圖 7-4 利用 getoff 程式尋找開放讀架的精確位置

% more **tfiiia.orf**

>XLTFIIIA_1 [42 - 1073] X.laevis 5S RNA gene transcription factor (TFIIIA)
mRNA, complete cds.
atggggagagaaggcgctgccggtgtataagcggtacatctgctcttcgcgcactgc
ggcgcgtcgttataacaagaactggaaactgcaggcgcatctgtcaaaacacacaggag
aaaccattccatgtaaaggagaaggatgtgagaaaggcttacacctcgcttcatcactta
acccgcactcactcactcatactggcgaaaaacttcacatgtgactcgatggatgt
gacttgagatttactacaaaggcaaacatgaagaagcactttacagattccataacatc
aagatctgcgtctatgtgtgcatttggagaactgtggcaaagcattcaagaaaacacaat
caattaaagggttcatcagttcagtccacacacagcagtcgcatacgaatgtcctcatgaa
ggctgtgacaaggcggtttcttgcctccgtttaaaacgtcatgaaaaaagtccatgca
ggctatccctgcaaaaggatgattcttgcatttggaaagacttggacattatac
ttgaaacacgtggcagaatgcgcattcaggacctagcgtatgtgtgtatcgaaaa
ttcaggcacaaagattacttgaggatcatcagaaaactcacgaaaaagagcgaactgtg
tatctctgcctcgagatggctgtgaccgcctataccactgcattcaatcttagaagc
catatacaatcattcatgagaaacagagacccctttgtgagcatgctggctgcggg
aatgcttgcaatgaaaaaaaggcttagaaagacattcagttgtacatgtatccagagaag
aggaaagctgaaggagaaatgcctcgcccaaagagaagcctggcttcgcctcactgg
tacataccccccaagagcaaagaaaaatgcattccgttgcggaaacagaaaagactgat
tcattgtgaaaaataagccctctggcactgaaacaaatggctattggtagataaa
ttaactata**caa**

>XLTFIIIA_2 [1252 - 1518] X.laevis 5S RNA gene transcription factor
(TFIIIA) mRN
A, complete cds.
atgcctacaggtaaaggcacagtgttatggctacataccctcttacccatgtttgct
attaaaagtgggtgcagcggccactggctgtttatttacaatacattcatttagtaag
actctgtattcatttcaaaagaatcactaaggaaatgtgcaaaattgttatcactctac
tgtaaacacaaatgtactgctgcaccctgttgggtgggctttttggggagggtgact
qaccctgttttttttaacqqaattc

由這個例子中可看出，要是沒有圖形檔的協助，就必須瀏覽文字檔，計算開放讀架的長度，不但繁瑣，而且不容易知道結果出現在第幾個讀架。接下來我們可使用 EMBOSS 的「轉譯」程式 `transeq` 做一個其他程式也可使用的蛋白質序列檔。

練習 7-2 試將TFIIIA核酸序列中最長的開放讀架轉譯為蛋白質序列，並檢查pep1與pep2是否存在於轉譯出的蛋白質序列中。

Answer:

% transeq

Translate nucleic acid sequences

第七章 問題導向學習(I): 轉錄因子 TFIIIA 的選殖

```
Input sequence(s): xltfiiia.fasta
Translation frames
 1 : 1
 2 : 2
 3 : 3
 F : Forward three frames
-1 : -1
-2 : -2
-3 : -3
 R : Reverse three frames
 6 : All six frames
Frame(s) to translate [1]:
Genetic codes
 0 : Standard
 1 : Standard (with alternative initiation codons)
 2 : Vertebrate Mitochondrial
 3 : Yeast Mitochondrial
 4 : Mold, Protozoan, Coelenterate Mitochondrial and
     Mycoplasma/Spiroplasma
 5 : Invertebrate Mitochondrial
 6 : Ciliate Macronuclear and Dasycladacean
 9 : Echinoderm Mitochondrial
10 : Euplotid Nuclear
11 : Bacterial
12 : Alternative Yeast Nuclear
13 : Ascidian Mitochondrial
14 : Flatworm Mitochondrial
15 : Blepharisma Macronuclear
16 : Chlorophycean Mitochondrial
21 : Trematode Mitochondrial
22 : Scenedesmus obliquus
23 : Thraustochytrium Mitochondrial
Code to use [0]:
Regions to translate (eg: 4-57,78-94) [1-1518]: 42-1073
Trim trailing X's and *'s [N]:
Output sequence [xltfiiia.pep]: tfiiia.pep
```

Transeq 程式的使用很容易，程式可讓使用者對核酸序列上的任一讀架進行轉譯，但要注意的是，雖說開放讀架預測在 frame3 上，但是 getorf 所找到的核酸序列位置仍是由第一個鹼基開始算起，因此使用者仍須選擇轉譯 frame1，而非 frames3(圖 7-5)。

圖 7-5 轉譯出的 TFIIIA 蛋白質序列，粗體字部份是當初用來選殖 TFIIIA 基因的兩段 peptide

```
% more tfiiia.pep
>XLTFIIIA_1 X.laevis 5S RNA gene transcription factor (TFIIIA) mRNA, complete
cds.

MGEKALPVVYKRYICSADCAGAAYNKNWKLQAHLCCKHTGEKPFPCKEEGCEKGFTSLHHL
TRHSLTHTGEKNFTCDSDGCDLRFITKANMKKHFNRFHNNIKICVYVCHFENCGKAFKKHN
QLKVHQFSHTQQLPYECPEGCDKRFSLPSRLKRHEKVHAGYPCKDDSCSFVGKTWTLY
LKHVAECHQDLAVCDVCNRKFRHKDYL RDHQKTHEKERTVYLCPRDGCDRSYTTAFNLRS
HIQSFHEEQRPFVCEAGCGKCFAMKKSLERHSVVHDPEKRKLKEKCPRPKRSLASRLTG
YIPPKSKEKNASVSGTEKDSLVKNKPSGTETNGSLVLDKLTIQ
```

四. 序列資料庫的搜尋

雖然已知 TFIIIA 是一個轉錄因子，我們並不知道這蛋白質是否有其它的功能，因此要在核酸序列資料庫中找尋是否有與新序列相似的序列。因為 FastA 程式較靈敏，用 FastA 程式仍然比用 Blast 程式分析為宜，但在搜尋與比對的速度上較慢。在美國生物技術資訊中心(NCBI)的網頁上提供了 Blast 的服務，不但提供比對的資料完整，搜尋的速度亦較 FastA 快，因此使用者也日益增多。

NCBI 網址：<http://www.ncbi.nlm.nih.gov/>

圖 7-6 Blast 的使用

The screenshot shows the NCBI BLAST homepage. On the left sidebar, there are links for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area features a yellow box titled "What's NEW in BLAST®" with the text: "NEW March 5th 2002: New database linkouts from BLAST results. Results of a BLAST search will now link sequences from the BLAST results page to the NCBI LocusLink and UniGene databases. Links to additional databases coming soon". Below this, there are two sections: "Nucleotide BLAST" and "Protein BLAST". The "Nucleotide BLAST" section lists: "Standard nucleotide-nucleotide BLAST [blastn]", "MEGABLAST", and "Search for short nearly exact matches". The "Protein BLAST" section lists: "Standard protein-protein BLAST [blastp]", "PSI- and PHI-BLAST", and "Search for short nearly exact matches". A callout box points to the "blastp" link in the Protein BLAST section with the text: "以蛋白質序列搜尋蛋白質資料庫，請選擇 blastp".

The screenshot shows the protein-protein BLAST search interface. It has tabs for "Nucleotide", "Protein" (which is selected), "Translations", and "Retrieve results for an RIO". The main search area has a large text input field with a "Search" button. Below it are fields for "Set subsequence", "From" and "To", and a dropdown menu for "Choose database" set to "nr". A checkbox "Do CD-Search" is checked. At the bottom, there are buttons for "Now: BLAST!" and "Reset query" or "Reset all". A callout box points to the "Choose database" dropdown with the text: "選擇合適的資料庫". Another callout box points to the "Search" button with the text: "將欲搜尋的序列貼於此處".

使用 NCBI Blast 時，若要指定查詢序列的範圍，只須將序列中某一段貼至「search」的欄位中，不須貼上完整的序列。

如第六章所述，Blast 會根據使用者提供的序列(即 query)與指定的資料庫，自動選擇適當的程式來做資料庫搜尋。在這一組程式中，計有 Blastx、Blastp、Blastn、tBlast、tBlastx 等子程式，在表 7-1 中列出各子程式之功能。

表 7-1 Blast 之不同子程式的運作原理

程式名	查詢序列	資料庫	備註
Blastp	Protein	protein	-
Blastn	Nucleotide	nucleotide	-
Blastx	Nucleotide	protein	轉譯查詢序列
Tblastn	protein	nucleotide	轉譯資料庫序列
Tblastx	nucleotide	nucleotide	轉譯查詢序列與資料庫序列

五. 結語

決定純化的蛋白質中的幾個片段的序列，然後將其反轉譯為核酸序列以便設計雜交用的探針，是一種典型的分子選殖方法。這一章在敘述這個過程中所需要的電腦程式工具，並由此簡介輸入序列、使用序列與執行程式的一些技巧。若不知如何使用正確的檔案格式，根本無法做任何分析。此外，找尋開放讀架似乎是一件簡單的事，知道如何選擇適合的工具能收事半功倍之效，選錯了程式就要花些時間分析結果。又如同樣是搜尋資料庫的程式，FastA 與 Blast 的性能稍不同，如果用錯程式可能會錯失可看見的結果。從這個練習的過程中，你將對 EMBOSS 套組的使用更熟悉，也能體會到第一章中說使用程式是一種技術的原因。

參考網站

<http://www.ncbi.nlm.nih.gov/>

第八章 問題導向學習(II)：轉錄因子 TFIIIA 的性質分析

楊士德¹、楊永正²

¹ 陽明生物化學研究所、² 陽明大學生物資訊所

在確定所選殖出的基因是 TFIIIA 之後，將要設計實驗以多瞭解此蛋白質的性質。在做實驗之前，希望能利用序列分析的結果來指導實驗的設計。此外、目前在序列資料庫中有許多和此序列有關的基因，也希望經由序列比對能找到一些守舊的序列，或許由此可找到和 TFIIIA 活化轉錄有關的區域。

一、模組資料庫的搜尋

在找到開放讀架後，即可利用 Translate 程式轉譯出蛋白質序列，通常會希望將蛋白質再區分出具有不同功能的區域。用 Patmatmotifs 程式去搜尋 ProSite 資料庫，可檢查在這蛋白質中是否有已知的蛋白質模組，例如鋅指，ATP 接合模組等。這一系列的問題導向學習，是以 TFIIIA 為例介紹基因分析的標準步驟，所以在搜尋「現在的」模組資料庫的過程中會找到鋅指模組。在歷史的進展中，則是先發現 TFIIIA 才找到鋅指模組的。

範例 8-1 請檢查轉譯出的 TFIIIA 序列上，有哪些 ProSite 資料庫中的已知模組。

```
%patmatmotifs -full  
Search a PROSITE motif database with a protein sequence  
  
Input sequence: tf3a.pro  
  
Output report [transcription.patmatmotifs]: tf3a.mot
```

在 TFIIIA 的序列中只能找到鋅指這一個模組樣式，圖 8-1 中是節錄的搜尋結果。它顯示在 TFIIIA 中可以找到八個 C₂H₂ 型的鋅指，其省略的部份包括鋅指發現的重要過程及已知含有這種模組樣式的序列等。接下來列出的是共有的樣式(粗體字部份)，最後再列出相關的參考資料。

圖 8-1 TFIIIA 序列中所含的蛋白質模組

```
=====  
Sequence: Transcription      from: 1      to: 366  
HitCount: 8  
Full: Yes  
Prune: Yes  
Data_file: /usr/local/EMBOSS/share/EMBOSS/data/PROSITE/prosite.lines
```

```
Length = 23
Start = position 67 of sequence
End = position 89 of sequence

Motif = ZINC_FINGER_C2H2_1
EKFPCKEEGCEKGFTSLHHLTRHSLTHTGEKN
|           |
67          89

Length = 24
Start = position 97 of sequence
End = position 120 of sequence
Motif = ZINC_FINGER_C2H2_1
EKNFTCDSDGCDLRFTTKANMKKHFNRFHNIKIC
|           |
97          120
-----
Motif: ZINC_FINGER_C2H2_1
Count: 8
*****
* Zinc finger, C2H2 type, domain signature and profile *
*****
-----
-Consensus pattern: C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
[The two C's and two H's are zinc ligands]
-----
-Last update: December 2001 / Text revised; profile added.

[ 1] Klug A., Rhodes D.
    Trends Biochem. Sci. 12:464-469(1987).
-----
```

根據方盒 8-1 中之說明，鋅指模組的樣式是在守舊的兩個 Cys 間有 2- 4 個任意的胺基酸，在第二個 Cys 後則接著三個任意的胺基酸，其後在接著一個疏水性的胺基酸，它可以是括號中八個胺基酸中的任一種。這個疏水性的胺基酸之後接著八個任意的胺基酸，最後在兩個守舊的 His 間有三至五個任意的胺基酸。不同的系統，寫模組樣式的方式略有不同，因此特別在此介紹在 GCG 環境下的規則。

方盒 8-1 在 GCG/EMBOSS 環境下，寫模組樣式的規則

1. 氨基酸用一個字母的標準代碼(one-letter code)表示。
2. 在括號中若有多個氨基酸以逗點隔開，則代表該位置可能有這幾個氨基酸存在，例如(F,V)表示在此位置可能是 Phe 或 Val。X 是代表任意的氨基酸。
3. 氨基酸後若有大括號，則代表此氨基酸重覆的次數。一般而言，大括號中有兩個數字以逗點隔開，以表示一個範圍的起點與終點，例如 CX {2,4} 表示在 Cys 後有 2 至 4 個任意的氨基酸，然後再接一個 Cys。若大括號中第一個數字不寫，則表示由零個開始，所以 { , 4 } 與 {0, 4} 代表同樣的意義。若大括號中第二個數字不寫，表示在 GCG 中可接受的最大值 350,000。也就是 {2, } 與 {2, 350,000} 的意義相同。若前後兩個數字相等，例如 {12, 12} 就表示剛好重覆那麼多次。

要讓 GCG 程式直接讀取你所寫的模組樣式，必須將其寫成一定的格式。事實上其格式與紀錄限制酵素切割位置的檔案格式相同，其中最重要的三項資訊是模組樣式的名字、平移(offset)，與模組樣式。只要用空格隔開它們即可，並不需要填在固定的位置上。其中平移一項主要是用來指定輸出結果時，標出的位置與所給定序列間之距離，通常都用「1」。例如若 zf 出現在某序列的第 100 號氨基酸，當平移值為 5 時，輸出檔就會印出 105 而不是印 100。如第 9-8 頁的範例 9-2 所示，一個檔案中可列多個模組樣式，每個模組樣式佔一列，若想在模組樣式後在加上附註，只需在附註前加“!”，程式讀到驚歎號，就會忽略其後的文字，而繼續讀下一列。在第九章中，將以實例說明其應用(第 9-8 頁的範例 9-2 以及練習 9-2)。

Prosite 並不是唯一的模組資料庫，事實上在 GCG 的環境下，還有提供 Profile 資料庫，在網路上還可找到 Block 資料庫，這些都是收集模組樣式的資料庫。

二、蛋白質性質分析

1. 二級結構分析

蛋白質分析的第一步是預測其二級結構，在此書中將不討論如何使用 Pepinfo 程式及分析其結果，而只教產生明確結果的 Garnier 程式。如第四章所述，此程式的輸出結果，目前無法找到適當的 EMBOSS 程式來呈現圖形結果。但是在蛋白質物理化學性質分析上，可用 Pepinfo 程式繪成兩種不同形式(第 6-10 頁，圖 6-8 及第 6-11 頁，圖 6-9)的圖形。

範例 8-2 請用 Garnier 程式預測 TFIIIA 蛋白質的二級結構。

```
%garnier
Predicts protein secondary structure

Input sequence(s): tf3a.pro
Output report [transcription.garnier]: tf3a.gar
```

第八章 問題導向學習(II)：轉錄因子 TFIIIA 的性質分析

Garnier 計算的結果其內容摘錄如下：

圖 8-2 Garnier 的輸出檔案格式。

這個檔案只紀錄著每一個胺基酸的可能二級結構，一般來說二級結構預測的準確度大約為 70%~80%，為了加強預測的準確度，可能必須結合其他類似的程式一起來計算，找出共通被預測的二級結構。至於其它的物理化學性質分析，必須使用 Pepinfo，因此是個非常長的檔案，不易一眼看出數據的走勢。圖形的表示法可彌補此缺點，能夠輕易看出不同與化學性質的整體分佈情形，可是不易讀到精確的位置，必須參照著文字檔看。

練習 8-1 請用 Antigenic 程式，將抗原性(antigenic index)標幟於其上(修改第6-11頁，圖6-9所示)

Answer:
%antigenic
Finds antigenic sites in proteins
Input sequence(s): tf3a.pro
Minimum length [6]:
Output report [transcription.antigenic]: tf3a.antig

第八章 問題導向學習(II)：轉錄因子 TFIIIA 的性質分析

部分結果輸出如下：

```
#=====
# Sequence: Transcription      from: 1      to: 366
# HitCount: 18
#=====

Max_score_pos at "*"

(1) Score 1.287 length 13 at residues 121->133
*  

Sequence: NIKICVYVCHFEN  

|           |  

121         133

(2) Score 1.253 length 23 at residues 200->222
*  

Sequence: TLYLKHVAECHQDLAVCDVCNRK  

|           |  

200         222

(3) Score 1.176 length 19 at residues 27->45
*  

Sequence: ALPVVYKRYICSFADCGAA  

|           |  

27          45
-----
```

事實上，我們可以將鋅指模組範圍和二級結構的訊息位置，統整到一張表格(表 8-1)找到鋅指模組和蛋白質二級結構的位置關係。在目前 EMBOSS 的程式並沒有提供像 GCG 程式 PlotStructure，可以達到如此的功能。因此，我們可以將 Patmatmotifs 和 Garnier 程式輸出結果彙整起來，找到鋅指模組和蛋白質二級結構的位置關係。

表 8-1 自 tfiiia.p2s 中所得可能具有 α -螺旋的區域。

鋅指數目	鋅指模組範圍	helix(H) α -螺旋範圍
1	13-37	10-30
2	37-59	43-46, 50-56
3	67-89	68-81
4	97-120	105-118
5	129-151	131-146
6	159-181	173-180
7	216-236	206-222
8	254-268	none
9	276-298	282-306

由這個練習中可知，要使用程式預測抗原性是非常容易的，一個不瞭解預測原理的人也可以繪出一張漂亮的圖。問題是要怎樣由繪出的圖來找出最適合做抗原的區域，這種預測的成功率又有多高？這些問題在線上輔助系統中談的很少，好在有列參考文獻供需要使用的人查閱。

2. 兩性性質分析

由表 8-1，找到鋅指模組和蛋白質二級結構的位置關係。

表 8-1 中可看出在 TFIIIA 的結構中有許多可能的 α -螺旋，幾乎每一個鋅指中都有。這些區域是否具有兩性(amphipathic)的特性，可用 Hmoment 與 Pepwheel 兩程式來驗證。在這入門的課程中將不討論前者的原理，而只用 Pepwheel 去測試 α -螺旋的特徵，在範例 8-3 中示範繪出 6-13 頁，圖 6-11 之方法。

範例 8-3 試練習使用 Hmoment 程式繪圖。

```
%hmoment -window 10 -aangle 100 -bangle 160 -baseline 0.35 -plot -graph png  
Hydrophobic moment calculation
```

```
Input sequence(s): tf3a.pro
```

由執行程式時所問的問題可看出第 6-12 頁，圖 6-11 是以曲線圖(curve graph)的形式，繪出在固定角度的疏水性矩 (Hydrophobic moment) 的大小。因為在第四個鋅指中，形成 α -螺旋的趨勢很強，因此特別看一下這一段的性質。兩種不同的預測方法所得的二級結構稍有出入，所以取整個鋅指的後半做分析的範圍。再將 α -螺旋角度做各種的調整並將曲線圖彼此相互比較，在第 137-155 個胺基酸的位置，疏水性胺基酸在縱軸的方向上分佈很分散，若繪出 Pepwheel，可能也不會集中在 α -螺旋的一側。

練習 8-2 試利用 Pepwheel 程式檢視 TFIIIA 中第 137 到 155 個胺基酸的疏水性官能團分佈特性。

Answer:

```
%pepwheel -sbegin1 137 -send1 155 -sprotein1 -wheel -steps 18 -turns 5 -amphipathic -  
nodata  
Shows protein sequences as helices
```

```
Input sequence: tf3a.pro
```

```
Graph type [x11]: png
```

上述練習的結果顯示在第 6-12 頁圖 6-10，疏水性胺基酸(以方框表示)的確呈分散分佈的狀態。如果將圖 6-11 中做出從 70° 到 180° 的範圍，可看出在第 90 到 100 個胺基酸附近都有較大的疏水性矩，只是這一區域不是 α -螺旋，因此不適合用 Pepwheel 的預設值作分析。

三、多序列排比

1. 自動分析 – Emma

在前面章節的分析中顯示序列資料庫中有許多與 TFIIIA 相似的序列（第三章與第五章）。雖然這些序列都是在非洲有爪水生蛙的 TFIIIA 被選殖出來之後才找到的，在此我們假設 *Xenopus* 的 TFIIIA 是比較後期才找到的，以便現在討論多序列排比的應用。

範例 8-4 請利用 Emma 程式，將所有與 TFIIIA 有關的 DNA 序列做多序列排比。

```
%emma -osformat msf
Input sequence(s): @tf3adna.list
Output sequence [xbb2aa.aln]: tf3adna.msf
Output file [xbb2aa.dnd]: tf3adna.dnd
..clustalw -infile=7879A -outfile=7879B -align -type=dna -output=gcg -
pwdnamatrix=iub -pwgapopen=10.000 -pwgapext=0.100 -newtree=7879C -
dnamatrix= -gapopen=10.000 -gapext=5.000 -gapdist=8 -hgapresidues=GPSNDQEKR
-maxdiv=30..

CLUSTAL W (1.81) Multiple Sequence Alignments

Sequence type explicitly set to DNA
Sequence format is Pearson
Sequence 1: XBB2AA           1753 bp
Sequence 2: XBFINAA          1735 bp
Sequence 3: XBFINAB          1331 bp
Sequence 4: XBTF3A           1377 bp
Sequence 5: XLFINAC          1428 bp
Sequence 6: XLTFIIIA         1518 bp
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 49
Sequences (1:3) Aligned. Score: 54
Sequences (1:4) Aligned. Score: 53
Sequences (1:5) Aligned. Score: 55
Sequences (1:6) Aligned. Score: 53
Sequences (2:3) Aligned. Score: 56
Sequences (2:4) Aligned. Score: 53
Sequences (2:5) Aligned. Score: 85
Sequences (2:6) Aligned. Score: 53
Sequences (3:4) Aligned. Score: 98
Sequences (3:5) Aligned. Score: 54
Sequences (3:6) Aligned. Score: 86
Sequences (4:5) Aligned. Score: 52
Sequences (4:6) Aligned. Score: 84
Sequences (5:6) Aligned. Score: 50
Guide tree      file created: [7879C]
Start of Multiple Alignment
There are 5 groups
Aligning...
Group 1: Sequences: 2      Score:24914
Group 2: Sequences: 3      Score:22196
Group 3: Sequences: 4      Score:13770
Group 4: Sequences: 2      Score:23671
Group 5: Sequences: 6      Score:12405
Alignment Score 63377
GCG-Alignment file created      [7879B]
```

在這個範例之中，引入了一個新的指定序列的方法，因為需要同時指定多個序列檔，所以在檔名之前要加一個「@」(唸做「at」)，以表示這檔案的特殊性。在這檔案中列出各序列貯存的路徑與檔名，因此稱之為「檔名檔案」(參閱方盒 8-2)，程式能根據這檔的內容而自動尋找所需的序列檔。

方盒 8-2 檔名檔案(list file)的產生與應用

在安裝 EMBOSS 的電腦上，每個使用者可用的磁碟空間都有限制，若是每次要做序列排比時都要把序列取回自己的工作子目錄，不但要花時間，而且很容易超出磁碟空間的配額(quota)，以至於無法工作。一個最好的辦法就是直接指定資料庫中的序列，甚至分析的範圍，以便讓程式自動去公用的磁碟機取用序列資料。這種紀錄序列檔名字的檔案稱之為「檔名檔案」。

有一些程式可以自動產生檔名檔案，以便做後續分析，例如第五章中所述 FastA 的輸出檔案可作為檔名檔，而不需要使用者自己鍵入。可是有時產生的檔案並不理想在後續的 Emma 分析時可能造成困擾。若你不想將這些檔名刪除，可在該筆資料前加上「！」。所有的 GCG/ EMBOSS 程式看到驚嘆號後，會自動跳過其後的資料而繼續讀下一行。因此驚嘆號也可被用來在檔案中加註解。

詳細 List File 的使用說明可參照 (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Doc/Tutorial/node15.html>)。也許你會問，對一個初學者而言，怎麼會知道哪些程式的輸出檔案可以做為檔名檔呢？這有兩種方法：

- (1) 有需要使用檔名檔時，追問這個程式的前一步驟是否可產生檔名檔，因此去查閱前一程式的輔助說明，再做一下測試。
- (2) 在學習其他程式時，在線上輔助系統中順便瀏覽一下其他有趣的功能，比如在「User Documentation (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/userdoc.html>)」下的「Report Format (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Themes/Report Formats.html>)」，就有討論哪些程式可產生檔名檔或特定的格式，要產生時需加何種指令等。

在讀入序列後，Emma 就會開始運作，若使用者希望知道序列排比時相互比較的順序，可繪一張如第 6-7 頁的圖 6-7A，此圖並不代表真正的親緣關係，只代表序列排比的先後次序。好在 EMBOSS 套組中有一系列的程式可用來分析親緣關係，但在此入門課程中不做深入的介紹。此外，程式也會產生一個副檔名為 msf 的輸出檔，其格式如圖 8-3 所示。

圖 8-3 msf 檔案格式。

```
NA_MULTIPLE_ALIGNMENT 1.0

tf3adna.aln MSF: 2067 Type: N 09/12/02 CompCheck: 8954 ..

Name: XBFINAB Len: 2067 Check: 2735 Weight: 8.50
Name: XBTF3A Len: 2067 Check: 9253 Weight: 10.90
Name: XLTFIIIA Len: 2067 Check: 4049 Weight: 15.80
Name: XBB2AA Len: 2067 Check: 7157 Weight: 28.00
Name: XBFINAA Len: 2067 Check: 2600 Weight: 18.20
Name: XLFINAC Len: 2067 Check: 3160 Weight: 18.20
//                                1                               50
XBFINAB ~~~~~
XBTF3A ~~~~~
XLTFIIIA ~~~~~
XBB2AA ~~~~~
XBFINAA CCAACTTTGATCAGGGTCACCTGGGTGCCTGTGATTAGGATCC
XLFINAC ~~~~~
```

做蛋白質多序列排比時，所用的參數與核酸分析時用的不同，程式會自動建議數值，請做練習 8-3來比較一下哪些參數不同。

練習 8-3 請利用 Emma 程式，將所有 Swiss-Prot 中與 TFIIIA 有關的蛋白質序列做多序列排比，並比較所用的 penalty 與 DNA 多序列排比(第 8-7 頁，範例 8-4)之差別。

Answer:

```
%emma -osformat msf
Multiple alignment program - interface to ClustalW program
Input sequence(s): swissprot:tf3a_*
Output sequence [tf3a_bufam.aln]: tf3apro.aln
Output file [tf3a_bufam.dnd]: tf3apro.dnd
..clustalw -infile=8395A -outfile=8395B -align -type=protein -output=gcg -
pwmatrix=blosum -pwgapopen=10.000 -pwgapext=0.100 -newtree=8395C -
matrix=blosum -gapopen=10.000 -gapext=5.000 -gapdist=8 -
hgapresidues=GPSNDQEKR -maxdiv=30..
```

```
CLUSTAL W (1.81) Multiple Sequence Alignments

Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: TF3A_BUFA 339 aa
Sequence 2: TF3A_HUMAN 423 aa
Sequence 3: TF3A_ICTPU 322 aa
Sequence 4: TF3A_RANPI 335 aa
Sequence 5: TF3A_XENBO 339 aa
Sequence 6: TF3A_XENLA 366 aa
Sequence 7: TF3A_YEAST 429 aa
Start of Pairwise alignments
Aligning...
-----
```

2. 顯示共有序列 - Showalign

使用 Emma 雖能將序列並列比較，可是在 msf 的格式中(參見圖 8-3)，少數有差異的序列位置不易被觀察到，此外也未列出共有(consensus)序列，這問題可用另一程式 Showalign 解決。如範例 8-5所示，在使用 Showalign 時只要在指令行上加上「-show N」與「-show I」的選項，即可產生如第 6-4 頁圖 6-4 之格式(不過預設值為「-show N」及「Non-identities between the sequences」)。

範例 8-5 請利用 Showalign 程式，將TFIIIA的蛋白質多序列分析結果表示成第6-4頁，圖6-4之形式。

```
%showalign  
Displays a multiple sequence alignment  
Input sequence set: tf3adna.msf  
Output file [tf3adna.showalign]:
```

在使用此程式時是以 Emma 的輸出檔「tf3adna.msf」為輸入資料，因為在此檔中有許多個序列，在給檔名時必須在檔名後加上「{*}」，以表示要分析此檔中所有的序列，例如「tf3asw.msf{*}」。如果只要分析其中的某幾個序列，可將序列名稱寫在大括號中。因為在 Emma 中已將序列排好，在使用 showalign 程式時只要計算共有序列，並重新將序列格式化，所以速度很快。在計算共有序列時，不同的序列可以有不同的「重量 (weight)」，這樣一些不重要或不確定的序列就比較不會影響到共有序列。若以「tf3asw .msf」為輸入資料，則可在檔頭中各序列的名字後指定其比重，在執行程式時程式會根據序列數目，建議使用者"plurality"的設定，此值會影響共有序列的產生。在運算時，若只有小於 plurality 的序列數相同，則不寫出共有序列，而以「-」代表(參閱第 6-4 頁圖 6-4)。

在序列差異很大時，我們希望要很快能看到多個序列完全相同之位置，此時就不宜使用「-show N」，而應使用「-show I」。若每一序列的某一位置，序列完全相同，則程式會在共有序列這一行中以「*」來表示，因而在眾多不同的序列中突顯出相同之處。請根據此觀念做下列練習。

練習 8-4 請將TFIIIA蛋白質的多序列分析結果表示成第6-4頁，圖6-3之形式。

```
Answer:  
%showalign  
Displays a multiple sequence alignment  
Input sequence set: tf3apro.msf  
Output file [tf3apro.showalign]:
```

3. 人工分析 - MSE

多序列排比只是根據給定的運算法則，計算出最符合這些法則的結果。如果法則給的不恰當，就不可能把相似或相同的序列完全對齊。例如，在練習 8-3與範例 8-4的比較中

可看到蛋白質與核酸序列列使用不同的計分表(scoring table)和空隙分離距離(gap separation distance)方法，使用錯誤就不可能得到好的結果。即使使用正確，也不保證就能完全對齊，因此需要人工調整。MSE(Multiple Sequence Editor-Jalview)程式是一個多序列編輯器，此工具只有在 Jemboss 的環境介面下使用，而在 EMBOSS 這套工具組並沒有提供相關的程式，在本次的介紹不納入說明。

對初學者而言，要將很長的序列人工並列在一起是很困難的，因此，在這一章中並不提供任何練習。

四、結語

這一章中簡單地介紹了 6-13 頁，圖 6-12 中，基因分析標準步驟的主要部分。在第四章中只是建立一個架構，讓大家知道在怎樣的狀況下，應使用怎樣的程式。在這一章中，以實例介紹幾個主要程式的用法。在這一章中，以實例介紹幾個主要程式，希望藉此讓讀者更加了解 EMBOSS 套組的使用方法。。

參考網站

<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/userdoc.html>

第九章 問題導向學習(III)：鋅指模組的發現與分析

蔡健偉¹、楊永正²

¹ 賽亞公司生物資訊部、² 陽明大學生物資訊研究所

一、重覆序列的尋找

在基因體 DNA、傳訊者核糖核酸、或蛋白質上重覆出現的序列，通常是一些具有功能的區域。英國 MRC 實驗室的 McLachlan 對蛋白質中重覆出現的序列(repeats)很感興趣，因此他注意到在已發表的 TFIIB 序列中似乎有一些重覆的序列。為了進一步確認這觀察，他就採用點矩陣(dot matrix)運算法(algorithm)來尋找這些重覆區域的位置。點矩陣的概念很簡單，在圖 9-1 中將相同的序列分別寫在垂直方向與水平方向的座標軸上，若兩序列上有相同的核苷酸，就在座標交錯的欄位中填入「1」，若不相同則留空。在分析完成後，在圖中對角線的位置全部是「1」，這是因為在縱軸與橫軸上放的是相同的序列所致。有趣的是在對角線之外的區域也可以看到一些與對角線平行的短線，(圖中粗體字部份)。這些線代表在序列中有一小段序列是重覆的。例如 GCGT(在橫軸序列上用底線標示處)在序列中連續出現兩次，橫軸上第一個 GCGT 與縱軸序列上的第一個 GCGT 序列相同，所以會在對角線上出現；橫軸上第二個 GCGT 與縱軸序列中的第一個 GCGT 序列也相同，因此在離開對角線的位置另有一條平行於對角線的短線出現(粗體字)。因此若比較完全相同的序列時，所畫出點矩陣中若有離開對角線的短線，那就代表序列中有重覆的區域。如果在兩軸上放的是不同的序列，則平行於對角線的線，所代表的是相似的序列(homologous sequence)，對角線則不會出現。

在 EMBOSS 中的「Alignment dot plots」項目下，有 Dotmatcher、Dotpath、Dottup 與 Polydot 等四個程式是運用點矩陣的觀念來比較兩序列的相似性，並將結果以圖形的形式呈現出來。上列程式包含有兩種運算的方式，一個是比對速度慢卻靈敏的 window/ threshold 法，代表程式為 Dotmatcher；另一種是較速度快的 word match 法，但較不靈敏，其餘三種皆屬此種方法，代表程式為 Dottup。在此入門課程中不討論運算法，僅建議大家使用 window/ threshold 法。在提出鋅指結構的原始文獻⁽¹⁾中，雖不是利用 EMBOSS 套組做的，這些程式卻可達到同樣的目的，範例 9-1 即是利用此法尋找 TFIIB 中的重覆序列。

圖 9-1 以「點矩陣」法尋找重覆序列

	A	T	G	C	G	T	A	C	G	C	G	T	G	A	C
A	1						1								1
T		1				1								1	
G			1	1				1	1	1			1		
C				1			1		1						1
G					1			1	1	1			1		
T						1						1			
A							1							1	
C								1	1						1
G									1	1					1
C										1					1
G											1	1			
T												1			
G													1		
A														1	
C															1

方盒 9-1 序列重覆出現的意義

基因體 DNA 雖具有保存遺傳特性的重責，仍須允許有限度的變異 (variation)，以維持生物族群的多樣性 (diversity)。基因突變 (mutation)、複製 (duplication)、重組 (recombination)、--- 等都是產生變異的機制。在基因重組的過程中，如果生物巨分子不能維持執行功能所需的結構，就有可能會被淘汰，因為基因重組所發生的位置並不固定。在經過長時間的演化後，具有功能的區域在折疊成三級結構時，就會逐漸具有獨立折疊 (independent folding) 的性質。這是因為當一生物巨分子重要功能區域具有 IFM 性質時，如在此一功能區域外部發生突變，功能區域因具 IFM 性質，其結構較不受其周遭的突變所影響，功能仍能維持；只有當功能區域內部產生突變時，功能區域結構才有較大之機率遭到破壞，致使功能喪失而遭淘汰。而當生物巨分子重要功能區域不具有 IFM 性質時，在此區域外的突變較可能造成功能區域結構的破壞，而使此一基因遭到淘汰。兩相比較之下，在生物巨分子中具有 IFM 性質的功能區域實具有較高的突變耐受力，在演化的過程中要比不具 IFM 性質的功能區域要佔優勢。長久天擇下來，生物巨分子中具有重要功能區域將逐漸得到 IFM 的性質。

這些獨立折疊的區域不因周圍序列的不同而影響其功能，所以即使被重組到其他的位置，仍能執行其原有的功能，例如接合 DNA 或 ATP 等。而且它周圍的序列發生突變時，也不易影響到這些獨立折疊單元 (IFM, independent folding motif) 的結構與功能。更重要的是在有了許多不同種的 IFM 之後，生物體可以利用組合的方式產生各種新的蛋白質，或是增加族群的多樣性。當我們比較多個不同的蛋白質序列時，有時會發現一小段相似的序列。這些序列通常都具有功能，而且是 IFM，一般稱之為模組樣式 (pattern)。在 DNA 序列上也可能有重覆序列，只是它們的功能不是以三級結構表現的，通常是蛋白質的辨識位置。因此巨分子上的重覆序列所代表的是一些可能具有功能的區域。

此處的策略是在縱軸與橫軸放置同樣的序列，希望能找到一些與對角線平行的短線。使用圖形分析的好處是一次可看到整個蛋白質序列 (1 至 344) 的分析結果，不像圖 9-1 只看到 15 個核苷酸的比較而已。

範例 9-1 利用點矩陣法來尋找TFIIIA中的重覆序列

```
%dotmatcher -sask -opt
Displays a thresholded dotplot of two sequences
Input sequence: tf3a.pro
Begin at position [start]:
End at position [end]:
Second sequence: tf3a.pro
Begin at position [start]:
End at position [end]:
window size over which to test threshhold [10]:
threshold [50]:
Matrix file [EBLOSUM62]:
Display as data [N]:
Graph type [x11]: ps
Created dotmatcher.ps
```

在這個範例中，有兩個特別的名詞會影響到觀察到的結果，一個是 Window size。另一個是 threshold。前者的主要目的是希望表現一個區域的平均特性，後者相當於一個可濾掉雜訊的過濾器。如圖 9-2 所示，若訊號變化很大，則不易看出曲線的走勢。若以 5 個數據點做一個 window，將平均值記錄在第一個數據點的位置；然後在將 window 移動一個數據點，將平均值記錄在第二個數據點；依此類推，即可畫出一條較平滑，可看出走勢的曲線。可是若將 window 放大到 25，則原有的一些訊號就被平均掉了。在 Window size 接近重覆區域的大小時，訊號最能真實反應重覆區域間的相似性，問題是做這類型的分析時並不知到相似的區域有多長，通常必須試用數個不同的條件。鋅指的長度大約為 30 個胺基酸，因此在範例中以 30 為 Window 大小是很恰當的。

因為有些胺基酸性質相似，在演化的過程中可互相置換而不影響功能，因此比較蛋白質序列時所用的計分方式和核酸序列的不同。目前較常用的方式是 BLOSUM62 矩陣，在這入門課程中不討論其由來，只要知道某些在演化上非常守舊的胺基酸(例如 Cys)相同時，得分很高(9 分)，一些可由其他胺基酸取代的胺基酸(例如 Ala)相同時得分較低(0 分)。若是 Ala 被置換為 Leu 時，則變成負分(-1 分)。程式內部會根據你所選取的 window size 建議一個 threshold。想瞭解這名詞的意義，可將點矩陣中每一格的得分視為在 Z 軸方向的高度，而每一格的分數則是以這一格為中心，以 Window 大小為範圍的數個沿對角線方向的格子的平均值。Threshold 就相當於以垂直於 Z 軸的不同平面切割這三維空間的立體圖，而比平面高的那些點就會紀錄在 Dotmatcher 所繪出的圖上，根據這種想法，threshold 越高，能看到的點與線就越少。在圖 9-3 中，我們將可看到不同 threshold 的效果。

圖 9-2 利用移動窗使曲線變平滑的示意圖

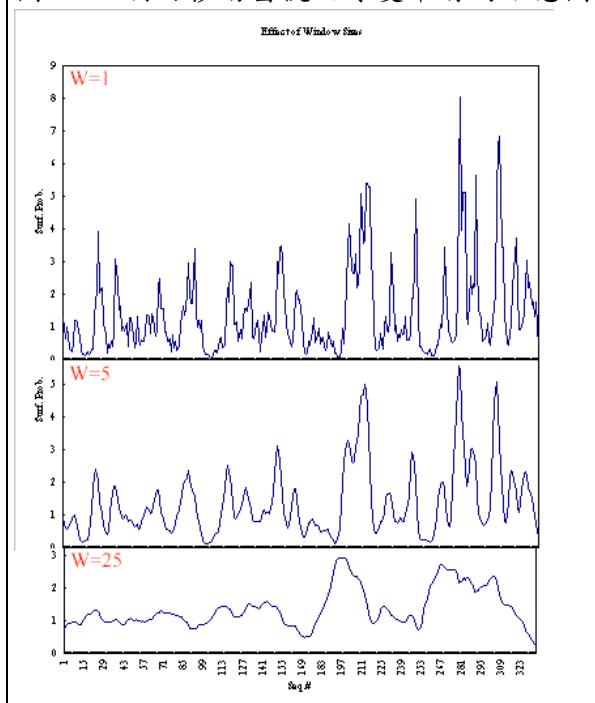
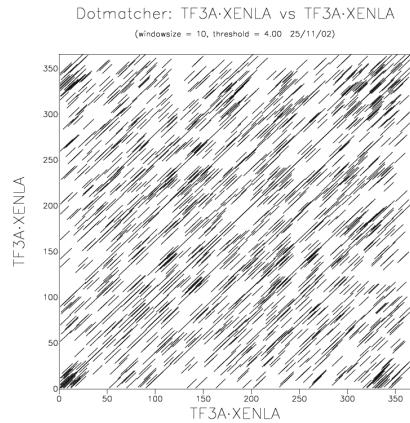
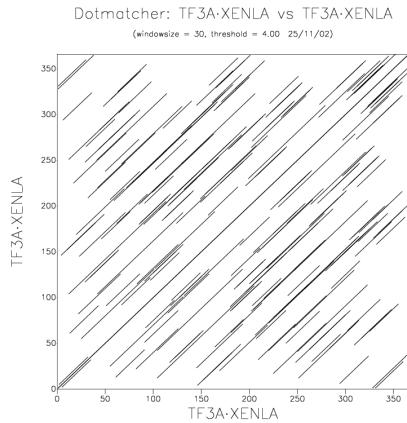


圖 9-3 不同 Window size 與 Threshold 對點矩陣表示法的效應

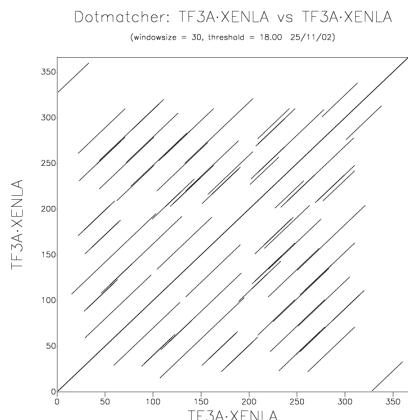
$W=10 \quad T=4$



$W=30 \quad T=4$



$W=30 \quad T=18$



$W=30 \quad T=50$

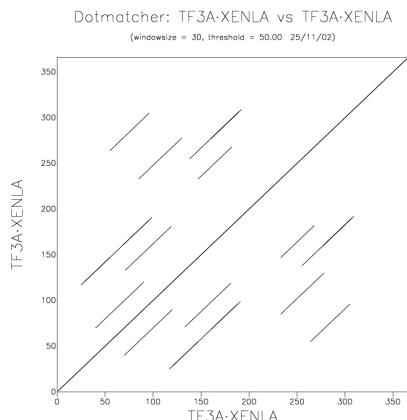


圖 9-3 列出四種不同比較條件的圖，讓讀者比較一下 window size 與 threshold 變化時的效應。在 threshold 提高時，雜訊降低，較易看清楚代表訊號的線，可是相對地也損失了一些訊號。反之，threshold 低時訊號雖很多，但與雜訊混在一起，比較不易區辨訊號的真偽，同時，重覆序列的長度似乎也長一些。Window size 的效果與圖 9-2 中所示的相仿。在這個例子中，讀者可很清處地看到調整參數有助於結果的判讀。若只用一種條件，不易判定那一條線才是真的訊號，也不清楚重覆序列的大小。

圖 9-4 對 TFIIIA 內重覆序列做序列排比所需的檔案

```
% more zf.lst
# searching for: "tfiiia"
Swissprot:Tf3a_Xenla[35:67]
```

```
Swissprot:Tf3a_Xenla[65:98]
Swissprot:Tf3a_Xenla[96:129]
Swissprot:Tf3a_Xenla[127:159]
Swissprot:Tf3a_Xenla[157:188]
Swissprot:Tf3a_Xenla[186:214]
Swissprot:Tf3a_Xenla[212:246]
Swissprot:Tf3a_Xenla[244:276]
```

根據 Dotmatcher 中各重覆序列的大概範圍，可進一步對這些重覆序列做序列排比，以便尋找共有序列。在第五章所示的序列排比都是不同序列之比對，要如何將同一序列的不同部份互相比對呢？如圖 9-4 所示，只要在序列名後加上給定的範圍[Begin:End]。因為這裏的目的是比較同一序列的不同位置，所以前面的序列名稱全部是 TFIIIA 的蛋白質序列。照理說應採用自己組合的序列來做，可是因為 TFIIIA 的序列已發表了，所以此處直接取用資料庫中的序列做練習，以免引入不必要的錯誤。而 zf.lst 檔可利用 vi 編輯器製作，或在個人電腦上寫好後再傳送過來使用。請利用圖 9-4 中的檔名檔 zf.lst 及 Emma 程式做多序列排比。

練習 9-1 請利用多序列排比顯示TFIIIA中各重覆序列的相似性。

```
% emma -opt

Multiple alignment program - interface to ClustalW program
Input sequence(s): @zf.lst
Do you want to produce only the dendrogram file? [No]: 
Do you want to use an old dendrogram file? [No]: 
Insist that the sequence type is changed to protein [No]: Y
Do you want to carry out slow or fast pairwise alignment
s : slow
f : fast
Please select one [s]:
Input value for gap open penalty [10.0]: 2
Input value for gap extension penalty [0.1]:
DNA pairwise alignment matrix options
i : iub
c : clustalw
: own
Select matrix [i]:
Nucleotide multiple alignment matrix options
i : iub
c : clustalw
: own
Select matrix [i]:
Enter gap penalty [10.0]:
Enter variable gap penalty [5.0]:
Use end gap separation penalty [Yes]:
Gap separation distance [8]:
Cut-off to delay the alignment of the most divergent sequences [30]:
Output sequence [tf3a_xenla.aln]: zf210.aln
Output file [tf3a_xenla.dnd]:
..clustalw -infile=22234A -outfile=22234B -align -type=dna -output=gcg
-pwdnamatrix=iub -pwgapopen=2.000 -pwgapext=0.100 -newtr
e=22234C -dnamatrix= -gapopen=10.000 -gapext=5.000 -gapdist=8
-hgapresidues=GPSNDQEKR -maxdiv=30..
```

```
CLUSTAL W (1.8) Multiple Sequence Alignments
```

```
Sequence type explicitly set to DNA
Sequence format is Pearson
ERROR: Multiple sequences found with same name, TF3A_XENLA (first 30 chars are
significant)
```

在這練習中，讀者會發現程式會在螢幕上顯示錯誤訊息。這是因為在輸出檔案中，Emma 程式會在序列左側寫出檔名。而我們所給的每一個序列都是 TFIIIA 序列，因此程式會告訴使用者檔名並不是唯一的，同時程式也自動在檔名後加上序號做區別，因此並不影響結果的正確性。

圖 9-5 TFIIIA 內各重覆序列的序列排比結果

```
% more zf210.showalign
```

```
10      20      30      40      50      60
-----|-----|-----|-----|-----|-----|-----|
Consensus  xxcccxxcxKXfxxxXx1kxhxxxxxxekxxxXXXX
TF3A_XENLA_4 -T.DS.G-.DLR.TTKANm.K.FNRFHNI.ICVYVC
TF3A_XENLA_9 -L.PR.G-.DrSyTTAFN.rS.IQSFHE.qR-PFVC
TF3A_XENLA_5 YV.HFeN-.G.A.KKHNQ..V.QFSHTQqLPYEC--
TF3A_XENLA_6 YE.PHeG-.D.R.SLPSR..R.EKVHAGYPCKK---
TF3A_XENLA_3 FP.KEeG-.E.G.TSLHH.TR.SLTHTG..NFTCD-
TF3A_XENLA_2 YI.SFAD-.GAAYNKNWK.qA.LCKHTG..PFPC--
TF3A_XENLA_8 DLAVC.V-.NrK.RHKDY.rD.QKTHEK.rTVYLCP
TF3A_XENLA_7 --.KK.DS.SFVGKTWTLYLK.VAECHQdLA-----
Consensus  xxcccxxcxKXfxxxXx1kxhxxxxxxekxxxXXXX
```

在上圖中 Emma 的結果並不理想，如果以第六章中所述的 MSE 程式處理 Emma 程式的輸出，則可將未對齊之處以人工方式對齊(圖 9-6)。

圖 9-6 Mse 後的鋅指序列排比。

```
% more zf210.pre
```

```
10      20      30      40      50      60
-----|-----|-----|-----|-----|-----|-----|
Consensus  xxcccxxcxKXfxxxXx1kxhxxxxxxekxxxXXXX
TF3A_XENLA_4 -T.DS.G-.DLR.TTKANm.K.FNRF.NI.ICVYVC
TF3A_XENLA_9 -L.PR.G-.DrSyTTAFN.rS.IQSF.E.qR-PFVC
TF3A_XENLA_5 YV.HFeN-.G.A.KKHNQ..V.QFS-.TQqLPYEC-
TF3A_XENLA_6 YE.PHeG-.D.R.SLPSR..R.-EKV.AGYPCKK--
TF3A_XENLA_3 FP.KEeG-.E.G.TSLHH.TR.-SLT.TG..NFTCD
TF3A_XENLA_2 YI.SFAD-.GAAYNKNWK.qA.-LCK.TG..PFPC-
TF3A_XENLA_8 AV.--.V-.NrK.RHKDY.rD.QKT-.EK.rTVYLCP
TF3A_XENLA_7 --.KK.DS.SFVGKTWTLYLK.VAECHQdLA-----
Consensus  xxcccxxCxKXfxxxXx1kxHxxxxxHxeKxxxXXXX
```

在決定出 TFIIIA 基因序列的前一年，吳成文院士的實驗室發現在 TFIIIA 的蛋白質中有鋅離子存在⁽²⁾。學過普化的人都知道過渡性金屬離子很容易形成錯合物，而 Cys 中的硫與 His 中的氮都有成為鋅離子的配位子(ligand)的潛力。因此在 MRC 的一群科學家就將 TF IIIA 中守舊的 Cys 與 His 和鋅連想到一起。它們提出一個如圖 9-7 所示的結構模型，它們並據此建議了 TFIIIA 和 DNA 交互作用的可能模式(圖 9-8)。因為在它們所提出的鋅指模中，N

圖 9-7 MRC 研究群提出的鋅指結構

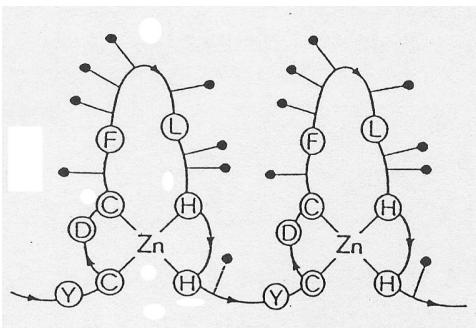
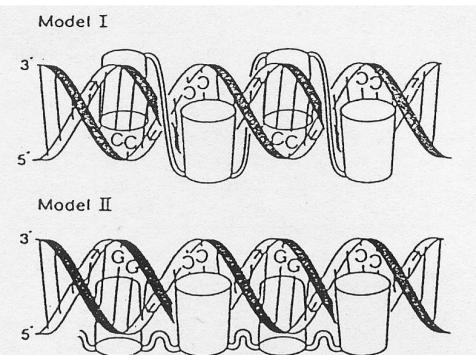


圖 9-8 鋅指與 DNA 交互作用的可能方式



端到 C 端的方向在鋅指啟始和結束處是反向的(圖 9-7，箭頭)。若想在不扭曲結構的前提下，將連續的幾個鋅指嵌入 DNA 的主溝槽中，就必須採交錯嵌入的方式。在圖 9-8 所示的兩種可能性中，它們認為第二種作用方式較可能。

二、以模組搜尋序列資料庫

在 MRC 的研究群提出鋅指的結構後，美國約翰霍普金斯大學的 Berg，首先用這一段重覆的序列搜尋資料庫，找出當時在資料庫中含有這種重覆的蛋白質。問題是在這重覆序列中變異很大，守舊的只有 Cys 和 His 兩個胺基酸、幾個疏水性的胺基酸及他們的相對位置(參閱圖 9-5)。想在資料庫中找到符合這條件的序列，就必須有辦法將這些共有的特徵用有系統的表示出來。這些共同特徵稱之為模組樣式，有人也稱其為意涵序列(signature)。這就像是汽車有大，有小，有不同之形狀與顏色，可是我們一看到它的一些特徵就知道它是汽車。尋找模組樣式與尋找限制酵素切割位置的原理是相同。有些限制酵素可辨識幾種不同的序列，因此在表示切割位置時必須有某些彈性。換言之，程式不是做簡單的字串搜尋(string search)，而是在尋找一定的模組樣式。

表示模組樣式有兩大要件，一個是在某位置上胺基酸的類別，一個是這些胺基酸出現的次數。只要掌握住這兩大要件，各胺基酸間的相對位置與距離就自然呈現出來了。可是在 TFIIIA 中的重覆序列，在特定位置的胺基酸種類不止一個，而且重覆的次數在各重覆序列上也不完全相同，必須有一些描述它們的規則。不同的軟體可能用不同的符號來表示模組樣式，但是規則都是一樣的，EMBOSS 環境下所使用的符號與規則如下所述。

基本上表示模組樣式的規則與 PROSITE 資料庫的表示方式相同：

胺基酸使用標準 IUPAC one-letter code 來表示。

符號'X'代表此一位置可為任何一個胺基酸。

中括弧'[]'中的符號代表此一位置可接受之胺基酸種類。舉例來說：[ALT]代表 Ala 或 Leu 或 Thr。

大括弧'{ }'中的符號代表此一位置「不」接受之胺基酸種類。舉例來說：{AM}代表接受 Ala 或 Met 以外的任一胺基酸。

每一個位置皆以'-'隔開。

重複序列的表示是在胺基酸右手邊小括弧'()'中，以數字表示次數或次數範圍。舉例來說： $X(3)$ 代表 $X-X-X$ ，而 $X(2,4)$ 則代表 $X-X$ 或 $X-X-X$ 或 $X-X-X-X$ 。

若要將模組樣式限制在序列的 N-端或 C-端上，則以'<'符號開頭來代表 N-端或以'>'符號結尾來代表此模組樣式位於 C-端。

Berg 當時並沒有很注重細節，只是將他認為重要的一些特徵表示出來。範例 9-2 將 Berg 所列出的模組樣式寫成 EMBOSS 的格式。

範例 9-2 請將Berg所寫的模組樣式表示為EMBOSS的格式

Name	Pattern
C2H2	C-X(2,4)-C-X(2,15)-H-X(2,4)-H
C2C2	C-X(2,4)-C-X(2,15)-C-X(2,4)-C
H2C2	H-X(2,4)-H-X(2,15)-C-X(2,4)-C

若以 C_2H_2 為例，寫出的模組樣式表示在 Cys 後有二到四個任何的胺基酸，然後再接 Cys；在這 Cys 與下一個守舊的 His 之間，有二到十五個任何的胺基酸；在 His 與下一個守舊的 His 之間，有二到四個任何的胺基酸。

練習 9-2 試根據圖 9-5，將TFIIIA中各重覆序列的共有序列用符號表示出來。

Answer:

Name	Pattern
TFIIIA	C-X(2,5)-C-X(12)-H-X(2,4)-H

若在 EMBOSS 的環境下要看那些資料庫中的序列含有上述的模組樣式，可使用「Nucleic motifs」或「Protein motifs」這類別下的程式 Fuzznuc、Fuzzpro 以及 Fuzztran。因為 EMBOSS 的環境下無法同時指定 PIR 與 SwissProt 兩個資料庫，每檢視一個資料庫序列是否含有欲搜尋的模組樣式，就必須再執行一次程式。或者將要搜尋的資料庫或序列檔以 USA 的方式寫在一個檔名檔中，接著才執行 Fuzzpro 程式。

範例 9-3 以Berg的模組樣式搜尋Swiss-Prot序列資料庫

```
%fuzzpro -opt &

Protein pattern search
Input sequence(s): swissprot:tf3a_*
Search pattern: C-X(2,4)-C-X(2,15)-H-X(2,4)-H
Number of mismatches [0]:
Output report [tf3a_bufam.fuzzpro]: c2h2.fuzzpro
```

為避免在螢幕前久候，在範例中是將工作送到背景執行(&)。

在圖 9-9 中顯示的是利用 C_2H_2 為模組樣式所搜尋到的部份結果。這個檔案有 1050Kbyte 若印到 A4 紙上要印 400 多頁，像這樣大的資料量顯然必須用電腦來分析。

圖 9-9 利用 C_2H_2 為模組樣式所搜尋到的節錄結果

```
#####
# Program: fuzzpro
# Rundate: Wed Nov 27 15:02:55 2002
# Report_format: seqtable
# Report_file: tf3a_bufam.fuzzpro
#####

=====
#
# Sequence: TF3A_BUFAM      from: 1    to: 339
# HitCount: 11
#
# Pattern: C-X(2,4)-C-X(2,15)-H-X(2,4)-H
# Mismatch: 0
#
=====

Start      End Mismatch Sequence
14          36      . CSFPDCNATYNKNRKLQAHLCKH
44          62      . CTYEGCEKGFTLHHLNRH
44          62      . CTYEGCEKGFTLHHLNRH
44          66      . CTYEGCEKGFTLHHLNRHVLSH
74          97      . CETENCNLAFTTASNMRHLHFRAH
106         128     . CYFADCGQQFRKHNQLKIHQYIH
136         158     . CSHEGCDKCYASPSRLKRHEKTH
141         158     . CDKCYASPSRLKRHEKTH
192         212     . CSICNRTFKRKSFLKEHKKIH
221         244     . CPRENCDRTYTTKFNLKSHILTFH
252         274     . CEHEGCGKTFAMQSLDRHFNTH

=====
=====
#
# Sequence: TF3A_HUMAN      from: 1    to: 423
# HitCount: 9
```

```

#
# Pattern: C-X(2,4)-C-X(2,15)-H-X(2,4)-H
# Mismatch: 0
#
=====

Start      End Mismatch Sequence
100        122      . CSFPDCSANYSKAWKLD AHLCKH
130        148      . CDYEGCGKAFIRDYHLSRH
130        152      . CDYEGCGKAFIRDYHLSRHILTH
160        183      . CAANGCDQKFNTKSNLKKHFERKH
222        244      . CTQEGCGKHFASPSKLKRHAKAH
249        271      . CQKGCSFVAKTWELLKHVRETH
277        297      . CEVCRKTFKRKDYLKQHMKTH
306        329      . CPREGCGRTYTTVFNLQSHILSFH
337        359      . CEHAGCGKTFAMKQSLTRHAVVH
...以下省略

```

事實上連使用的資料庫及輸出檔的名稱都可在指令行指定，這留給大家做練習用。

練習 9-3 試在指令行下條件，以TFIIIA的鋅指模組樣式搜尋Swiss-Prot資料庫。

Answers:

```

%fuzzpro -sequence sw:l* -pattern 'C-X(2,4)-C-X(2,15)-H-X(2,4)-H' -outfile tfiiia.fuzzpro -auto &
[1] 22480
Protein pattern search
%
```

因為 TFIIIA 的模組樣式要求較嚴，所以練習 9-3 的輸出檔案大約只有 423Kbyte，即使如此，要用人工分析這數據還是很辛苦的。當然，大家可在命令列中指定參數「-rformat」選擇不同的報告輸出格式來幫助判讀。

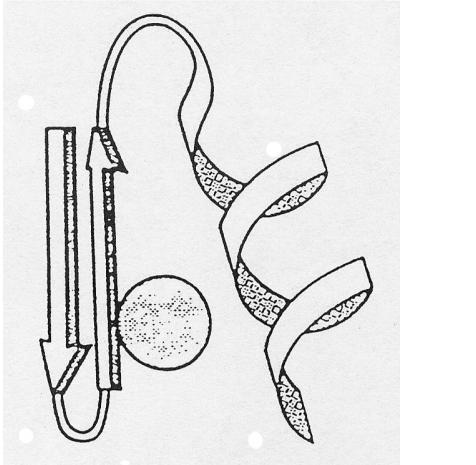
三、立體結構的預測

在 1986 年，Berg 根據他對蛋白質化學的瞭解，不單找出了 C_2H_2 形式的鋅指，還將鋅指分為五類：(1) Cys-X₂-Cys-X₄-His-X₄-Cys，例如 HIVgag protein；(2) Cys-X₂-Cys-X₁₃-Cys-X₂-Cys，例如 adenovirus E1A protein；(3) Cys-X₂-Cys-X₉-Cys-X₂-Cys，例如 tRNA synthetases；(4) Cys-X₂-Cys-X₁₁₋₁₃-His-X₂-His，例如 SV40 T antigen；(5) Cys-X₃-His-X₅-Cys-X₂-Cys，例如噬菌體 T4 的 helix-destabilizing protein。

他也注意到在 C_2H_2 形式的鋅指中，有一些序列也可在蛋白質結構資料庫(pdb, protein databnk)中找到。例如 rubredoxin (Fe) 與 aspartate transcarbamoylase (Zn) 等均有 Cys-X₂-Cys 這序列，這些序列的一個共同的結構特徵就是都出現在一對反平行的 β -鏈中，並且其上的疏水性胺基酸都在金屬離子的一側。而在 thermo-lysin (Zn)、hemerythrin (Fe) 與 hemocyanin (Cu) 上都可找到 His-X₃-His 這序列，它們全部出現在 α -螺旋中，並且以 His 的 α -N 和金屬離子作用。

在前一章預測 TFIIIA 的二級結構時亦發現在每個重覆中都有 α -螺旋(參閱第 8-5 頁，表 8-1)，而且許多 DNA 接合蛋白都是靠 α -螺旋嵌入 DNA 的主溝槽(major groove)中進行辨識，因此 Berg 大膽地提出鋅指的可能結構⁽³⁾ (參閱圖 9-10)。後來 Lee⁽⁴⁾等人在 1989 年以 NMR 研究在 Xfin 蛋白質中的一個鋅指的結構時，驗證了 Berg 的預測。這種預測方法假設序列相似的兩段 peptide 很可能具有相同的結構，事實上也是建立在這些片段能獨立折疊的概念上的(參閱表示 DNA 序列的方式對解讀 DNA 語言的影響)。這種利用序列相似的已知結構來推測新蛋白質結構的策略是目前預測蛋白質三級結構較準確的一個方法，也是未來的趨勢。

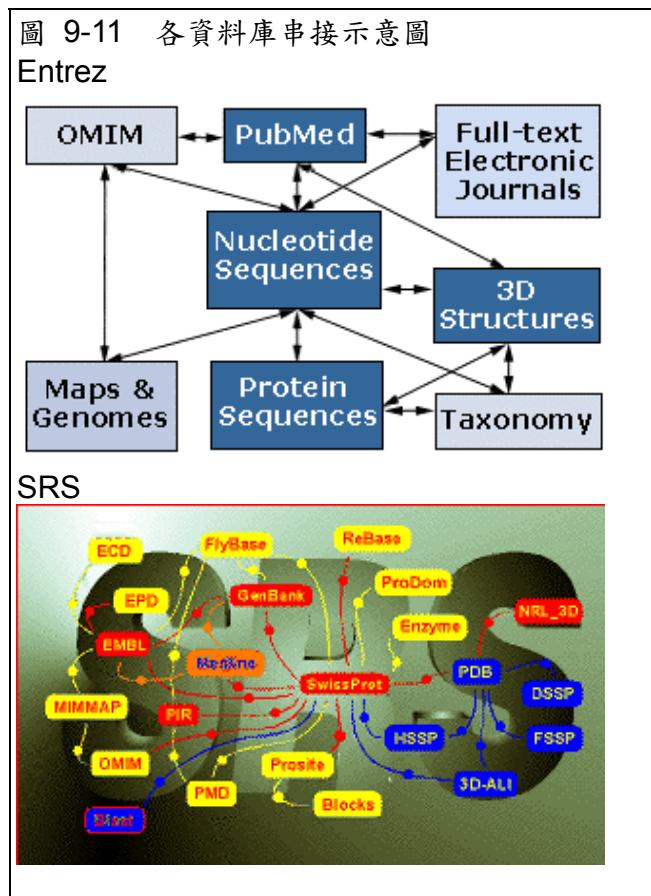
圖 9-10 Berg 所提出的鋅指模型



四、結語

在國家高速電算中心中有 Homology 這種軟體，可以協助分子生物學家利用有蛋白質的序列相似性來預測新蛋白質的結構。這將在進階的課程中才會討論。在搜尋蛋白質結構資料庫方面，現在 GenBank 資料庫中，不但紀錄序列與描述序列的資訊，也開始將序列與其他資料庫中的關係建起來。例如 Xfin 中 GenBank 序列檔，不但紀錄著蛋白質資料庫中的索引代碼(PIR : S00647)，若在 PDB 中有這蛋白質的結構，也會收錄其代碼(IZNF)。而且在紀錄中所引用的文獻(medline : 89378224)也可直接連到 Medline。美國國家醫學圖書館的生物技術資料中心(NCBI)發展出的 Entrez 系統或者歐洲 EBI 的 SRS 系統(第四章)就可利用這些代碼將各資料庫串接起來(參閱圖 9-11)。所以不論查任何一個資料庫都可看到其他相關資料庫中的資料。另外，日本的 GenomeNet 下的 dbGET 也提供類似的線上查詢系統。

Berg 所提出的預測，不但將原來所提出的鋅指模型由平面擴展到立體，其中一個最主要的差別在於 N 端到 C 端的方向在鋅指的起始與結束處是平行的，因此在不扭曲結構的前提下，最好是像一條繩索那樣，繞在 DNA 這雙螺旋結構的主溝槽中(圖 9-8)。由這裏我們可以看到，不同的模型直接影響到思考分子間交互作用的模式。此外，序列分析的工具，不單可處理一些分子生物學上的問題。只要瞭解蛋白質的特性，也能成功地預測分子的性質與結構。若自己的預測不成功，不能一味地怪工具(軟體)不好，也要問一問自己對蛋白質的性質瞭解的是否像 Berg 那樣深入。



參考文獻

1. Miller J., McLachlan A. D., Klug A. (1985). EMBO J. 4(6):1609-14.
2. Hanas J. S., Hazuda D. J., Bogenhagen D. F., Wu F. Y., Wu C. W. (1983) J Biol Chem. 258(23):14120-5.
3. Berg J. M. (1988). Proc Natl Acad Sci U S A. 85(1):99-102.
4. Lee M. S., Cavanagh J., Wright P. E. (1989). FEBS Lett. 254(1-2):159-64.

第十章 問題導向學習(IV)：TFIIIA 基因體序列的分析

陳淑美¹、楊永正²

¹ 國科會科教處、²陽明大學生物資訊研究所

在選殖出 TFIIIA 基因的 cDNA 後，希望也能選殖其基因體序列，尤其是可能調控其表現的基因上游序列(upstream sequence)。史丹福大學的 Korn 與其同僚在 *Xenopus* 的基因體基因庫(genomic library)中找到兩個長約 15kb 的 λ -噬菌體，均包含 TFIIIA cDNA 的 5' 端與 3' 端序列，因此它們針對其中多個片段做霰彈槍(shotgun)式的序列分析。其中有一個片段，包含 cDNA 的 5' 端與其上游的序列(Accession # X15785)，另一片段則有幾個 exon 與 intron(Accession #X03736)。

一、尋找 mRNA 的起點

Roeder 與其同僚在選殖出 TFIIIA cDNA (accession # K02938)時，曾利用引子延伸(Primer extension)的方式決定 mRNA 的啟始點位置，結果發現其 cDNA 少了大約十七個鹼基對。現在若想比較基因體序列與 cDNA 序列之相對位置，就必須使用雙序列排比。雙序列排比(sequence alignment)有兩種方式，一種是尋找整體性(global)的最佳排列，另一種是尋找區域性(local)的最佳排列。在 EMBOSS 下的 needle 與 water 分別是這兩種分析方式的代表。整體最佳排列分析的特色是插入空隙，使兩序列幾乎是頭與頭、尾與尾配對；而區域性最佳排列分析是取出這兩個序列最相似的一段作序列排比。這兩種分析方式所用的 algorithm 相似卻不相同，為使其運作較成功，使用者還可以透過參數的設定，使並列的結果產生一些變化。

範例 10-1 請利用已知的序列資訊找出TFIIIA mRNA的起始位置

1. 區域性序列排比 (結果參閱圖 10-1)

```
% water -sequencea C_WINDOWS_chen_xeltfiiia.fasta -sbegin1 1 -send1 1518  
-seqall C_WINDOWS_chen_xltf3a5.fasta -gapopen 100.0 -gapextend 0.5 -brief  
-aformat srspair -auto
```

2. 區域性的序列排比可以顯示 cDNA 的起點與基因體序列之間的關係(圖 10-1)，向上游方向算 17 個核苷酸即為 TFIIIA mRNA 的起點(如下圖中*部份所示)

*	GGAAGC	cDNA sequence
CAGTGGCTTC	TACAAGTTCA	GAGGAAGC
421	431	441

在此選擇 water 程式的原因是這兩個序列的長度不同，如果用 needle 程式做相同的分析，也同樣使用程式的預設值，則找不到任何密集的相似區域(圖 10-2)，所以也無法找到 mRNA 的起點。若修改 needle 程式的參數雖有可能改善序列排比的結果，可是仍然無法清楚地看出最像的區域的邊緣，在以後將會討論空隙罰分(gap penalty)對序列排比的影響。

圖 10-1 用 water 找尋 TFIIIA 的 cDNA 序列和基因體序列中共有之序列

```
#####
# Program: water
# Rundate: Fri Dec 06 22:36:23 2002
# Align_format: srspair
# Report_file: xltfiiia.water
#####
=====
#
# Aligned_sequences: 2
# 1: XLTFIIIA
# 2: XLTF3A5
# Matrix: EDNAFULL
# Gap_penalty: 100.0
# Extend_penalty: 0.5
#
# Length: 70
# Identity: 66/70 (94.3%)
# Similarity: 66/70 (94.3%)
# Gaps: 0/70 ( 0.0%)
# Score: 314.0
#
#
=====

XLTFIIIA      11 agccgagggtgttcagttcgtgaaggagatggggagaaggcgctgc      60
                |||||||..|.|||.|||||||||||||||||||||||||||||||
XLTF3A5       446 agccgagggtcagcttagttactgaaggagatggggagaaggcgctgc   495
                |||||||..|.|||.|||||||||||||||||||||||
XLTFIIIA      61 cggtgtgttataaagcggtac      80
                |||||||..|.|||.|||||||||||||||||||
XLTF3A5       496 cggtgtgttataaagcggtac      515
                |||||||..|.|||.|||||||||||||||||||


=====
#-----
```

圖 10-2 用 needle 找不到 TFIIIA 的 cDNA 序列和基因體序列中共有之序列

```
#####
# Program: needle
# Rundate: Fri Dec 06 22:46:05 2002
# Align_format: srspair
# Report_file: xltfiiia.needle
#####
=====
#
# Aligned_sequences: 2
# 1: XLTFIIIA
# 2: XLTF3A5
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1562
# Identity: 382/1562 (24.5%)
# Similarity: 382/1562 (24.5%)
# Gaps: 1091/1562 (69.8%)
# Score: 550.0
#
#
=====

.
.
.
.
.

XLTFIIIA      151 acacaggagagaaaccattccatgtaaaaaggatgtgagaaaggc      200
```

第十章 問題導向學習(IV)：TFIIB 基因體序列的分析

XLTF3A5	1	... agatctat-tgagaaaagg-	17
XLTFIIIA	201	tttacctcgcttcactcaacttaaccgcactcactcactcatactggcga .	250
XLTF3A5	18	-----gc-----cttactg---	26
XLTFIIIA	251	gaaaaacttcacatgtgactcggatggatgtgacttgagat---ttact . . .	296
XLTF3A5	27	-----tgt-----gctgtaa--ttagatgctgtta--	49
XLTFIIIA	297	acaaggcaaacatgaagaaggcactttaacagattccataacatcaagat . . .	346
XLTF3A5	50	-----gttatcgcactcc-----	62
XLTFIIIA	347	ctgcgtctatgtgtg--ccattttgagaactgtggcaaagcattca--ag .	392
XLTF3A5	63	-----tgtgtggaccatt-----gcatccatcac	86
XLTFIIIA	393	aaacacaatcaattaaagggttcatcagttcagtcacacacagcagctgcc 	442
XLTF3A5	87	attcacaa-cagttacagttctcca-----acac-cagcagctgc-	125
XLTFIIIA	443	atacgaatgtcctcatgaaggctgtg-acaagcggtttcttgccctcc 	491
XLTF3A5	126	-----tgcaca-----c	132
XLTFIIIA	492	cgtttaaaacgtcatgaaaaagtccatgcaggctatccctgcaaaaagga .	541
XLTF3A5	133	cgttt----cctc-----ggct-----	145
XLTFIIIA	542	tgattcttgcatttggaaagacttggacattatacttgaaacacg .	591
XLTF3A5	146	-----tcatgt-----attat-----cacg	160
XLTFIIIA	592	tggcagaatccatcaggacctagcagttatgtgatgtgtgaaatcgaaaa 	641
XLTF3A5	161	tg-----ctccactaggac-----	174
XLTFIIIA	642	ttcaggcacaaagattacttgagggg---atcatcagaaaactcagaaa 	687
XLTF3A5	175	-tcaaccactaaga--ac--gaggggagtgtc--cagaaa---cac---	210
XLTFIIIA	688	aagagcgaactgttatctctgcgcctcgagatggctgtgaccgctcstat . .	737
XLTF3A5	211	----ccaaact-----tgtga-----aat	224
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
XLTFIIIA	1307	ttgctattaaaagtgaggtgcagcagccactggctgtttatataata .	1356
XLTF3A5	516		515
XLTFIIIA	1357	cattcattttagtaagactctgtattcatttcaaaagaatactaaggga .	1406
XLTF3A5	516		515
XLTFIIIA	1407	atgtgcaaaattgttatcactctactgtaaacacaaatgtactgcttgca .	1456
XLTF3A5	516		515
XLTFIIIA	1457	ccctgttggggcttttggggagggtgactgaccctgttttttt .	1506
XLTF3A5	516		515

```
XLTFIIIA 1507 ttaacggaaattc 1518
XLTF3A5      516           515
#-----
#-----
```

做整體序列排比的程式在比較兩個序列長度不同的序列時，不可避免的要插入許多空隙以尋找基因的整體相似性。在區域性的比較中，則是先取最像的一段，再加入空隙使其更像，例如上述 cDNA 序列只與含上游序列的基因體序列有部分重疊，程式必須能找到其重疊區域，而不是在其中插入空隙，使這兩個序列並列在一起。

二、基因上游序列的分析

一旦決定了 mRNA 的起點，即可分析基因上游序列的性質，通常需要知道有哪些可能的轉錄因子接合位置。這類分析其實與尋找限制酵素切割位置很像，在 emboss 之下可以使用 Tfscan 程式，來找一個序列中轉錄因子接合位置。

範例 10-2 請估計 TFIIIA 基因上游序列中的 TATA 盒相對於 mRNA 起點的位置

%tfscan xltf3a5

在圖 10-3 中第 426 個核苷酸的位置是"1"，第 425 個核苷酸則為"-1"，所以 TATA 盒約出現在-30。

圖 10-3 以Tfscan進行TFIIIA的上游基因序列分析的結果

HS\$CDC25C_03	R04367	379	380	GG
HS\$CDC25C_03	R04367	323	324	GG
HS\$CDC25C_03	R04367	322	323	GG
HS\$CDC25C_03	R04367	317	318	GG
HS\$CDC25C_03	R04367	310	311	GG
HS\$CDC25C_03	R04367	309	310	GG
HS\$CDC25C_03	R04367	295	296	GG
HS\$CDC25C_03	R04367	249	250	GG
HS\$CDC25C_03	R04367	194	195	GG
HS\$CDC25C_03	R04367	193	194	GG
HS\$CDC25C_03	R04367	192	193	GG
HS\$CDC25C_03	R04367	171	172	GG
HS\$CDC25C_03	R04367	142	143	GG
HS\$CDC25C_03	R04367	68	69	GG
HS\$CDC25C_03	R04367	17	18	GG
HS\$CDC25C_03	R04367	16	17	GG
HS\$CDC25C_01	R04365	511	512	GG
HS\$CDC25C_01	R04365	500	501	GG
HS\$CDC25C_01	R04365	497	498	GG
HS\$CDC25C_01	R04365	488	489	GG
.
.
.
.

HS\$CDC2_04	R04341	224	227	TAAC
HS\$CDC2_03	R04340	255	258	TAAC
HS\$CDC2_03	R04340	224	227	TAAC
HS\$BG_22	R04295	497	501	GGTGG
HS\$BG_20	R04293	504	507	TATA
HS\$BG_20	R04293	396	399	TATA
HS\$BG_20	R04293	394	397	TATA
XENLA\$TFIIIA_03	R04006	174	206	CTCAACCACTAAGAACGAGGC
HS\$APOB_28	R03690	313	320	CACAGGAA
HS\$APOB_28	R03690	245	252	CACAGGAA
HS\$IL6_06	R03553	357	360	TTCC
HS\$IL6_06	R03553	136	139	TTCC
RAT\$A2UG_13	R03539	309	314	GGGACA
RAT\$A2UG_11	R03537	311	316	GACACA
MOUSE\$MT1_05	R03449	126	132	TGCACAC
MOUSE\$JUND_04	R03384	331	336	GCCAAT
MOUSE\$MT1_04	R03173	396	401	TATAAA
HS\$ASCC_04	R03054	504	508	TATAA
HS\$ASCC_04	R03054	396	400	TATAA
MOUSE\$THY1_07	R03047	487	490	AGGC
MOUSE\$THY1_07	R03047	403	406	AGGC
MOUSE\$THY1_07	R03047	294	297	AGGC
HS\$GHA_10	R02848	332	336	CCAAT
HS\$PL_08	R02790	362	368	GATGCAT
RAT\$NEU_01	R02465	332	336	CCAAT
MOUSE\$UPA_01	R02095	316	321	AGGAAA
HS\$GG_21	R02048	332	336	CCAAT
RAT\$TH2B_02	R01932	332	336	CCAAT
.
.
.

三、搜尋序列中字串的快速方法

尋找轉錄因子的結合位置，與區域性序列排比有些相似，可是有更快速的方法可以達到目的。要確保一定找得到存在的相似性，就必須使用有系統的方法，這樣若未找到相似的區域，才能很確定不是因為疏忽而未找到。這個尋找兩個序列間相似性的問題，其實和研究色層分析圖譜間的相似性，或是 RNA 結構等都是同類的問題。

從頭到尾掃描，來確定字串出現的位置固然一定可找目標，可是速度較慢，序列越長時就會花越多的時間。要找的字串數目越多，就需要越多的時間，在資料結構上雖會討論二元樹(binary search tree)的方法，它在序列分析上卻不如「混亂編碼(hashing)」法用的多。其實這個方法不但可用做字串搜尋，亦可用在轉譯遺傳密碼上，因為它的基本概念就是建立一對照表(look up table)。這有些像外文書在書末都有索引，因此使用者不必直接到書中瀏覽，而可經由查閱索引，直接跳到字串出現之處，書越厚，利用索引就比直接瀏覽越快找到主題。

圖 10-4 序列 A 的兩種表示方式，(A) 以鹼基形式表示；(B) 以雙鹼基 (dinucleotide) 形式表示

A.	5	10	15	20	
	TCGGA	TTCGT	ACGGT	ACGGA	TC
↓					
B.		5	10		15
	TC, CG, GG, GA,	AT,	TT, TC, CG,	GT, TA,	AC, CG, GG, GT,
		20			TA
	AC, CG, GG, GA,	AT,	TC		

以 DNA 序列為例，可選取 n 個字母為一個字，例如圖 10-4 中的序列，是以兩個鹼基為一個字，可將序列用字的形式寫出。因為 DNA 只有四種鹼基，所以兩個鹼基只能組合出 16 個字，列於表 10-1，而圖 10-4 在的位置亦可紀錄在表中。若想查詢 “CG” 出現的位置，只需比對這表上出現的字，即可知這個字出現在序列上 2、8、12、17 這四個位置。

當然使用者可以建立 n 個鹼基所形成的「字」之對表。例如在尋找限制酵素切割位置時，若每個酵素都要掃描一次序列會花許多時間。可是若先建立一個對照表，則多個酵素均可使用此對照表找到切割位置，這樣搜尋的速度就快許多了。

表 10-1 序列 A 的索引對照表

	Dinucleotide	Position		Dinucleotide	Position
1	GG	3,13,18	9	GA	4,19
2	TG	-	10	TA	10,15
3	AG	-	11	AA	-
4	CG	2,8,12,17	12	CA	-
5	GT	9,14	13	GA	-
6	TT	6	14	TC	1,7,21
7	AT	5,20	15	AC	11,16
8	CT	-	16	CC	-

另一個常用來比較序列相似性的方法是使用點矩陣，它的優點是計算速度快，可以配合利用視覺，在長序列中看到相似的區域。在 9-1 頁中曾說明此運作的原理，在此不再多言。只是要指出在垂直與水平方向的兩個序列若不相同時，只會看到偏離對角線的線段，而不會看到對角線。

四、預測 intron-exon 之交界點

直接由基因體序列預測出 intron-exon 的交界點是一門學問，現在雖然人類基因體計劃已完成，基因預測卻仍有不足之處，若能取得 cDNA 將對辨識基因有很大的幫助。此處由基因體序列與 cDNA 序列的比較來找 intron-exon 的交界點，此為辨識基因的重

要步驟。在上述找 mRNA 起點的例子中，兩序列只有一段相似的區域；如果一段序列上有多個 intron，則兩序列上會有多段相似的區域，這該怎麼處理呢？

練習 10-1 Korn的實驗室選殖到整個TFIIIA的基因體序列，可是只有決定了幾段序列，請利用其中的一段(accession # X03736)與cDNA 序列比對，找出intron-exon邊緣的位置。

Answer:

仍然使用water。如果你使用預設值(Gap opening penalty = 10; Gap extension penalty = 0.5)，你將得到圖10-5的結果，顯然無法找到表10-2中所列出的所有exon。

圖 10-5 利用water程式找尋exon-intron交界(使用預設值)。

```
#####
# Program: water
# Rundate: Sat Dec 07 00:03:01 2002
# Align_format: srspair
# Report_file: xltf3a36.water
#####
=====
#
# Aligned_sequences: 2
# 1: XLTF3A36
# 2: XLTFIIIA
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1707
# Identity: 753/1707 (44.1%)
# Similarity: 753/1707 (44.1%)
# Gaps: 810/1707 (47.5%)
# Score: 1672.5
#
#
=====

XLTF3A36      1 ct-----tggcaacatataatatattttttttcctctcagatgtgactc 45
                  |||  |||||.||.||.|||  |||  ||||.|||||||
XLTFIIIA     238 ctcatactggcgagaaaa----ct-tcacatgtgactc 271
XLTF3A36      46 ggatggatgtgacttgagatttactacaaggcaaacatgaagaagcact 95
                  |||||||||||||||||||||||||||||||||||||||
XLTFIIIA     272 ggatggatgtgacttgagatttactacaaggcaaacatgaagaagcact 321
XLTF3A36      96 ttaacagattccataacatcaagatctgcgtctatgtgtaatgttacac 145
                  |||||||||||||||||||||||||||||||
XLTFIIIA     322 ttaacagattccataacatcaagatctgcgtctatgtgtaatgttacac 358
.
.
.
XLTF3A36      296 tgataacttggcaatgtatcttggtttagtgccatggagaactgtgg 345
                  |||||||||||||||||||||||
XLTFIIIA     359 -----gtgccatggagaactgtgg 379
XLTF3A36      346 caaagcattcaagaaacacaatcaattaaagggttcatcagttcagtcaca 395
                  |||||||||||||||||||||||
XLTFIIIA     380 caaagcattcaagaaacacaatcaattaaagggttcatcagttcagtcaca 429
XLTF3A36      396 cacagcagctgccatacga-----gtaagaa----- 421
                  |||||||||||||||  |||.|||
XLTFIIIA     430 cacagcagctgccatacgaatgtcctcatgaaggctgtgacaagcggtt 479
XLTF3A36      422 -----ccttctactgttacctggtaatgtc---aaaaa----- 453
                  |||||  ||.|||  |||.|||  |||||
XLTFIIIA     480 tctttgccttc--ccgtt-----taaaacgtcatgaaaaagtccatgca 521
```

XLTF3A36	454	------caa----ggtaattcttg-gaatttctgctactgga--	485
		
XLTFIIIA	522	ggctatccctgcaaaaaggatg-attctgctcattgtg----ggaaa	565
		.	.
		.	.
		.	.
XLTF3A36	1117	ggcagtgttgtcttt-----ggttaagcaggctatgg---gttta	1154
		
XLTFIIIA	1300	ccca-tgtttg-ctattaaaagtgaggtgcagcagccactggctgttta	1347
XLTF3A36	1155	--tacaa-acaaacttattatactgatcaaggagtact-tgtttgcattg	1200
		
XLTFIIIA	1348	tttacaatacaca--ttcatt-tagt----aag---actctgtattcattt	1386
XLTF3A36	1201	acaaaatgtaac----atggattgtg---atttgaacctttatgtcttag	1243
		
XLTFIIIA	1387	tcaaaagaatcactaaggaaatgtcaaaattg----ttatcact---	1427
XLTF3A36	1244	gctatccctgcaaa---aaggatg--attcttgctc-----	1274
		. . .	
XLTFIIIA	1428	-cta---ctgtaaaacacaa--atgtactgcttgcaccctgttggggc	1471
XLTF3A36	1275	-atttgtggaaagacttg---gac---attatacttgaacacgtggc	1316
		
XLTFIIIA	1472	tttttttggggagg--ttgactgaccctgttttttttaa-----c	1511
XLTF3A36	1317	agaatgc 1323	
		.	
XLTFIIIA	1512	ggaattc 1518	
<hr/>			
#-----			
#-----			

在此使用區域性的序列排比是一個正確的選擇，因為不知道基因體序列的頭尾是否與 cDNA 序列相似。使用預設值雖然能在 cDNA 序列中央插入連續的空隙，來跳過 *intron* 的區域，為何這樣的條件只能找到兩段相似的區域呢？你可能已注意到此處的空隙很小，而有些 *intron* 的長度很長(參閱表 10-2)。程式的預設值是每產生一個空隙扣 10 分，延伸一個空隙再扣 0.5 分，所以空隙越大扣分越多，大的空隙不易產生。依此推論，只要延伸空隙不扣分或扣很少的分數，即有機會找到所有的 *intron*。

表 10-2 TFIIIA 基因體序列中 exon、intron 之範圍。

features	listed ranges	water ranges	Comments
intron	1-36	1-50	
exon 3	37-133	51-132	第 3 號鋅指
intron	134-325	133-324	
exon 4	326-414	325-414	第 4 號鋅指
intron	415-980	415-980	
exon 5	981-1054	981-1052	第 5 號鋅指
intron	1055-1243	1053-1241	
exon 6	1244-1330	1242-1330	第 6 號鋅指
intron	1331-1337	none	

據此原則再重做一次 water，插入空隙的罰分改成 100，避免空隙加的太零散，但將延伸空隙的罰分改變為 0.1，果然得到不錯的結果(參閱圖 10-6)。預測的 Exon 範圍列在表 10-2 中，與真的交界點很接近。

圖 10-6 利用water以找尋exon-intron交界(將Gap opening penalty 改為100；Gap extension penalty 變為0.1)。

XLTFIIIA	521	-		aggctatcc	529
XLTF3A36	1251	ctgcaaaaaggatgattttgcattgtggaaagacttggacattat		1300	
XLTFIIIA	530	ctgcaaaaaggatgattttgcattgtggaaagacttggacattat		579	
XLTF3A36	1301	acttgaaacacgtggcagaatgccatcagg		1330	
XLTFIIIA	580	acttgaaacacgtggcagaatgccatcagg		609	
#-----					
#-----					

有趣的是上述每個 **exon** 轉譯出的蛋白質片段都含有一個鋅指模組，事實上除第八個 **exon** 有兩個鋅指，第九個 **exon** 沒有鋅指外，其它的 **exon** 都與鋅指的出現有很好的關係。這個觀察支持獨立摺疊模組(IFM)在基因重組的過程中不被破壞的論證。

如果用 **needle** 程式做相同的預測時，改變參數會有怎樣的結果呢？在此不多敘述，只將其結果列在表 10-3 中提供使用者參考。由數值上可見相似性會隨罰分的減少而增加，可是其效果仍然不如使用 **water** 程式。

表 10-3 改變參數對用 **needle** 預測 Exon-Intron 連接處的影響。

	G=10 ; L=0.1	G=10 ; L=0.3	G=10 ; L=0.5	G=10 ; L=10
Length	2006	1986	1964	1636
Gaps	57.7%	56.2%	54.6%	25.5%
% Identity	36.2	37.4	38.2	43.8

* G = gap opening penalty; L = gap extension penalty。

五、序列排比的原理

凡是自己試著用人工的方法做小段序列排比的人，可能都知道這不是件容易的事。因為不同的排列方式的得分可能很相近，用人工不易確定是否找到所有可能的並列方式。在另一方面，若需考慮不同胺基酸間的相似性來給分，則計分亦變成一繁複的工作。

既然使用者要找的是最佳的排列方式，最安全的方法就是列出所有可能的並列方法，一一計分，再找出得分最高的排列方式。這個過程，有些像在預測 RNA 二級結構時，畫出所有的結構，再計算出所有結構的自由能，以便尋找能量最低的結構。這兩件看似全然不同的工作的確是相通的，可以用類似的策略來解決這兩個問題。

在真正運算時當然不希望一一列出所有的可能性，然後才算出最佳並列方式。一般常用的是一種叫「dynamic programming」的運算法，這種方法在幾個不同的領域各自獨立地被發展出來，解決該領域的問題，到後來才知道他們用的其實是同一個方法。在生物學上是由 Needleman 與 Wunsch 所提出來的，目的是在做廣域性的序列排比，因此許多人以「Needleman and Wunsch algorithm」來表示廣域性的序列排比。這種方法的基本原理是用「分而治之(divide and conquer)」的策略，將大的問題化為小的而且相似的問題，這樣即可用遞迴(recursive)法，逐一解決小的問題。在圖 10-4 中的序列中

的前十二個鹼基含兩個不完美的重覆，這兩段序列有九種可能的並列方式(圖 10-7)。

圖 10-7 序列 A 中不完美重覆序列之序列排比

A.	D.	G.
TCGGAT TCGTAC	TCG-GAT TCGT-AC	TCG-GAT- TCGT-A-C
B.	E.	H.
TCGGAT- TCGTA-C	TCGG-AT TCG-TAC	TCGG-AT- TCG-TAC-
C.	F.	I.
TCGGA-T TCGTAC-	TCG-GA-T TCGT-AC-	TCGG-A-T TCG-TAC-

在這九種並列方式中，可以分為允許與不允許誤配兩種情形。在不允許一個鹼基對誤配的情形，又有兩種可能的排列方式比較圖 10-7，B 與 C)。因為有兩個可能產生誤配的位置，所以又可組合出許多並列的方式。由這個例子可看出不同的計分方式可能會得到不一樣的結果，例如使用者若覺得在演化的過程中比較容易產生點突變而不易產生插入式突變，可在並列時每加一個空隙，就扣若干分。可是使用者若認為某一種類型的突變不易發生，則可在產生誤配時扣較多的分數。由此可看出序列排比的程式就有如邏輯論證，而使用者的模型則是假設(*assumption*)，如果假設錯誤，即使邏輯正確，也會得到錯誤的答案。

在此先假設你有一個正確的模型，我們要問的是程式要怎樣替你做最佳化的。這個過程有點像使用算盤，可以分為兩個階段，第一個階段是做機械的運算，使用者根據題目與口訣撥動珠子計分；第二個階段則讀取結果。整個過程不需思考，只是反射性地做機械動作，所以計算速度有可能比使用計算機還快。在「dynamic programming」中，第一個階段計分，第二個階段則反向尋找(back trace)最佳路徑，而這路徑可代表一個序列排比，或一個 RNA 結構，甚至兩個色層分析圖譜的相似性。

為什麼一定要做反向尋找才能找到結果呢？這是因為在計分的過程中有太多可能的選擇，這有點像爬山時遇到三叉路，如果走錯了要再回頭走另一條路，雖可通往想去的地方，卻浪費了許多時間。可是若有地圖，就很容易規劃出最佳的路徑。地圖所代表的是地景的全貌，而先計算出所有較佳配對方式之得分，就能看到序列排比的全貌，也能找到最佳的序列排比方式。

假設兩序列上指定的位置有相同的鹼基可得一分，其它的狀況，包括誤配或插入空隙都給負一分，則序列比對的計分結果可表示成一個矩陣的形式(圖 10-8)。矩陣中每一格的得分是由比較相鄰的三格而得到的，以序列的第一、二個位置為例(圖 10-9 上方)，

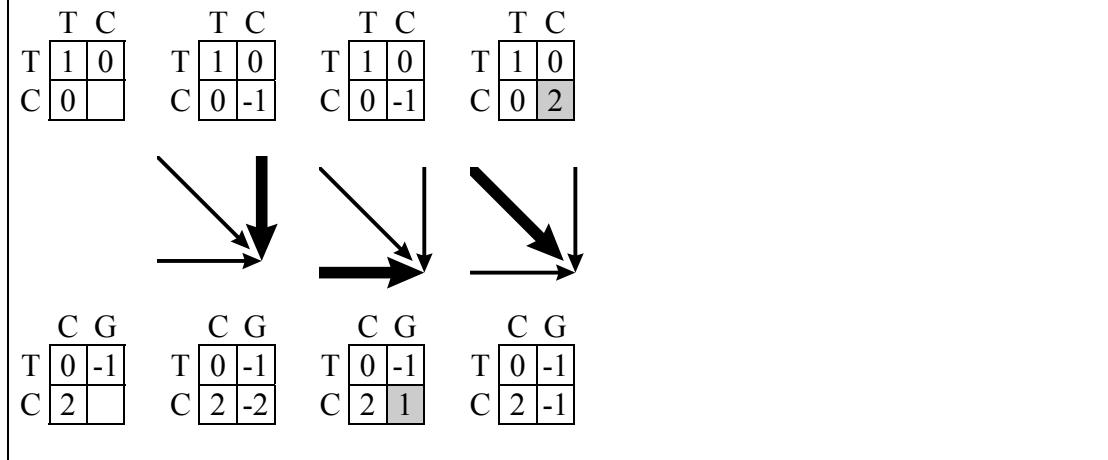
有三種不同的方式可以做並列。因為 C 與 C 配對可以得一分，不配對則被扣一分，所以沿對角線方向移動的這種並列方式分最高。同樣的道理，若考慮第一個序列的第 2,3 個鹼基與第二個序列的第 1,2 個基鹼基的比較，三種方式均扣一分，因此可直接取週圍三格中最高的得分，減一分填入此空格(圖 10-9 下方)。若每次都在空格中填入三種移到方式中得分最高的得分，在未來做反向尋找路徑時自然會找到最佳的路徑。

圖 10-8 動態程式設計的計分方式

	T	C	G	G	A	T
T	1	0	-1	-2	-3	-4
C	0	2	1	0	-1	-2
G	-1	1	3	2	1	0
T	-2	0	2	2	1	2
A	-3	-1	1	1	3	2
C	-4	-2	0	0	2	2

根據上述的計分方式，可以有系統地由矩陣的第一列由左計算至右，跳到第二列由左至右，……依此類推，直到計算出所有的分數(圖 10-8)。此時可由起點，挑得分高的路徑，有時會遇到多個路徑都有同樣的得分，這代表有幾種同分的並列方式。以上述的計分方式，就可能產生圖 10-7 所示的九種可能的並列方式。如果調整計分的方式，可能會有利於某幾種，甚至一種並列的方法。換言之，是使用者的模型，決定了最後的結果，使用者必須要測試假說是否成立，而不是怪程式程式無法得到自己想要的序列並列結果。

圖 10-9 每個格子的得分都是相鄰的三個格子的最高得分與該格子的分數之和



以上所描述的廣域性的序列排比，兩個序列的頭尾都是對齊的。區域性序列分析也運用同樣的策略，只是不要求序列末端對齊，而在矩陣中尋找最大的並列區域。這種方法必須要配合計分系統才能成功，因為扣分的值若與加分的值相當，則越長的序列得分越高，換言之全序列的分數會最高，這就變成了廣域性的序列排比。因此在做區域性序列排比時計分系統要偏向負值，這樣只有好的並列區域才會變正，而抵銷掉長度所造成的偏差。這種運算法最先由 Smith 與 Waterman 提出，因此當文章中提到「Smith-Waterman algorithm」時，所代表的是區域性序列排比。

六、 結語

本章所討論的字串搜尋，與序列排比都在為討論更複雜的資料庫搜尋與多序列排比做準備工作。由本章所討論的原理與所舉的例子中，讀者應能瞭解該怎樣思考調整插入空隙罰分與延伸空隙罰分來達到自己的需求。

第十一章 問題導向學習(V)：5S rRNA 的二級結構分析

楊永正

陽明大學生物資訊研究所

5S rRNA 是核糖體中的一種結構性 RNA，它和其它的 rRNA 會與核糖體蛋白質作用，而組合成轉譯作用所必須的核糖體。這個基因的表現受到 TFIIIA 的調控，雖然 TFIIIA 是一個會與 DNA 作用的轉錄因子，在卵母細胞中，它也能與 5S rRNA 接合，形成 7S 粒子以貯存大量的 5S rRNA，供胚胎早期發育之用。直到受精卵形成後，這些 5S rRNA 才會與 TFIIIA 分離，而進入核糖體參與轉譯作用。在此我們要以 5S rRNA 為例，討論預測 RNA 二級結構的方法。目前預測 RNA 二級結構的方法可分為兩類：第一種方法，假設自然界存在的 RNA 之二級結構是自由能較低的結構，而由序列預測能量最低的二級結構(參閱方盒 11-1)；另一種方法則是假設在演化的過程中，為了維持結構，序列雖可能產生變化，配對的關係卻能保存下來，因此利用尋找多個序列的親緣關係以預測二級結構(參閱第 11-16 頁，方盒 11-3)。不論是哪一類的研究方法都需要電腦程式協助分析。因為 EMBOSS 中尚無此類程式，所以本章是以 GCG 中的程式為例做說明。

一、由 RNA 序列預測 RNA 的二級結構

在 GCG 中有兩個由單一序列預測 RNA 結構的程式，分別是 FoldRNA 與 MFold，這兩個程式都有能力預測能量較低的結構。前者所輸出的結果較明確，但產生的結果只能當作參考，還需用實驗數據修正預測的結果，才能找到比較可信得結構。後者一次可列出多個結構，瞭解預測原理的使用者，可由這一群預測結構中，拼湊出一個比較可信的結構，再用實驗數據修正或佐證結果，會比第一種方法容易得到正確的結果。

方盒 11-1 由序列預測能量最低二級結構之原理

決定 RNA 結構穩定性的主要因子，是鹼基對(base pair)間的堆疊(stacking)作用。不同的鹼基對，在堆疊時釋出的自由能具有加成性，因此可計算不同二級結構的 RNA 之自由能⁽¹⁾。以圖 11-1 中的結構為例，不同的鹼基堆疊形式之間的能量變化可自表 11-1 查到。例如圖 11-1 中，方框中的 (U, A) 與 (U, A) 堆疊能 (stacking energy) 為 -1.2 Kcal/mole，可是因為有一個 bulge，因此要在加上 3 Kcal/mole。表 11-1 中列出所有可能的堆疊能，各位可試著練習圖 11-1 中 RNA 能量之計算。

既然只要畫出結構就可計算自由能，應可比較不同結構之自由能大小。因此電腦程式可以直接從單一的序列，預測能量最低的 RNA 二級結構。因為預測的經驗有助於修正預測理論，這比蛋白質二級結構的預測全靠統計的方法要邏輯一些。從 1960 年到現在，熱力學上測得的實驗值已經數次修正，使其更精確。可是在 RNA 分子越大時，累積的誤差越大。目前常用的係數，包括室溫的 Salser-Cech 係數⁽²⁾與 37°C 用的 Turner 係數⁽³⁾。這些參數還沒有準確到可以判斷一個長數百個核苷酸的 RNA 片段的最穩定結構，因為不同結構之間的能量差異，可能小於係數累加所產生的誤差。

當 RNA 的長度變長時，可能組合出的二級結構就變多，預測 RNA 二級結構就愈複雜，複雜到必須用電腦程式來比較那一個結構能量最低。這類程式也採用 dynamic programming 的方法有系統地尋找能量最低的結構，所以不允許結(knot)或假結(pseudo-knot)的拓樸結構形成。

圖 11-1 不同結構對鹼基堆疊自由能的影響

Improved Estimation of Secondary
Structure in Ribonucleic Acids

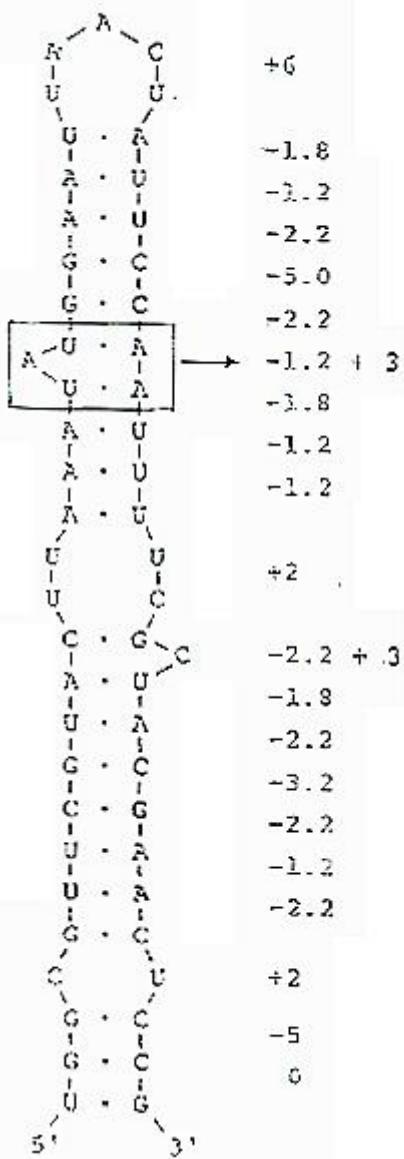


表 11-1 25 °C 下，RNA 的鹼基堆疊自由能⁽⁴⁾

Free energies at 25 °C for RNA structures	
Base paired regions	$\Delta G(\text{kcaloric}) \pm 10\%$
-A-A-	
-U-U-	-1.2
-A-U- • -U-A-	-1.8
-U-A- • -A-U-	
-A-C- -C-A- • -A-G- -G-A-	-2.2
-U-G- • -G-U- • -U-C- • -C-U-	
-C-G- -G-C-	-3.2
-G-C- • -G-G- -C-G- • -C-C-	-5.0
-G-U- -U-G-	-0.3
-G-X- • -X-G- -U-Y- • -U-Y-	0
Unbonded regions	
$\Delta G(\text{kcaloric}) \pm 1 \text{ kcaloric}$	
Number of bases unbonded	
2-6	+2
7-20	+3
$m(>20)$	$+1+2 \log m$
Interior loops	
1	+3
2-3	+4
4-7	+5
8-20	+6
$m(>20)$	$+4+2 \log m$
Bulge loops	
3	+3
4-5	+4
6-7	+5
8-9	+6
10-30	+7
$m (>30)$	$+4+2 \log m$
Hairpin loops	
Closed by G • C Closed by A • U	
3	+8
4-5	+5
6-7	+4
8-9	+5
10-30	+6
$m (>30)$	$3.5+2 \log m$ $5.5+2 \log m$

1. FoldRNA 與繪圖程式的使用

對使用者而言，預測 RNA 二級結構時希望能以圖形表示輸出結果，這樣一眼即可看出哪一個區域有 stem-loop 的結構，以判斷結構與功能是否相關。在執行 FoldRNA 這程式時，會產生一個副檔名為「fld」的文字檔，這是一個可用任何印表機列印的表示方式。為避免印出序列時文字重疊，有時不易看出序列連接的方式。因此「fld」檔除列出序列的好處外，不但不適合做發表的用途，連自己分析時都困難重重。一般預測結構的程式並不直接繪圖，因為計算二級結構需較長的時間，可是一但計算完成，即可用其他程式很快地繪出不同的圖。FoldRNA 會計算鹼基間連接的方式，並將結果紀錄在一個副檔名叫「connect」的文字檔中。

範例 11-1 試以Foldrna 程式預測Xenopus 5S rRNA的結構

```
%foldrna
FoldRNA predicts a single optimal secondary structure for an RNA molecule
by the older method of Zuker.

FOLDRNA on what sequence ?  xen5s.dna
What is the structure output file (* xen5s.fld *) ?
What is the base-by-base output file (* xen5s.connect *) ?

      Begin (* 1 *) ?
      End (* 120 *) ?
%
```

在第二階段中，使用者根據「connect」檔中紀錄的結果與自己的目的，繪成特定的圖（參閱圖 11-3）做進一步的分析。在分析結構與功能間的關係時，最好使用折曲圖(squiggles)，因為它可直接顯示 stem-loop 的結構。

範例 11-2 請繪出上述範例中所預測的結構之折曲圖(結果參見11-2A)

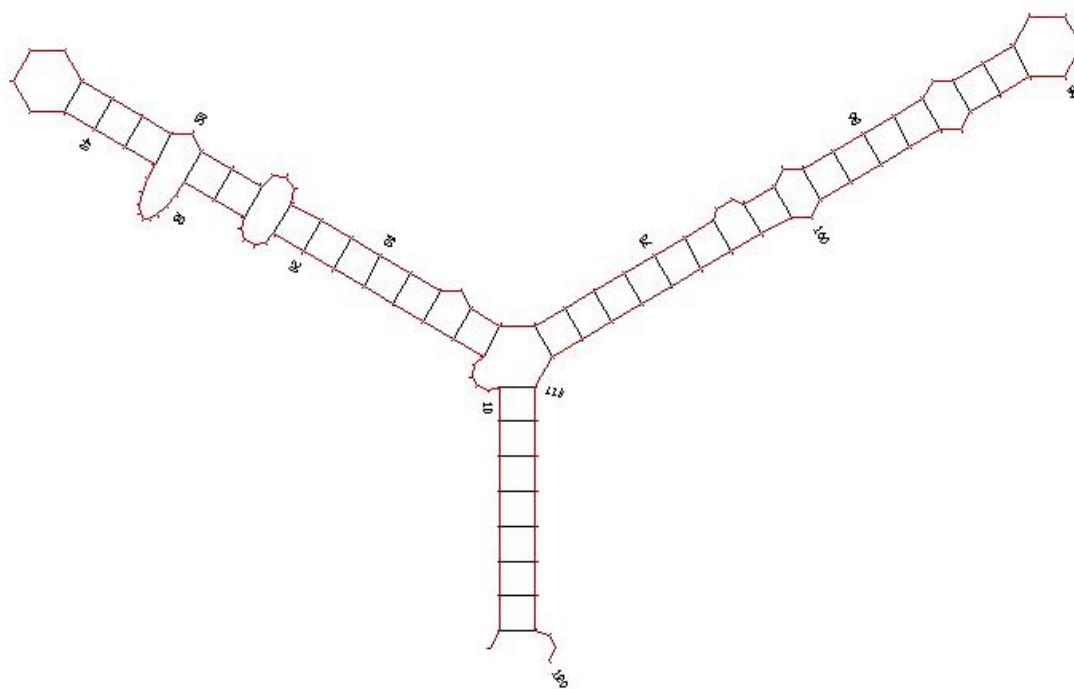
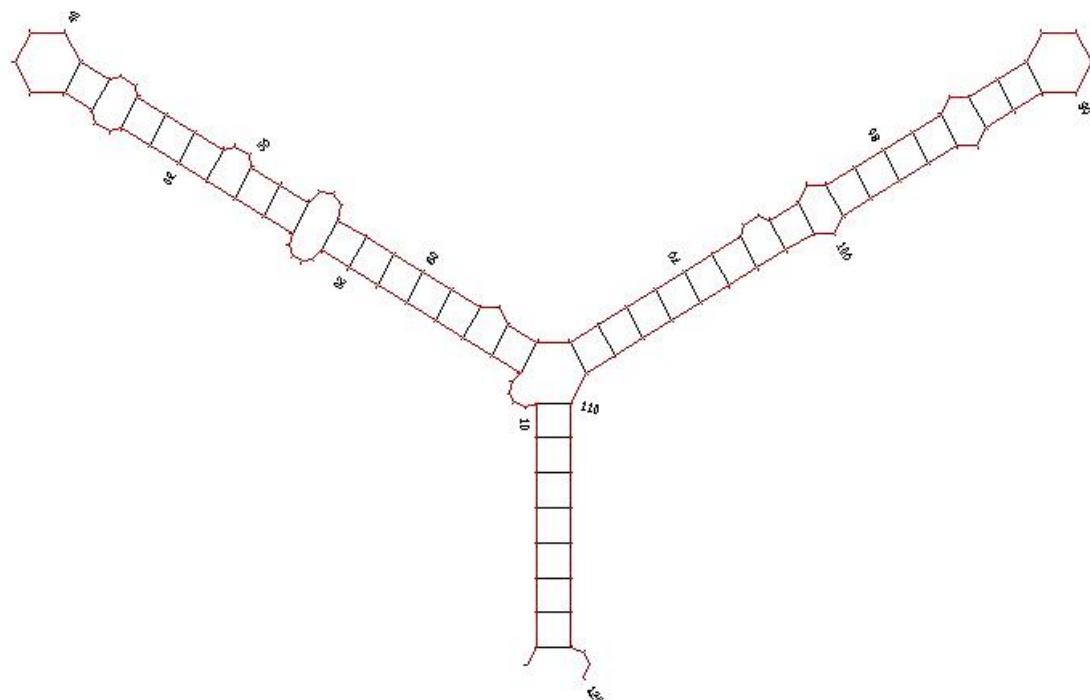
```
% hpgl laser graphics.dat
Plotting Configuration set to:
  Language: hpd
  Device: LASERJETIII
  Port or Queue: graphics.dat

% squiggles
Squiggles uses an output file from FoldRNA to make a plot of an
RNA secondary structure.

Process set to plot with LASERJETIII attached to graphics.dat
using the hpd graphic interface.

SQUIGGLES of what FOLDRNA output file ?  xen5s.connect
HPGL instructions for a LASERJETIII are now being sent to graphics.dat.
```

圖 11-2 在不同溫度預測 *Xenopus* 5S rRNA 二級結構之結果。A. 37°C；B. 25°C。

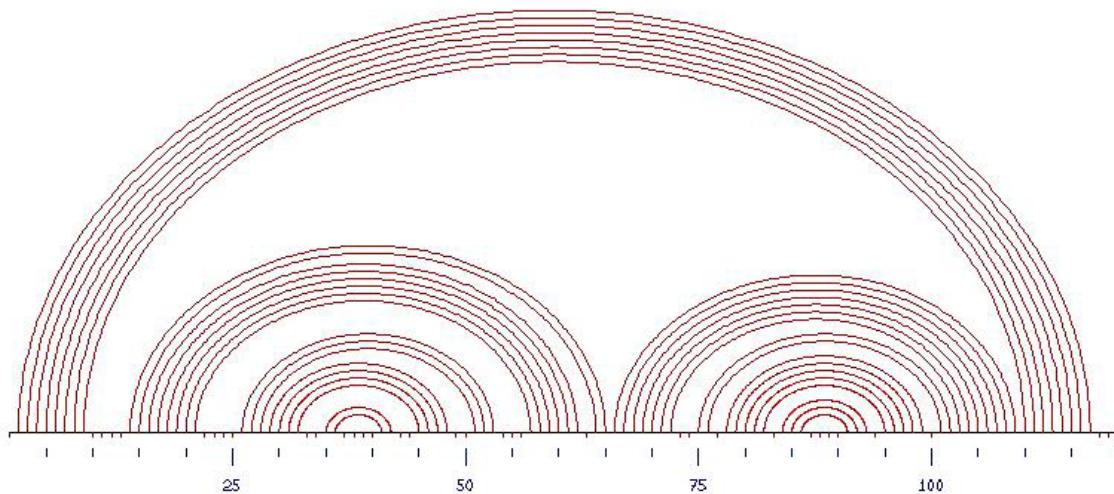


可是若想比較多個結構的相似性時，不同結構的折曲圖，可能因為 stem-loop 繪製的方向不同而不易找到相似性，此時可用 Domes(圖 11-3 A 與 B)、Circles(圖 11-3C)、或 Mountains(圖 11-3D)等方式來繪製圖。以 5S rRNA 為例，在比較圖 11-3 A 與 B 時，比較容易看出預測的與被接受的結構間之相似性。

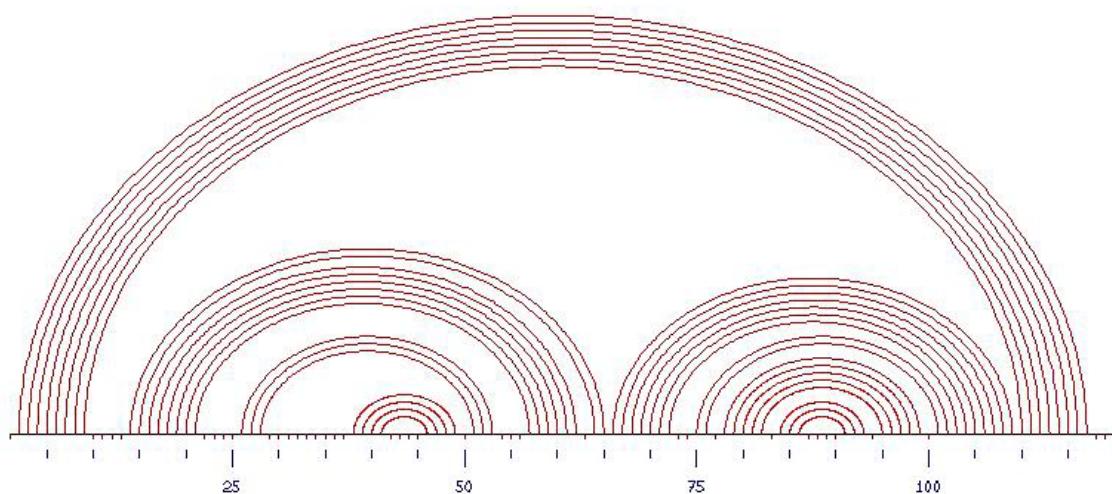
圖 11-3 以 5S rRNA 的結構為例，說明 RNA 結構可用不同的方法表現。

A. Domes 的輸出；B. Domes 的輸出；C. Circles 的輸出；D. Mountains 的輸出（除 B 為 25°C 外，其它均為 37°C 的結構）

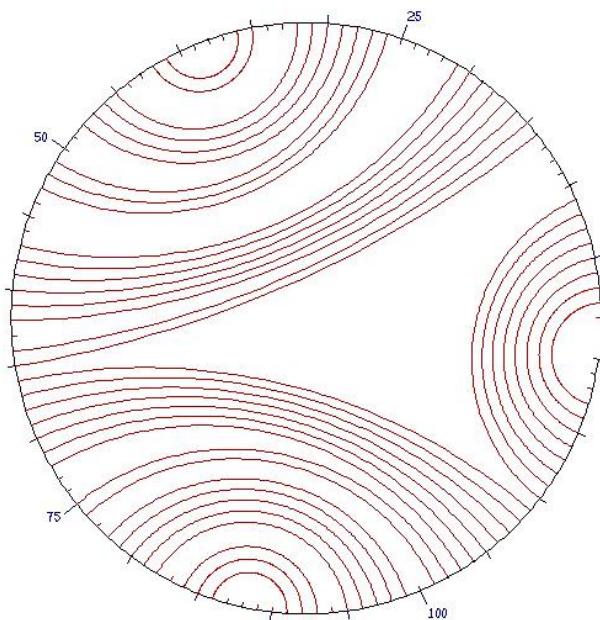
A.



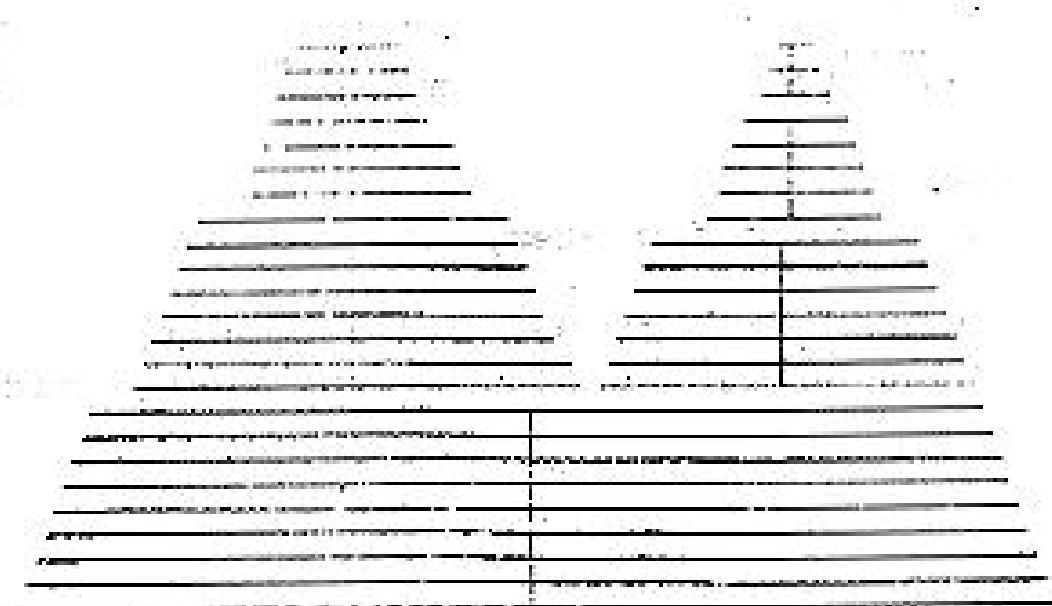
B.



C.



D.



在使用 FoldRNA 時要知道程式是預設 37°C 時的熱力學參數，來計算不同結構的能量。而 *Xenopus* 是體溫隨環境溫度而變的生物，因此應以 25°C 的熱力學參數預測結構較合理。

練習 11-1 請找出一個方法，預測 *Xenopus* 5S rRNA 在室溫的結構，並繪出其折曲圖

Answer:

1. 取回適當的參數檔，並將其命名為「25.en」。
% **fetch salser_cech.energy 25.en**
2. 以「25.en」預測 *Xenopus* 的 5S rRNA 結構
% **foldrna -dat=25.en**
3. 確定圖形環境已設置完成，再利用 squiggles 程式繪圖(結果參見圖 11-2B)
% **squiggles**

在圖 11-2A 中顯示的是 37°C 時所預測的 5S rRNA 的結構，而圖 11-2B 中則繪出 25°C 下預測的結果。由比較此圖之 A 與 B 可知，溫度似乎對預測的結構有些影響，其中 37°C 的結構與已被接受的結構幾乎相同，而 25°C 的結構則與已被接受的結構有一段差距，在後面將討論兩者結構為何有很大的差異。若仔細分析細部的結構，則發現圖 11-2B 中斜線部份的結構與被接受的結構相同，這部份的比例約佔全長的 59% (=71/120)。

2. MFold 與 PlotFold 的使用

溶液中存在的結構不見得就是最穩定的結構，可是組成整個結構的多數 stem-loop 則可能是最穩定的，因此就產生了預測 sub-optimal 結構的需要。在 GCG 下的 MFold⁽⁵⁾可以預測出比最低能量稍不穩定的所有結構，以協助 RNA 二級結構的預測。溶液中的結構經常是由幾個較穩定的 stem-loop 組成的，在求最低能量的過程中，可能因為所取的區域不完整，或是未考慮在 RNA 摺疊過程中可能的動力學障礙，甚至蛋白質的作用，所以使得某些相當穩的區域結構無法形成。MFold 這個電腦程式列出多種在給定能量範圍內的結構，使用者可組合這些數據，以期得到最可能的二級結構。

練習 11-2 請利用 MFold 程式預測 *Xenopus* 5S rRNA 在室溫中之二級結構

Answer:

```
% mfold xen5s.dna -temp=25
```

```
MFold predicts optimal and suboptimal secondary structures for an RNA
molecule using the most recent energy minimization method of Zuker.
```

```
Begin (* 1 *) ?
End (* 120 *) ?
```

```
What should I call the energy matrix output file (* xen5s.mfold *) ?
```

```
Folding .....
```

```
CPU time: 04.67
```

```
Output file: xen5s.mfold
```

因為 Mfold 可印出在能量最低點附近的多個 RNA 結構，所以必須用新的方式呈現結構的相似性，以免自己要比較多個結構相似之處，此時就需要用 energy dotplot 及 P-Num plot。為分析 Mfold 的輸出結果，GCG 提供 PlotFold 程式做所有的繪圖工作。

範例 11-3 請利用 PlotFold 繪出「xen5s.mfold」的energy dotplot

```
% plotfold
PlotFold displays the optimal and suboptimal secondary structures
for an RNA molecule predicted by MFold.

Process set to plot with LASERJETIII attached to graphics.dat
using the hpd graphic interface.

PLOTFOLD with what saved energy matrix file ? xen5s.mfold

Maximum size of interior loop = 30
Maximum lopsidedness of an interior loop = 30

Do you want to display:

SURVEY OF OPTIMAL AND SUBOPTIMAL FOLDINGS

A) energy dotplot
B) p-num plot

SAMPLING OF OPTIMAL AND SUBOPTIMAL FOLDINGS

C) circles
D) domes
E) mountains
F) squiggles
G) text output
H) connect file output

Please choose one (* A *):

Energy of optimal structure = -53.6
Plot base pairs at what energy increment (* 2.7 *) ?

How many color levels in the energy plot (* 1 *) ?

The minimum density for a one-page plot is
138.6 bases/100 platen units on each axis.

What point density would you like (* 138.64 *) ?

PLOTFOLD will take 1 pages. Would you like to:

P)lot the points
D)ifferent density

Q)uit
Please select one (* P *):
```

繪製 energy dotplot 及 P-Num plot 時，程式會問「能量增加值（energy increment）是多少？」，它代表比最低能量的結構能量高出的數值。程式會自動計算出預設值，以範例 11-3 中的例子而言，能量的增加值（2.7 kcal/mole）除以最低能量結構之能量絕對值（-53.6 kcal/mole）約為 0.05，即 5%。程式會將在能量在-50.9 kcal/mole 至 -53.6 kcal/mole 間的所有結構用來作 energy dotplot。對其他形式的圖而言，使用者可指定要畫多少個（例如 n 個）結構，程式就會由最低能量起，依次繪出能量較高的結構，直至繪出你所指定的 n 個結構為止。其他的問題多在控制繪圖的方

式，選用程式的預設值一般均可繪出令人滿意的結果。

3. MFold 的結果分析

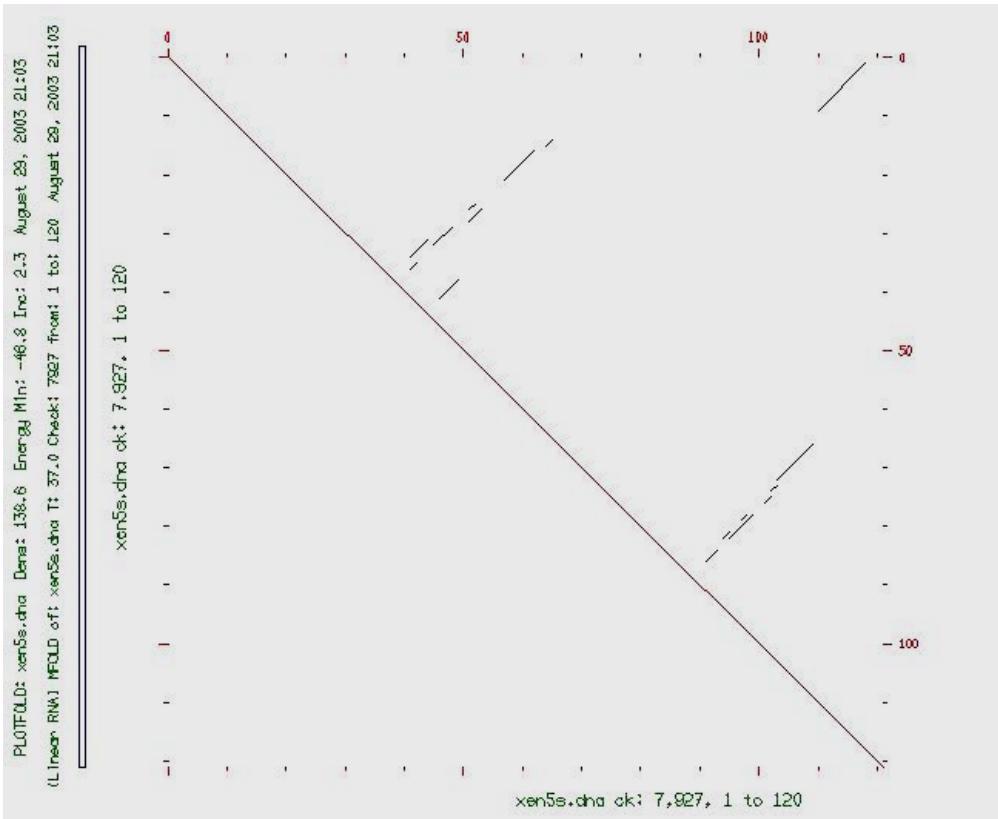
Foldrna 程式只產生單一的結構，因此只需將預測的結果與實驗結果相比較即可；可是 MFold 會輸出多個結構，必須瞭解怎樣閱讀『energy dotplot』與『p-num plot』之意義。在此有一個很重要的觀念，就是有一些比較穩定的小結構（sub-structure）在 suboptimal 的結構中也會出現，這些小結構出現的次數越多，這片段可能越穩定，因此真正的結構很可能就是這些小片段組合出來的。如果要一一繪出給定之 energy increment 中所有的結構，再用眼睛比較就太辛苦了，所以在 PlotFold 中提供上述兩種圖形，協助使用者很快就能找到這些較穩定的小片段之位置。

在 energy dotplot 中(圖 11-4A)，橫座標與縱座標都代表測試序列的位置，假設在兩個位置配對處打一個點，就會隨著配對方式的不同，而在圖上出現許多和對角線垂直的線，因為在對角線兩側之圖形是對稱的，所以只繪出一半。與對角線垂直的每一條線都代表一個小結構，若此線有不連續、甚至偏離的現象產生，所代表的是出現在 stem 中的 interior loop 或 bulge loop，此線與對角線之間的空白則代表出現在 stem 末端的 terminal loop。在給定之 energy increment 中所出現的各結構都會出現在這張圖上，因此能量增加值越大，圖上的線就會越多，在圖 11-4A 中，線的數目並未佈滿整張圖，這暗示在不同的結構中，某些小片段的結構是重覆出現的，所以圖上只看到有限的結構，每個小結構的位置及配對方式可直接由圖上的縱、橫座標讀出來。例如圖右上角的小線段在縱軸方向的範圍是 1-10；它在橫軸上的範圍是 1111-120。這就表示在序列中 1-10 附近的區域與 1211-110 附近的區域配對，形成二級結構。

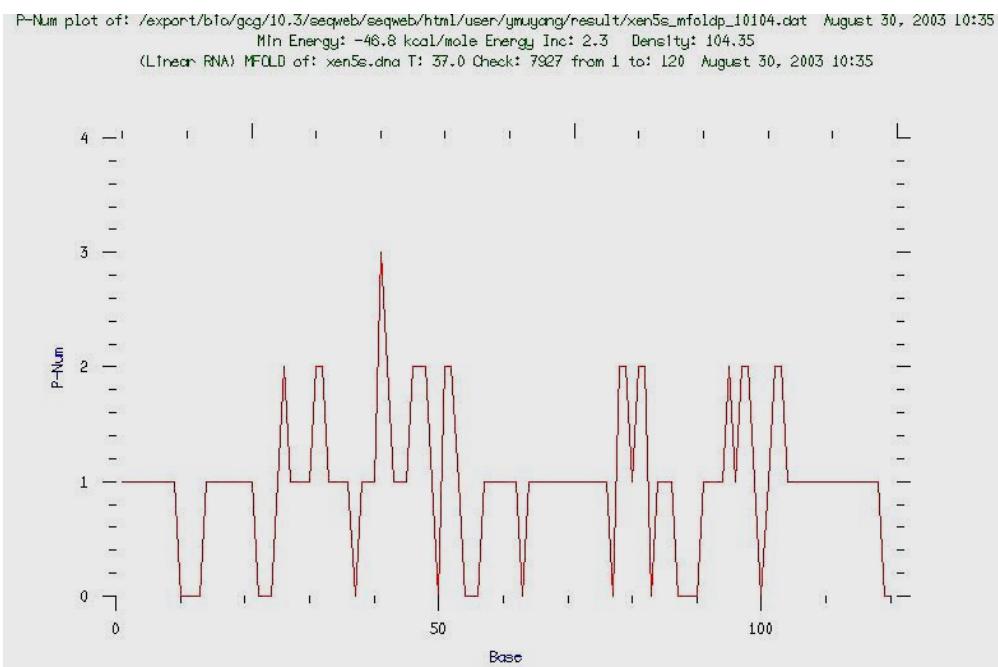
P-Num plot (圖 11-4B) 是另一種尋找穩定小結構的方式，此圖橫座標是序列的位置，縱座標則代表在給定能量增加值內之所有結構中，每一個位置上有多少種配對的方式：如果 P-Num 是零，代表這區域不參與配對；P-Num 若是壹（例如第 5 個鹼基處至第 10 個鹼基處），則代表在所有不同的結構中，均具有相同的配對方式，也就是結構都一樣的意思。其配對之方式可由 energy dotplot 中某一點的縱橫座標看出來。若是 P-Num 大於壹，則代表這個鹼基在不同的結構中可能會與不同之鹼基配對，若有 n 種配對方式，則 P-Num 就是 n，因此 P-Num 的最大值就是在給定之 energy increment 中的所有結構總數。我們有興趣的是這些 P-Num 值小的區域，因為這些區域代表的是穩定的結構，綜合言之，P-Num plot 顯示那些區域之結構較穩定，energy dotplot 則顯示這些穩定區域內之鹼基配對方式。

圖 11-4 分析 MFold 結果所用的 energy dotplot (A) 與 P-Num plot (B)

A.

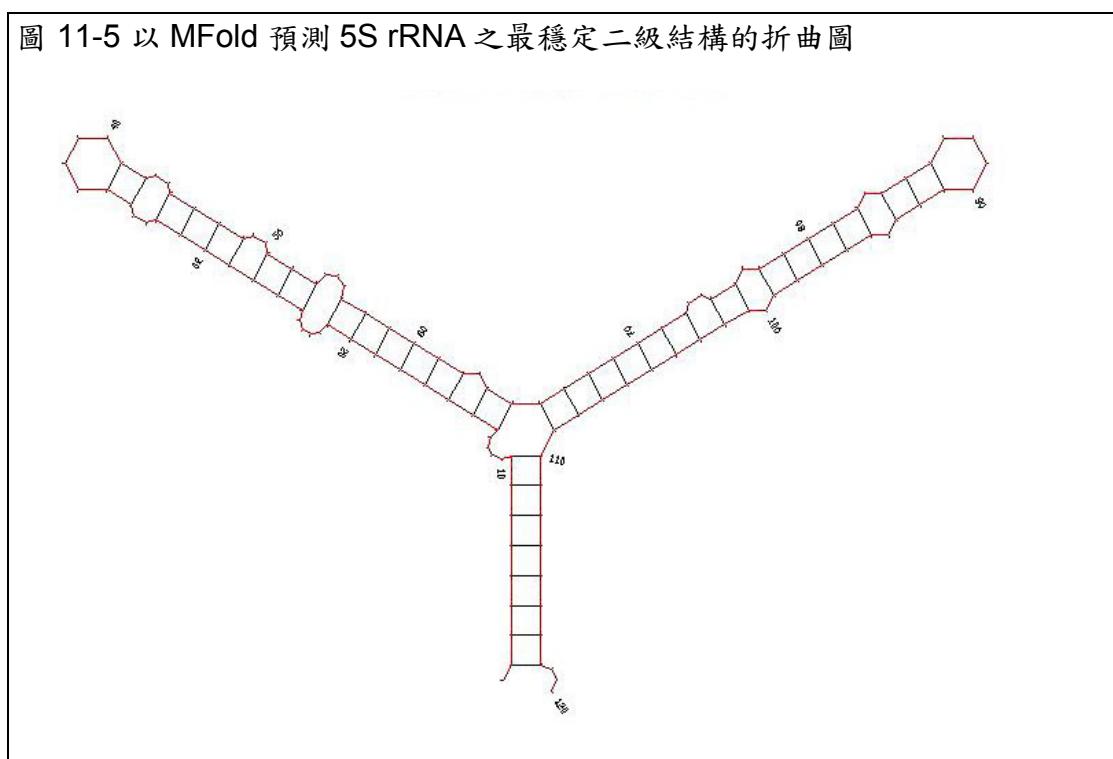


B.



若將 MFold 的結果以折曲圖的形式呈現（圖 11-5），其結果與目前已接受的結果相符，而且不論在 25°C 或 37°C 都得到同樣的結構。對 5S rRNA 而言，溫度由 25°C 升到 37°C，雖然所算出的自由能由 -53.6 kcal/mole 上升到 -39.4 kcal/mole，可是相對於其他的結構而言，它仍是最穩定的結構。這結果暗示 FoldRNA 在 25°C 所用的能量參數可能有問題，所以預測的結果比較不符合酵素切割的結果。

圖 11-5 以 MFold 預測 5S rRNA 之最穩定二級結構的折曲圖



雖然溶液中存在的結構，通常是由一些很穩定的區域結構(local structure)組成，可是整個分子的結構卻不見得是最穩定的。因為在 RNA 摺疊過程中可能有動力學障礙(kinetic barrier)，甚至有蛋白質參與結構的穩定作用，所以使得某些理論上相當穩定的區域結構無法形成（圖 11-6）。在求最低能量結構的過程中，可能因為用來分析的區域不完整，或是未考慮上述因素，使預測出的二級結構仍有許多錯誤。因此必須有其它的方法來驗證和改善預測的正確性。

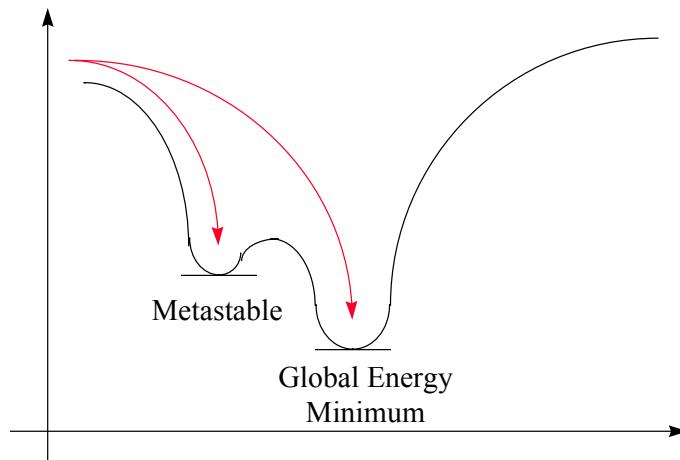
4. 限制條件的應用

最常用來探測在溶液中 RNA 二級結構的方法，是利用一些能區別 RNA 雙股區域(double-stranded region)與單股(single-stranded)區域的酵素或是化學物做 RNA 部份水解或是修飾。配合上高解析度的凝膠分析後，可以精確地判定某一個核苷酸具有單股或雙股的構形(conformation)，因而推論其二級結構。利用此法的好處很多，例如，反應通常只需微量的 RNA (pmole 之量即可)，而且可選用的酵素或是化學物很多⁽⁶⁾，不同的酵素或是化學物所得的結構資訊可相互印證。

此時可利用實驗上的觀察作為預測結構的限制條件(constraint)。例如，若由酵素切割的專一性知道某區域是處於單股狀態，即可在預測時禁止它與其他區域配對。再如親緣分析的結果若顯示某兩個鹼基會配對，亦可在預測時強迫這兩個鹼基配對。在執行程式時有一定的規則(convention)，來加入限制條件(參見方盒 11-2)。在表 11-2

中，列出部份酵素切割 5S rRNA 的結果，而根據親緣分析亦可判斷出的可能配對關係。在下面的範例與練習中，就將利用這些實驗結果，練習設定限制條件的方法。

圖 11-6 生物體內的 RNA 不一定是最穩定的結構



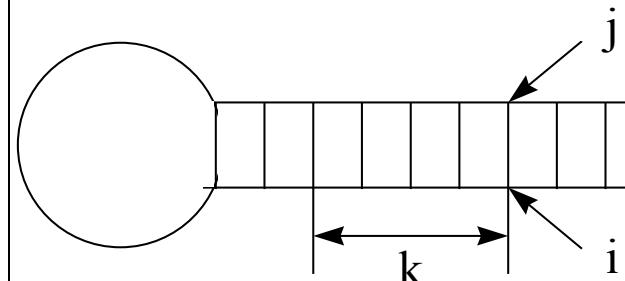
方盒 11-2 加入限制條件的規則

FoldRNA 與 MFold 共有的限制條件有兩類，分別是 `force` 與 `prevent`，前者是強迫某一區域形成雙股區，若已知是那一個區域形成配對，可用 `force=i, j, k` 表示，由圖 11-7 中可看出 i, j 分別代表一區域之起始點與終點的位置， k 代表自 i 起連續有 k 個鹼基會與終點處之各鹼基配對，在利用 RNase V 酶素做圖譜時，只知道某個區域是雙股區（即被切割），卻不知道這區域與何處配對，這樣可將 j 設定為零，程式就會找最恰當之區域與 i 到 $(i+k-1)$ 之間的區域配對，強迫其形成雙股。使用另一參數 `prevent` 時，其 i, j, k 之定義與 `force` 相同，而其效果卻與 `force` 相反。它禁止一區域與另一區域配對 ($j \neq 0$) 或禁止某一區域形成雙股 ($j=0$)。

在預測二級結構時為避免太複雜，有不允許結與假結 (pseudoknot) 之形成的假設。因此，一旦一個區域的起始點 (位置 i) 與終點 (位置 j) 形成配對，在此範圍內的鹼基就不可能與這個區域外之鹼基產生配對，也就是不會影響這區域外的 RNA 摺疊。在預測時可利用這特性除去某一區域，以減少預測所需的时间。例如根據實驗上之證據若已知某一區域是獨立摺疊的，即可將此區域除去，而預測剩餘區域之結構，其效果與利用 GCG 下的 `assemble` 程式除去序列中的某一段再來預測結構是相同的。這功能在 FoldRNA 與 MFold 使用不一樣的參數名稱，在 `foldrna` 中使用 “`-REMOve=i,j`” 而 `mfold` 中使用 “`OPENexcise=i,j`”，在使用時可使用多個，但必須依序使用。

在 MFold 中有另一種與 “`-open=i,j`” 相對的參數 “`CLOSEexcise=i,j`”，使用這參數時指定區域（通常是一個已預測成功的區域）的結構仍會出現在結果中，但不會受到其他區域的影響。其主要的目的是防止在強迫 (`force`) 或禁止 (`prevent`) 形成某些雙股區域時，使原已預測正確的區域受到影響。在 MFold 中還有另一獨有的參數是 “`-CIRcular`”，它會接受一個圓形（即首尾相接的序列）的 RNA 做摺疊預測，FoldRNA 不提供這些功能，因此在使用限制條件時不如 MFold 好用。

圖 11-7 雙股區域(Stem)上 i, j, k 的定義。



範例 11-4 請利用表 11-2 中數據微調25°C下所預測出的二級結構。

1. 編輯一名為「1013.ini」起始檔，其內容如下：

..	表示此行以上是comments
-infile=xen5s.dna	輸入檔檔案名
-beg=1	要分析的範圍起點
-end=120	要分析的範圍終點
-dat=25.en	指定自用數據檔為25°C時用的熱力學參數
-outfile1=junk	將表示結構的文字檔命名為「junk」
-outfile2=1013.con	將表示結構的連接檔命名為「1013.con」
-prevent1=10,0,4	禁止自第十個核甘酸起連續四個核甘酸與其它核甘酸配對

2. 以下列方式執行FoldRNA程式

% foldrna -init=1013.ini

3. 以Squiggles程式繪圖

% squiggles 1013.con (其結果與圖 11-2A完全相同)

用酵素或化學物做切割的缺點是不同位置被切割的效率會因三度空間之障礙而有差異，因此唯有被切割才能獲得資訊，換言之，某些區域卻可能會得不到二級結構的資訊。好在不同的酵素或化學物，有不同的專一性與動力學特性，所以使用多種酵素或化學物，即可同時探測到分子中各部份的構形。這些酵素或是化學物也能以footprinting的方式，分析 RNA 與蛋白質作用的位置^(7,8)。有時兩種甚至多種的酵素或化學物，可能有互相矛盾的結果（例如表 11-2 中灰底部分），造成許多困擾。在排除不相容的數據後，這些實驗上的證據可當做預測 RNA 結構時的限制條件，使程式的預測能與實驗結果配合。除了切割圖譜的資訊外，預測時的限制條件，亦可由親緣分析取得，例如圖 11-10 中雖只能看出兩個協同變異，當使用更多的序列做序列排比時，可看出 (68,107) 與 (69,106) 亦為協同變異。

表 11-2 *Xenopus laevis* 卵母細胞 5S rRNA 的切割
酵素位置(節錄自 Andersen et. al. (1984)⁽⁹⁾ 表一)

	A	T ₁	T ₂	V ₁	ss	ds	Conflicts
C ₁₀	+				1		1
A ₁₁			+		1		
C ₁₂	+				1		1
A ₁₃			+		1		1
C ₁₅				+		1	
G ₃₇		++			1		
U ₃₈			+		1		
C ₃₉	++		+		1		
C ₄₆				+		1	
A ₄₉			++		1		
A ₅₀			+		1		
A ₅₆				+		1	
U ₇₃				+		1	
U ₇₆				+		1	
A ₈₃			+	+	1	1	?
G ₈₇		+			1		1
A ₈₈			++		1		
G ₈₉		++			1		
C ₉₁				+		1	1
U ₁₀₂				+		1	1
C ₁₀₅				+		1	
C ₁₁₀				+		1	
C ₁₁₂				+		1	

* RNase A 切割 C and U; RNase T₁ 切割 G; RNase T₂ 喜歡
切割 A; RNase V₁ 切割雙股區域

練習 11-3 請利用(68,107)與(69,106)兩對協同變異為限制條件，預測出5S rRNA之二級結構。

1. 編輯68107.ini

```
.. 表示此行以上是comments
-infile=xen5s.dna 輸入檔檔案名
-beg=1 要分析的範圍起點
-end=120 要分析的範圍終點
-dat=25.en 指定自用數據檔為25°C時用的熱力學參數
-outfile1=junk 將表示結構的文字檔命名為「junk」
-outfile2=68107.con 將表示結構的連接檔命名為「68107.con」
-force1=68,107,2 強迫(68,107)配對、(69,106)配對
```

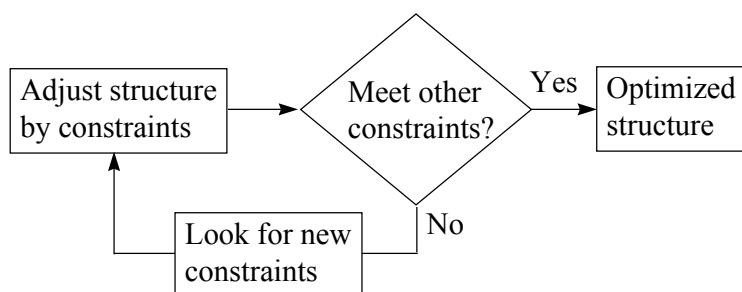
2. 執行FoldRNA與Squiggles

```
% foldrna -init=68107.ini
% squiggles 68107.con
```

3. 預測出的結構與圖 11-5 中的結構完全相同

因為 5S rRNA 是一個很小的 RNA，因此在給簡單的限制條件後就能很快得到符合大部份實驗數據的結構。可是對較大的 RNA 而言，可能需要反覆做許多次微調 (refine)。不論採用那一種方法找出限制條件，不要一次將所有的限制條件放入，不然就沒有其他的數字可驗證預測出的模型是否正確。因此在放入一些限制條件後，應檢視新預測的結構是否會更符合實驗數據，如果仍不理想，則應換其他區域再試。這一系列的步驟，要做到已無法再增加它與實驗數據的符合程度為止。以 5S rRNA 為例，在圖 11-11 中，標示「E」的區域附近有 A-A, A-G 等非傳統性 (noncanonical) 的配對。雖然此處有 V₁ 的切割（參閱表 11-2），程式根本不可能讓其配對，因此預測的結構不可能與實驗數據完全相合。

圖 11-8 RNA 二級結構的微調是一個遞迴的過程。



二、利用親緣分析預測 5S RNA 的結構

使用親緣分析法的先決條件是這個 RNA 在不同的生物中相似，卻有恰當的變異，太相似或變異太大都會造成困擾。以 HCV 的 IRES 為例，不同類別(type)間序列的差異僅 0.7%，未產生變化的區域就找不到協同變異，也就是沒有結構資訊。換句話說，用親緣分析法推測結構時，產生變異的區域才有機會提供確定的資訊，可以用來證明模型是否正確。而所謂守舊的區域則多少有些不確定性，因為現在不變，不代表將來不會變，所以它只能用來顯示相容性，而不能用來排除一些模型。

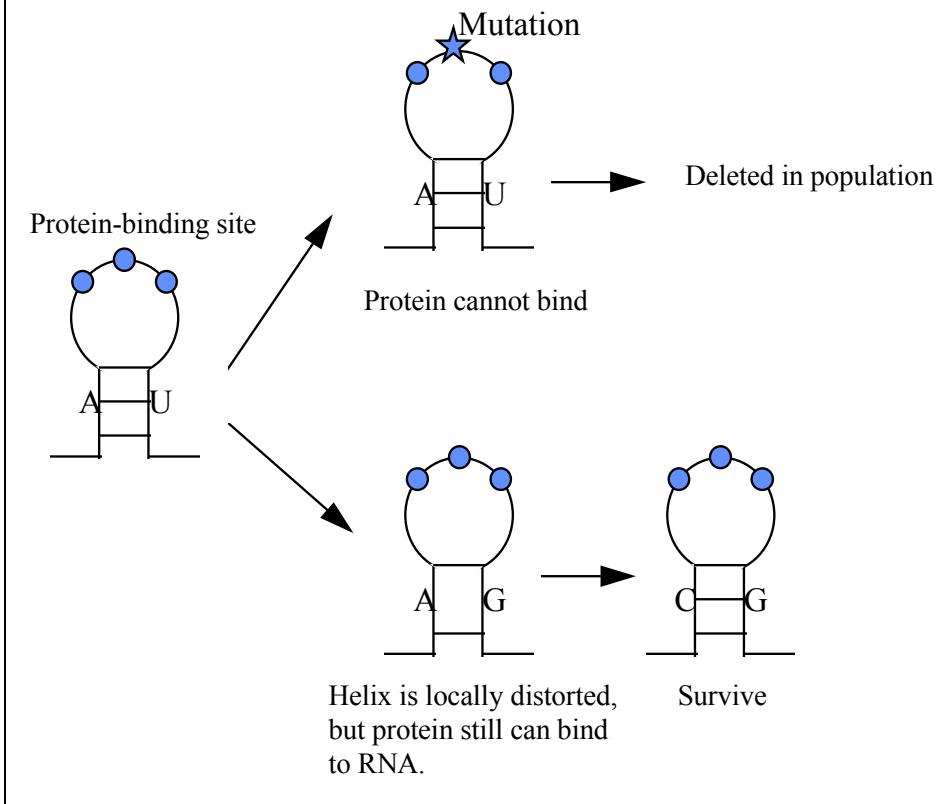
方盒 11-3 以親緣關係預測二級結構的原理

此法假設在演化的過程中，參與功能的 RNA 不易產生結構變化，否則會因喪失功能而無法流傳到後代。如果在結構中的雙股區有突變發生，為了維持 RNA 分子雙股結構區的特性，在演化的過程中可能會在相配對的鹼基上產生補償性的突變(compensatory mutation)，這一對突變稱之為協同變異(covariation)。例如，原本是 A-U 配對的雙股 RNA 經過補償性的突變之後可能產生 C-G 配對(圖 11-9)。一種預測結構的方法是以找到的協同變異為中心，尋找周圍區域配對的可能性，以建立二級結構。另一種做法是到電腦預測的諸多結構中，選擇符合突變數據的結構，再加修改而得到最佳的結構。

因為 RNA 分子的演化是在活體內進行的，所預測出的結構較有可能代表在細胞內的 RNA 結構，所以親緣分析被認為是目前預測 RNA 二級結構最準確的方法。利用親緣分析，已成功的預測出相當多的 RNA 二級結構包括 tRNA, hammerhead catalytic RNA，以及核糖體(ribosomal) RNA 的二級結構。除了預測 RNA 二級結構之外，利用親緣分析亦可分析在已知二級結構中，是否存在一些三級結構的作用，例如 Hoogsteen 鹼基對以及兩個 loops 之間的配對等。其預測的原理與 RNA 二級結構的預測相同，也是假設在 RNA 中，具有 triple base interaction 的鹼基對上，其中一個鹼基發生突變；而 RNA 為了保持原始結構的完整性，也會在相對位置上產生補償性的突變。group I 與 group II intron 產生三級結構的位置即是以親緣分析預測出的結果⁽¹⁰⁾。

在由序列預測二級結構時，可利用親緣分析的結果，調整預測的結構，若有足夠的親緣分析資訊，亦可直接由此分析預測結構。因為在資料庫中已有許多 5S rRNA 的序列，在此將以其為例，來說明利用親緣分析預測 RNA 二級結構的步驟。

圖 11-9 產生協同變異的可能機制



範例 11-5 自資料庫中取得5S rRNA序列。

```
利用FastA搜尋GenEMBL資料庫
% fasta -bat

FastA does a Pearson and Lipman search for similarity between a query
sequence and a group of sequences of the same type (nucleic acid or
protein). For nucleotide searches, FastA may be more sensitive than BLAST.

FASTA with what query sequence ? xen5s.dna

      Begin (* 1 *) ?
      End (* 120 *) ?

Search for query in what sequence(s) (* GenEMBL:* *) ?

What word size (* 6 *) ?

Don't show scores whose E() value exceeds: (* 2.0 *):

What should I call the output file (* xen5s.fasta *) ? xen5s.fas

** fasta will run as a batch or at job.

** fasta was submitted using the command:
      " atnow "

      warning: commands will be executed using /bin/sh
job 871415160.a at Tue Aug 12 12:46:00 1997
```

在做親緣分析時，太多變異會在序列排比時產生問題，太少變異又找不到足夠的協同變異，所以在分析時必須慎選序列的種類。在此使用 FastA 的主要原因是它能提供序列排比的結果，有助於選擇適當的序列做分析。此外、自 GCG 第九版起，FastA 的輸出中加列出各資料庫序列中，能與查詢序列並列的區域(練習 11-4 中 Begin 與 End 的部份)，更能增加做 PileUp 的成功的機會。在使用 Fasta 程式時，若在指令行加上「-noalign」，所產生的檔案不顯示序列排比的結果，這樣雖然可以減少檔案的大小，可是卻失去了選擇適當序列的依據，使用者可以根據自己的需求決定要不要用此功能。在此勿用 StringSearch 程式來找序列，避免產生困擾。

練習 11-4 編輯「xen5s.fas」，取得分最高的29個序列做多序列排比

Answers:

1. 修改FastA的輸出檔，以「//End of list」使第29個序列之後的序列改變為「說明」。

!!SEQUENCE_LIST 1.0

(Nucleotide) FASTA of: xen5s.dna from: 1 to: 120 August 12, 1997 12:58

...

```
GB_OV:AMTRGA Begin: 1 End: 119
! L49407 Ambystoma salmoides 5S ribos... 523 523 523 580.7 1.4e-23
GB_OV:RP5SRRNA Begin: 1 End: 119
! X58368 R.pipiens gene for 5S riboso... 514 514 514 570.9 5e-23
GB_OV:BA5SRRNA Begin: 1 End: 119
! X58365 B.americanus gene for 5S ribo... 514 514 514 570.9 5e-23
GB_OV:XTRRN5SG Begin: 61 End: 180
! X12624 Xenopus tropicalis oocyte 5S... 510 510 510 563.1 6.1e-23
\\End of List
```

```
GB_OV:GATATAT Begin: 57 End: 176
! M74438 Gastrotheca riobambae 5S rRNA... 501 501 501 547.3 1.2e-22
GB_OV:XTRRN5SO Begin: 1 End: 120
! X12623 Xenopus tropicalis oocyte 5S... 500 500 507 563.2 1.3e-22
GB_OV:NVIRGAA Begin: 32 End: 151
! M13611 Newt (Notophthalmus viridescens) 495 495 501 553.8 2.3e-22
GB_OV:NV5SRRN Begin: 32 End: 151
! X06097 Notophthalmus viridescens 5S... 495 495 501 553.8 2.3e-22
GB_ST:MIPRRA Begin: 1 End: 120
```

.....

2. 利用PileUp與Pretty程式，尋找協同變異

% pileup @xen5s.fas

產生「xen5s.msf」檔

% pretty -con -dif=* xen5s.msf{*}

結果參閱圖11-10

FastA 的輸出檔案是一個檔名檔案，可利用「!」或「//End of list」將某序列或某一行之下的所有序列改變為「說明」，而不干擾下一個程式的運作。

圖 11-10 5S rRNA的序列排比之結果，由此可看出C18與G60，C30與G47有協同變異

Plurality: 2.00 Threshold: 1 AveWeight 1.00 AveMatch 1.00 AvMisMatch 0.00

PRETTY of: course.msf{*} August 12, 1997 13:19 ..

1

50

```
course.msf{RP5SRRNA} ***** ***** *****C***** ***** ***** ****t
course.msf{BA5SRRNA} ***** ***** ***cac*** ***** ***** *****
```

```

course.msf{XTRRN5SG} *****t***** *t***** C**** ******t***** ****
course.msf{AMTRGA} *****t***** a** ***t*c***t *****t***** a**
course.msf{XELRRAB} *****t***** *****t***** t*****atc tg****a***
Consensus GCCTACGGCC ACACCACCT GAAAGTGCCT GATCTCGTCT GATCTCGGAA

      51                               100

...
course.msf{RP5SRRNA} ***t*a**** ******t***** *****t***** ****
course.msf{BA5SRRNA} ***t*a**** ******t***** *****t***** ****
course.msf{XTRRN5SG} ***t*a**** a*****t** *****t***** ****
course.msf{AMTRGA} ***t*a****t *****a***** *****t***** ****
course.msf{XELRRAB} *****a**** **.***** *****t***** ****
Consensus GCGATGCAGG GTCGGGCCTG GTTAGTACCT GGATGGGAGA CCGCCTGGGA

...
course.msf{RP5SRRNA} *****C*** C*****~
course.msf{BA5SRRNA} *****~ ****
course.msf{XTRRN5SG} *****~ *t*****
course.msf{AMTRGA} *****~ ****
course.msf{XELRRAB} *****~ ****
Consensus ATACCAGGTG TCGTAGGCTT

```

這個練習的目的是提醒使用者選擇適當的序列是親緣分析法的重要步驟。在這個例子並未特別選擇差異適中的序列做序列排比，因此只找到少數有用的協同變異。

三、預測 RNA 結構的問題

1. 「結」及「假結」結構的出現

為了降低預測過程的複雜程度以利計算，RNA 二級結構預測程式排除具有結及假結結構的可能。然而根據目前的瞭解，結或假結的結構確實在生物體內存在，甚至在結構或功能上具重要角色⁽¹¹⁾，例如轉譯過程中的 frameshift 訊號⁽¹²⁾與 D 型肝炎病毒的 ribozyme⁽¹³⁾即是知名的的例子。於此狀況下，此種簡化方式顯然已經引入了結構預測失敗的因素。

2. 能量參數之準確性

如前所述，FoldRNA 係以找尋能量最低結構的策略來預測 RNA 二級結構。用以計算結構能量的能量參數顯然在結構預測之正確性上佔著重要的地位。目前用以預測結構之能量參數，均由實驗測量所得，任何測量均有誤差，此一測量自然也不例外。在預測較小的 RNA 分子之結構時，此類誤差影響不大。但在預測較大的 RNA 分子結構時，誤差將重覆累積為一可觀之數值，以致於能量最低點附近的數個不同結構間之能量差距小於誤差的累積值。此時，我們將無法斷定那一個預測之結構才是最為穩定的結構，更遑論此穩定之結構是否即為實際存在之結構。要解決此一問題，提昇實驗測量時之精確度無疑是最直接有效的方法。

3. 預測策略之合理性

FoldRNA 程式預測結構的策略乃是尋找能量最低之結構，但是這樣一個假設是否切合實際呢？以化學的觀念來衡量此一假設，其中大有可議之處。如果僅就熱力學部份進行討論，物質確有形成最低能量結構之傾向，然而此一最低能量狀態是否確能形成，則又牽涉了動力學的因子。如果形成能量最低結構的過程中必須克服極高之活化能障礙，在環境無法提供足夠能量的狀況下，所形成之結構將被局限在一個區域性的能量最低點(local energy minimum)。即使僅以熱力學討論，當形成兩個不同結構的自

由能落差均極接近能量最低點時，兩者理應相互達成平衡，此二結構各佔部份，二者間能量差距愈小，二者比例愈接近。以此觀之，要把能量最低之結構和真實存在之結構二者直接劃上等號，實屬過於理想。如能藉由其他方法(如親緣分析法或水解酵素切割圖譜分析)獲得關於結構的資訊，則可在預測結構時加入限制條件，此時所獲得之結構將是在此限制條件下能量最低的結構。因為此結構之能量必定高於未給限制條件前之最低能量，此一做法相當於尋找區域性的能量最低點(即自然狀態下的「次穩定狀態(metastable state)」)。所以經由實驗的幫助，我們可以克服理論與真實世界之差距，預測出水溶液中可能之二級結構。

4. 以分子生物學方法找尋結構可能遇到的問題

執行摺疊 RNA 的程式時，限制條件可以是一個構形狀態，也可以是一組配對的關係。圖譜分析可以提供單、雙股的構形資訊，因為較易獲得數據，在文獻中較常見到這種方法。而親緣分析法可以利用協同變異，來推測相互配對的鹼基。

(a) 以圖譜分析(mapping)的結果，微調(refine)預測的 RNA 結構之問題

不論是用酵素或是化學物來探測水溶液中的 RNA 結構，都是在區辨結構中不同鹼基的構形，可是它無法探測究竟是那兩個鹼基互相配對。因為圖譜分析法不知配對的對象，因此當結果顯示預測的結構能符合 80% 的圖譜分析資料時，並不表示這結構是 80% 正確。可能有些區域處於雙股狀態，卻與錯誤的區域產生配對。至於出現在單股狀態的區域，出現在髮簪狀(hairpin)結構的末端，與以突出(bulge)形式出現在雙股區的中央的兩種情形，很顯然代表不同的結構，可是圖譜分析的結果，卻無法區辨這兩種單股構形的差異。

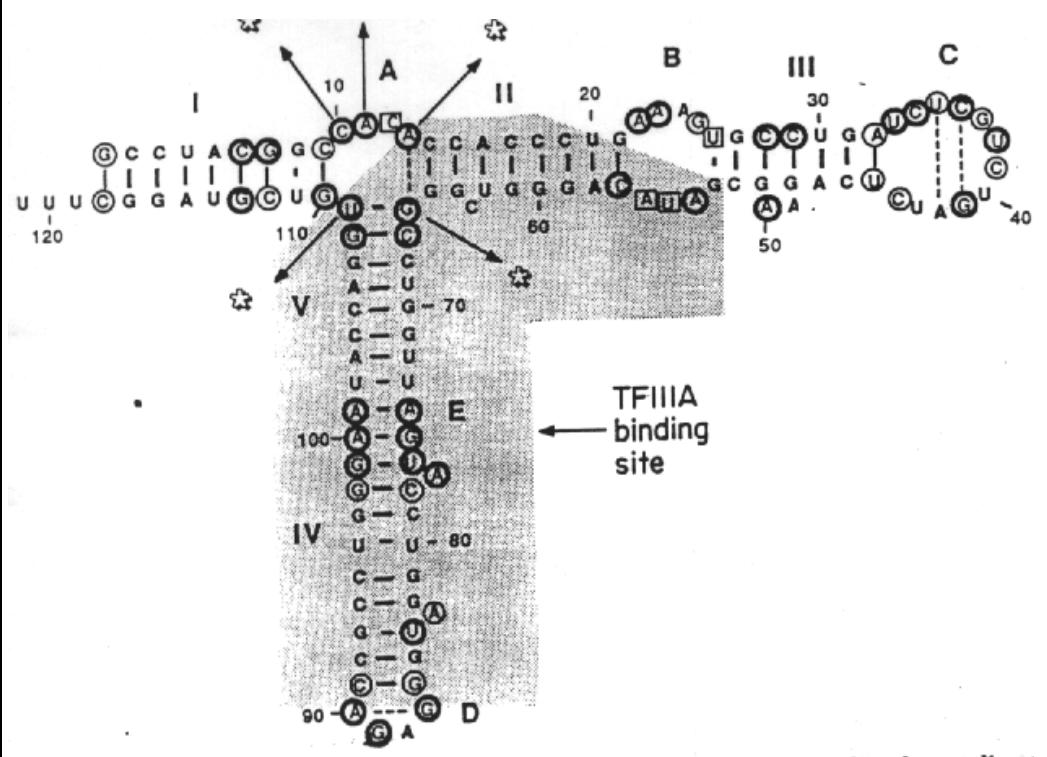
在使用圖譜資訊作為限制條件，微調預測的 RNA 結構時，經常在做到某種程度後，就無法再提高符合圖譜資訊的比例。此時一些不同的結構模型，可能都有相似的符合比例。在沒有新的資訊區辨這些結構模型時，沒有理由任選一個再繼續作微調。事實上對兩個結構很不相同的模型而言，在不同結構符合圖譜資訊的比例，相差在 5% 之下時，恐怕就無法判斷哪一個結構比較正確了。唯有在預測的不同結構模型都很相似時，也就是不需要耽心找到一個完全不同的結構時，才可用圖譜資訊做限制條件繼續微調。

(b) 以親緣分析的結果，微調(refine)預測的 RNA 結構之問題

要由親緣分析得到更多的結構資訊，就需要瞭解產生協同變異的機制。在理論上，有協同變異的區域的 RNA 結構應相當穩定，這樣在發生點突變時才不會喪失全部活性，而有機會等待在相配對的位置也發生補償性的突變(圖 11-9)。換言之，在預測 RNA 結構時，被用來做親緣分析的序列，也就是在演化過程中可以產生變異的區域，並不是天擇的主要對象。要找天擇的主要對象，就要找在演化過程中守舊的區域。

雖然自然界中也有雙股 RNA 接合蛋白存在^(14,15)，大部份的 RNA 接合蛋白都接合在 RNA 的單股區域⁽¹⁶⁾。這是因為 RNA 的雙股區域是 A 式雙螺旋，因此蛋白質不易嵌入主溝槽(major groove)與鹼基對形成氫鍵。單股的區域既是蛋白質的接合位置，在演化過程中，若無法與蛋白質作用，可能會失去某些功能而導致個體的死亡，因此我們認為 RNA 上的單股區域，可能比雙股區域守舊。這似乎與直觀的想法相違，因為單股區域的結構似乎不會造成 RNA 結構上的主要變化。檢視了許多同時有親緣分析和二級結構資訊的 RNA 後，我們發現守舊的區域幾乎都出現在單股區域，僅以圖 11-11 中的 5S rRNA 為例(幾百個序列的統計結果)說明單股區較守舊的事實。因此要如何應用守舊的區域預測 RNA 結構是一個值得深入思考的問題。

圖 11-11 守舊的核苷酸(○、□)多位於 RNA 的單股區域。



四、結語

Foldrna 的基本假設既是尋求最低能量的結構，在預測比較穩定的結構時，其預測成功的比例自然會較高。但問題是我們要怎樣區辨那一個 RNA 片段是比較穩定的結構，而且這樣的結構是否有生物上的意義？由各位的習題中 Vg1 mRNA 定位訊號 RNA 二級結構的預測，可知定位訊號 RNA 及其內的蛋白質接合區等具有重要功能的區域，能獨立的摺疊為特定的結構。使用 Foldrna 程式可預測出獨立摺疊的區域，因為在各種刪除式突變中，這區域的結構始終不變。從理論上分析這區域之所以會獨立摺疊，就是因為其結構很穩定，因此是很適合用 FoldRNA 或 MFold 來預測的。

參考文獻

1. Tinoco I. Jr., Borer P. N., Dengler B., Levin M. D., Uhlenbeck O. C., Crothers D. M., Bralla J. (1973). *Nat New Biol.* 246(150):40-1.
2. Cech T. R., Tanner N. K., Tinoco I. Jr., Weir B. R., Zuker M., Perlman P. S. (1983). *Proc Natl Acad Sci U S A.* 80(13):3903-7.
3. Turner D. H., Sugimoto N., Jaeger J. A., Longfellow C. E., Freier S. M., Kierzek R. (1987). *Cold Spring Harb Symp Quant Biol.* 52:123-33.
4. Tinoco I. Jr., Borer P. N., Dengler B., Levin M. D., Uhlenbeck O. C., Crothers D. M., Bralla J. (1973). *Nat New Biol.* 246(150):40-1.
5. Zuker M., Jaeger J. A., Turner D. H. (1991). *Nucleic Acids Res.* 19(10):2707-14.
6. Jaeger J. A., SantaLucia J. Jr., Tinoco I. Jr. (1993). *Annu Rev Biochem.* 62:255-87.
7. Stern S., Wilson R. C., Noller H. F. (1986). *J Mol Biol.* 192(1):101-10.
8. Murakawa G. J., Nierlich D. P. (1989). *Biochemistry.* 28(20):8067-72.
9. Andersen J., Delihas N., Hanas J. S., Wu C. W. (1984). *Biochemistry.* 23(24):5752-9.
10. Michel F., Umesono K., Ozeki H. (1989). *Gene.* 82(1):5-30.
11. Pleij C. W. (1995). *Genet Eng (N Y).* 17:67-80.
12. Brierley I., Digard P., Inglis S. C. (1989). *Cell.* 57(4):537-47.
13. Perrotta A. T., Been M. D. (1993). *Nucleic Acids Res.* 21(17):3959-65.
14. St Johnston D., Brown N. H., Gall J. G., Jantsch M. (1992). *Proc Natl Acad Sci U S A.* 89(22):10979-83.
15. Bass B. L., Hurst S. R., Singer J. D. (1994). *Curr Biol.* 4(4):301-14.
16. Michel F., Westhof E. (1994). *Nat Struct Biol.* 1(1):5-7.

第十二章 序列資訊之應用

黃彥華¹、楊永正²

¹英國劍橋大學桑格研究所、²陽明大學生物資訊研究所

一、 簡介

人類基因體研究計劃(Human Genome Project)已在 2003 年完成，這些大量的序列資訊，將對生物學的研究方法產生革命性的影響。這種影響，事實上在一些病毒的研究上早已出現端倪，以下將以對分子生物學發展有重要影響的腺病毒(adenovirus)為例，說明 DNA 序列對研究方法所產生的衝擊。

早在取得腺病毒的全序列之前，分子生物學家就已利用核酸雜交的技術，找到在感染細胞中所表現的病毒基因。這些表現的基因以 DNA 的複製作為分界點，分為早期(表現於複製之前)與晚期(表現於複製之後)兩種型式。利用生化的方法進行研究，發現最早表現的是 E1A 基因，其蛋白質產物會活化其他的基因表現。而利用遺傳的方法，亦發現在 E2 上有 DNA 接合蛋白與 DNA 合成酶的活性。而在晚期所表現的主要基因(major late)，更是發現 mRNA 剪接現象的一個重要關鍵。這些工作顯示，在沒有詳細的 DNA 序列的情況下，也能發現與解釋許多重要的生物現象。

可是在沒有序列資訊之下的生物學研究，其進展的速率，是遠低於知道序列後的。以遺傳方法找尋基因功能，必須要篩選突變株，再利用功能上的改變，以推論突變基因的功能；但是要找到具有明顯表現型、而且是單基因突變的突變株，是相當耗費時間及精力的。相反的，在知道腺病毒全序列後，對於所有開放讀架的位置都很清楚，因此可以很快的用一些簡單實驗加以驗證，哪一些開放讀架可真正轉譯出蛋白質；分子生物學家便可更進一步的用電腦，針對欲研究的序列，尋找適當的限制酵素切割位置，以便做分子選殖或設計探針。其次，也可以利用核酸探針與對特定蛋白質產物產生的抗體，輕易找出基因或蛋白質表現的先後次序。此外，也可根據序列設計突變，然後再利用探針與抗體來觀察突變所造成的影響，而不必單靠表現型的變化。利用這樣的方法可以找出各個不同基因的功能，這些都是在有序列之前很難做到的。

縱使有了全序列之後，要純粹以實驗的方式，來決定各個基因的功能，也不是一件容易的事；若能設法找出在序列上的特徵，加以註解，便可以使生物學家很快的掌握一些可能的方向。因此在這裡舉一個例子說明在電腦工具都很缺乏的狀況下，怎樣可以結合序列資訊與生化技術，漂亮地猜出基因的功能。在腺病毒的突變株中，有一株對高溫敏感的 ts1 突變株。利用所謂 marker rescue 的技術與序列分析，可將此突變的位置精確地定出來，恰巧是落在一個能產生 23kd 蛋白質產物的開放讀架上。在另一方面，實驗也發現，ts1 感染的細胞在它不能生長的溫度 (nonpermissive temperature) 下，幾個病毒的蛋白質分子量較大，似乎是有蛋白質切割的問題。因此有人利用被感染細胞的抽出物，證明在抽出物中具有切蛋白質中 Gly-Ala 鍵的能力。在序列上確實也可以發現 23kd 的蛋白質具有一般含 Ser 的蛋白酵素中可見的守舊序列。可是在野生型腺病毒感染的細胞中卻找不到 23kd 的蛋白質，只能看到一個約 19kd 的蛋白質。經序列比對，發現在 23kd 的開放讀架中也可以找到 Gly-Ala 鍵，若在此產生切割，則可產生分子量約 19kd 的蛋白質。Chatterjee 等人據此設計實驗，證明 19kd

蛋白質是一個蛋白酵素，而且它的 peptide 圖譜與被 *tsl* 感染的細胞中的 23kd 蛋白質之圖譜相似。因此推測在 *tsl* 中，Pro 因為產生點突變而變為 Leu，在 42°C 失去了自我切割與切割其他病毒蛋白質的能力。

這一個例子，很清楚的告訴我們，在有序列資訊的狀況下，很容易地就可以將產生突變的位置決定出來。配合上突變的表現型與序列資訊，馬上就可建立、並測試假說，證明一個開放讀架所產生的蛋白質的功能。可是，在沒有序列資訊的狀況下可能要花數倍的時間才能得到這樣確定的結果。讓我們想像，若要以差異顯示法來找突變株與野生型感染的細胞中病毒基因表現的差異，即使可找到差異，恐怕反而誤導了研究的方向，因為 mRNA 的量即使有變化，也只是由蛋白質未被適當切割所造成的次級效應。此外，以篩選突變株來發現基因的功能，而對於複雜而壽命較長的生物，如老鼠、甚或是人而言，是非常花費時間，而且較為困難。

在沒有基因體序列資訊的狀況下，對於致病基因，要做所謂的「定位選殖 (positional cloning)」是一件大工程，可能要花上數年至十年以上的功夫。這種窘境，將因為人類基因體計畫的完成而大為改觀。疾病研究者可將心力放在連鎖(linkage)分析，待將突變定到一小區域後，再由資料庫中查閱突變區的序列。因此可將做選殖與大量序列分析的心力，用來尋找致病基因造成疾病的機制。知道致病基因的序列，對於瞭解疾病產生的原因是有很大幫助的，例如導致纖維囊腫(cystic fibrosis)的基因是一個氯離子通道，就是透過序列比對而猜到，再經過實驗驗證的。雖然人類基因體序列的完稿已在今年完成，還是有許多基因的功能仍不清楚，可是核酸及蛋白質序列中的模組，有助於我們瞭解基因的可能功能，我們對於序列知道的越多，就越有機會找到所有的模組。

在以下的討論中，我們將針對 EST 序列的應用做進一步的說明，因為迄至 1995 年 7 月 15 日，序列資料庫中的 EST 序列已多於其他的序列的總和 (至 2002 年 10 月 15 日止，共計有 13,386,349 筆 EST 序列，佔 GenBank 全部 19,808,101 個序列的 67.6%)(欲得最新的統計資料，可以用 anonymous ftp 連線至 <ftp://ncbi.nih.gov/genbank/> 取得 *gbrel.txt*)。因此在利用差異顯示(differential display)，酵母菌雙雜交(yeast two hybrid)等方法尋找具有生物意義的基因時，有相當大的機會可以找到一些具有相似序列的 EST。

二、表現序列標幟 (expressed sequence tag , EST)

在開始進行基因體分析計劃的初期有兩派決定序列的策略，因為細胞中有許多「junk」DNA，有人主張直接尋找在細胞中會表現的 RNA，這樣可以直接找到有功能的序列。而另一種看法是認為決定基因體的序列較為直接，而且可以獲取所有的人類序列資訊，不必擔心是否誤失了表現量很低的 RNA。因此基因體分析計劃採用決定整個基因體序列的作法。

另外一方面，如果就序列資訊應用的角度來看，在不同的發育時間、不同的細胞所表現的 RNA，不但暗示著基因調控時的階層關係，也可能顯示不同細胞間交互作用的方式。既然已要決定基因體的序列，在研究基因表現時，沒有必要去決定整段基因的序列，只要決定一小段代表性的 cDNA 片段序列，足以顯示這個表現基因的特性，待確定了該序列可能和哪種重要疾病有高度相關之後，再查閱基因體序列資料庫，便可快速的取得該基因全序列。因此，表現序列標幟(EST, expressed sequence tag)的

觀念，便因而產生。它是利用聚合酵素連鎖反應(PCR)以隨機引子夾出的部份 cDNA 序列，因為轉譯區的序列可以和蛋白質序列比對，不但降低了序列比對時的雜訊，也增加了推測基因功能的機會。

經由比較正常細胞與不正常細胞基因表現的差異，或許可以找到產生病變的原因，或是達到改良品種的目的。因為在醫藥研發上有很高的潛在價值，所以在 1990 年代，許多生物科技公司(如 Incyte)有系統地決定 EST 的序列。又如 Merck 公司支持的研究計劃，每日產生超過一千個 EST 序列。目前資料庫中 EST 的序列數目已超過其他所有序列的總數，因為有這些 cDNA 已被有系統地整理出來，目前對於生物學的研究已有非常大的助益。

1. 「表現序列標幟」的分析是解讀藍圖時的“法眼”

生命現象是展現在巨分子的交互作用上的，因此建立交互作用圖譜有助於瞭解各反應的機制。想在短時間內建立交互關係，最實際的方法，不是由序列去預測，而是由實驗去決定交互作用圖譜。目前有許多方法可協助我們找尋和已知巨分子有交互作用的分子，例如 far-Western、two-hybrid system 等，可用來找尋相互作用的蛋白質；而 South-Western、North-Western、*in vitro* evolution 等，可用來找尋和核酸交互作用的蛋白質分子。此外、差異顯示(differential display)或差異基因庫(differential library)的篩選等方法也有決定是哪一種巨分子開啟或關閉了一系列的基因表現的潛力。過去使用這些方法時，共同的困擾是必須花時間去決定所找到的基因的整個 cDNA 全長序列。雖然基因體序列仍然不完全，若發現有生物意義的序列和「表現序列標幟(EST, expressed sequence tag)」相似，即可訂購含此 EST 的 cDNA 以增加決定 cDNA 序列的效率。

2. EST 資訊的使用

因為多數的 EST 是別人用 random priming 的方法由 cDNA 基因庫(library)中所決定的部份 cDNA 序列，而含 cDNA 的菌株都已經過整理，只需要提供含此 cDNA 的菌株名即可買到，這樣可以省去許多篩選 cDNA 基因庫的精力。目前可由下列三個經過認可的機構購得這些含 cDNA 的菌株：ATCC(American Type Culture Collection)、屬於 Invitrogen 的 Research Genetics 及 Genome System。在此將介紹如何由 Research Genetics 的全球資訊網站上取得資訊。請查看國家衛生院的全球資訊網來瞭解其使用方式。

三、應用 EST 資訊的實例

1. 人類的 eIF4G 與 EST 中的序列相關但不相同

一個在肝癌與正常組織有差異表現的一個序列片段為例，在(1997 年)用 Blast 搜尋 EST 資料庫時，發現它與編號 W39270 的序列在 265 bp 的重疊區域內有 92.5% 的核苷酸相同；再用 Blast 搜尋 GenEMBL 資料庫，則發現它與一個小鼠(mouse)的 eIF4G2 相關基因具有高度相關性，經分析後發現在 498 bp 的重疊區域內有 91.2% 的核苷酸相同(在蛋白質序列上，則在 161 個胺基酸的重疊區域內有 90.1% 的胺基酸相同)。因此這個具差異表現的序列片段，以及找到的 EST 序列可能都是源自於同樣的一種人類的 eIF4G2 相關基因。

為確定這種人類的 eIF4G2 相關基因沒有被進一步分析，除了利用 E-mail 向 info@image.llnl.gov 確認沒有更新的資訊外，也以小鼠的 eIF4G2 相關基因为查詢序列，用 FastA 程式查閱序列資料庫，結果可找到人的 eIF4G2 基因。若交互比較不同來源的 eIF4G2 相關基因之間序列的相似性，則可以發現人類和兔子的 eIF4G 基因，mRNA 有 89.1% 相同(表 12-1A 第二列所示)，蛋白質則有 84.7% 相同(表 12-1B 第二列所示)，兩者應是同源基因；而小鼠的 eIF4G2 相關基因則和人類或是兔子的基因呈低度相關，只在蛋白質序列上有 27% 左右相同(表 12-1B 第三列所示)，mRNA 的序列則因差別太大，只能在非常侷限的範圍內有相似序列(表 12-1A 第三列所示)。

此外，人類的 EST W39270 的序列則只有和小鼠的 eIF4G2 相關基因較為類似(表 12-1A 第四列所示)，兩者的 mRNA 序列有 97.1% 相同。綜合以上的分析，人類應具有相對於小鼠的 eIF4G2 相關基因的一種新的基因，而這種基因可能會在肝癌組織有差異性表現。我們認為這種差異表現可能具有生物意義，因此希望做進一步的分析，這當然包括要得到基因的全序列。

表 12-1. 各種不同來源的 eIF4G 相關序列的相似性比較。

A. 以 mRNA 序列互相比較。B. 以蛋白質序列互相比較。

A.

DNA 排比		Human 5018 bp	Rabbit 4690 bp	Mouse 3792 bp
Rabbit 4690 bp	Identity	89.1%		
	Aligned Length	4675 bp		
Mouse 3792 bp	Identity	58.7%	60.4%	
	Aligned Length	252 bp	288 bp	
EST W39270 425 bp	Identity	93.8%	72.2%	97.1%
	Aligned Length	16 bp	36 bp	412 bp

B.

蛋白質排比		Human 1396 aa	Rabbit 1402 aa
Rabbit 1402 aa	Identity	84.7%	
	Aligned Length	1405 aa	
Mouse 907 aa	Identity	27.0%	27.2%
	Aligned Length	954 aa	960 aa

為了要加速研究的速度，便思考如何充分利用 EST 序列資料庫。由於 EST 序列資料庫資料增加得很快，藉由這些 EST 之間端點的相似性，來分類並且銜接各個序列，很可能可以連接成較長的基因接合序列，甚至由 EST 序列直接組合出某個特定人類基因的全序列。如此一來，便可以根據預測出來的接合序列設計引子，很快的由實驗上取得基因的全序列，進入功能性分析的階段；另一方面，在只能找到非常少數的 EST 序列的狀態下，則可能可以找出含有這個 EST 序列的 cDNA 菌株，進一步定出 cDNA 其餘的序列。

2. 以人類的 EST 資料重組一個新的 eIF4G2 相關基因

為了達到上述的目的，可以有兩種方式：一種是利用全球資訊網 (<http://www.ncbi.nlm.nih.gov/UniGene>) 查閱 NCBI 的 UniGene 資料庫 (Unique Human Gene Sequence Collection)(參閱方盒 12-1)，這是 NCBI 以電腦程式根據序列相似性，對 EST 序列自動進行分群所得到的資料庫，其目的是期望能由 EST 序列分群組合，來降低目前資料庫中序列的重複性(redundancy)。另一種利用 EST 序列資料庫的方式，則是以 Blast 直接去搜尋 NCBI 所搜集的最新 EST 序列。

方盒 12-1 生物學資料庫與 UniGene

GenBank 是最早開始蒐集 DNA 序列的資料庫，因為其主要目的是收集序列以供比對，資料中所列的許多欄位只是記錄著描述這個基因性質的資訊，而不是設計用來做搜尋的。在人類基因體研究計畫開始進行後，當人們要試著使用序列之外的這些資訊，發現 GenBank 中的重覆資訊太多，而且不能很容易地自其中選出想要的資訊。因此就開始有人重新整理 GenBank 所收錄的原始序列，並提供搜尋工具，使我們能由生物學的角度來使用與序列有關的資訊。

這些新的資料庫，有些是替使用者先做序列比對，除去重覆的資訊，並將相關的序列放在一起；另一些是整理 GenBank 中那些無法做搜尋的欄位，在經過整理後，使用者可利用一些有生物學意義的分類方式來搜尋，例如根據基因的俗名，基因在細胞中之功能等。在全球資訊網上 NCBI (<http://www.ncbi.nlm.nih.gov>) 有 UniGene 資料庫，TIGR (<http://www.tigr.org/tdb>) 有 HCD，EGAD 及 SST 等三個資料庫提供這方面的服務。

因為基因才是執行功能的一個基本單元，所以人類基因體研究計畫的最終的目的，是希望瞭解在不同的發育階段與不同的細胞中，各個基因表現的情形。據估計，在人類基因體之中大約只有 3% 是真正基因的序列。如果沒有一個有系統的方法可以定出各個基因在基因體之中獨特的位置，那麼將難以發揮基因體序列資料的真正用處。EST 的觀念的提出，其目的之一便是在於建立轉錄單元圖譜(transcript map)，以輔助人類基因體研究計畫，作為在基因體序列上直接進行「定位選殖」的關鍵工具。但是要達成此一目的，EST 的資訊必須要具有獨特性(unique)才行，否則，不同的 EST 序列對應至相近的實際基因體位置，就不容易分辨到底在這個區域之內具有數個基因，抑或是只有單一基因存在。

事實上，UniGene 就是將 EST 的序列，透過與功能已知的基因比對、及各個序列之間的相似性，建立了一些獨特(unique)的群組(cluster)，其目的是為了降低 GenBank 中資料的重複性，以提高資訊使用的效率。後續不斷新產生出來的 EST，即可與這些獨特的群組相比較，若不相同，即獨立為一個新的群組，以用來辨認尚未被發現的基因。因此，各個不同的群組可以獨特地(uniquely)代表單一基因的部分序列(或全序列)，這對尋找基因體中可以轉錄的區段，即建立「轉錄單元圖譜」非常有用。在決定 EST 序列時都會紀錄這個序列來自於那一個 cDNA 菌株，根據這個資料，不重疊的 EST 亦會被歸為一群。此外，在使用 EST 搜尋資料庫時，可以利用相似性找到這個 EST 與哪一個基因有關。(如果兩個來自於不同菌株的 EST 都與某一已知的基因相似，即使這兩個 EST 沒有任何重疊的區域，它們也可能有關。)

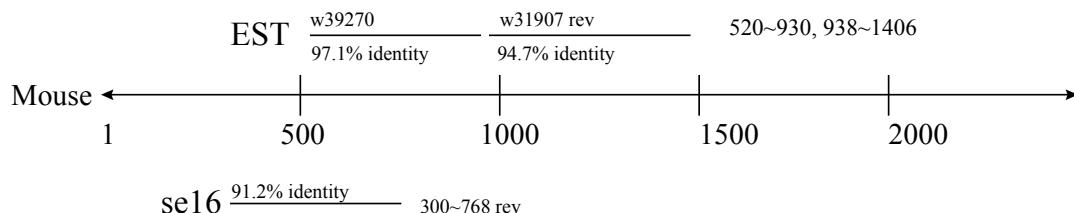
目前 UniGene 已蒐錄了約 121,062 個人類基因的獨特群組(迄至 2002 年 10 月)，其中含有已知基因序列的群組有 22,360。至於 UniGene 獨特群組數為何遠超過目前(2002 年)所公認的人類基因總數呢？其主要原因是因為 EST 序列品質不佳，導致 UniGene 分群困難。

由於 UniGene 資料庫內已包含了這麼多的 EST 群組，如果能找到其中的一群和我們所研究的序列片段是高度相關的，而且在這一群之中又含有不少的 EST 序列，便可以在電腦程式的輔助下組合成較長的接合序列，如此一來，我們就可以在決定 cDNA 序列之前，先得到更多的序列資訊，這些資訊將可以輔助並加速實驗上的定

序。若是所得的 EST 足以組合成一個完整的基因序列，而此基因的功能未知，那麼這種資訊可以指導實驗的進行方向，在最短的時間之內開始分析一個新基因的功能。

利用 UniGene 資料庫(1997 年)時，我們用已知的 EST 號碼 W39270 為關鍵字查詢時，找到分類在同一群之中的另一個 EST，其號碼是 W31907，來自於和 W39270 同一個 cDNA 菌株。UniGene 根據核苷酸序列相似性的比較，發現這兩個 EST 和人類的 eIF4G2 基因序列具有低度相似性，所以它們被 UniGene 歸類為一種人類的 eIF4G2 相關基因。利用 GCG 的 Bestfit 程式，來尋找我們所找到的片段與這兩個 EST 片段在小鼠的 eIF4G2 相關基因上的最佳位置，如圖 12-1 所示，這兩個 EST 在小鼠序列上僅相距 7 個 bp。我們的序列片段(se16 rev)雖使我們得到比 W39270 更向 5' 方向的序列，可是以小鼠序列作為參考時，仍無法得到全長的人類的 cDNA 序列。

圖 12-1 根據序列的相似性圖示 EST W31907 及 W39270，以及 se16 的序列分別對應至小鼠的 eIF4G2 相關基因各個不同的區域，各項位置是以小鼠的 eIF4G2 相關基因由 5' 端至 3' 端的位置為準。



註: rev 代表在進行並列時，該序列必須要轉成互補且方向相反的另一股序列

為輔助篩選 cDNA 基因庫或加速 RACE(Rapid Amplification of cDNA ends)的工作，我們先請中研院白果能老師替我們挑取菌株，經他查閱資料後，告訴我們這一株菌在第 745 盤，可是當時 NHRI 的菌株目前只有 600 多盤，所以我們在網路上向隸屬 Invitrogen 的 Research Genetics 公司(<http://www.resgen.com>)訂這個 cDNA 菌株，大約一週後即收到種在洋菜凝膠上的菌株，可供進一步的序列分析之用。

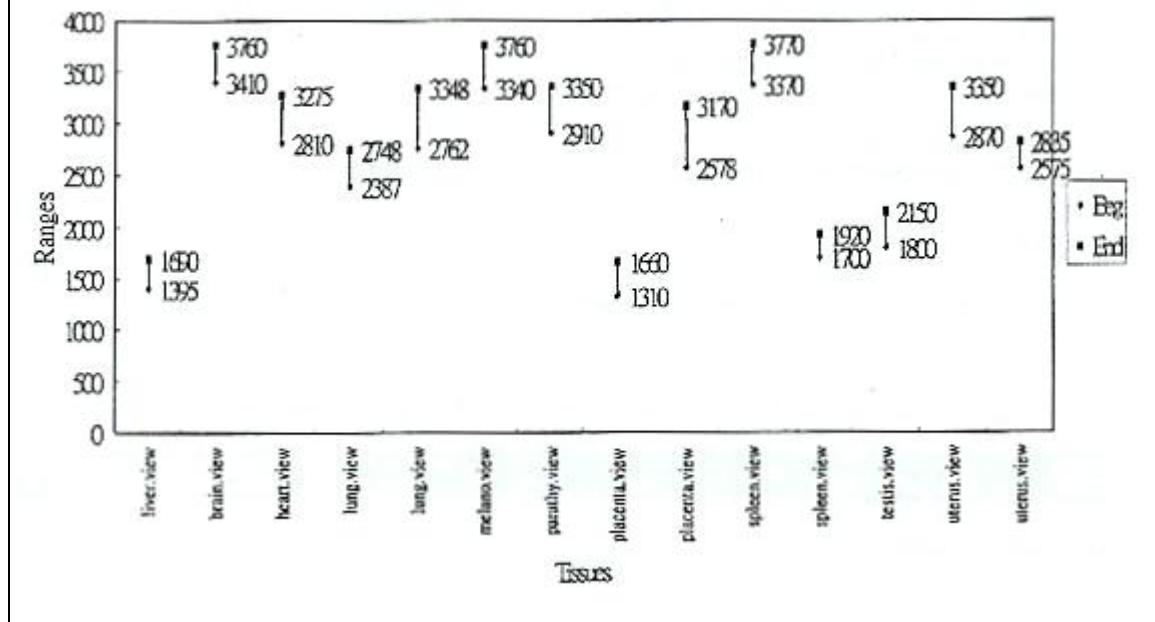
由於此時(1997 年)搜尋 UniGene 資料庫所得到的 EST 很少，並不能幫助我們把全序列組合出來。但是這並不意味著在資料庫之中相關的 EST 序列很少，不值得我們進一步去搜尋序列資料庫。其實 UniGene 本身自動運作的邏輯便會限制它能達成的目的：由於要將所有新的 EST 序列分別歸類至大約四萬五千個群組之中，每天新進超過一千個 EST 序列，要和四萬個群組的序列進行相似性比對，很明顯的會花費很多的運算時間；而且自動化的運作雖然比人工要有效率得多，但是缺乏彈性，很可能會失去一些很有價值的序列資訊。

在搜尋結果未能達成設定的目標之時，我們應該要考慮到每一種工具都有其先天上的限制，這不僅僅是電腦程式的特性，就連所有的實驗方法也具有同樣的特性。為了期望能夠很有效率的運用這些不斷增加的 EST，因此我們思考採取另一種策略。如果改以單一的序列片段直接搜尋相似的 EST，不但可以很有效率，而且可以很合理的推想：隨著 EST 資料庫快速的增大，我們可以陸續的找到新的 EST，只要 EST 數量超過某種臨界點，便可以趕在 UniGene 之前便預測出基因全序列。

因此我們採用 GCG 中的 Blast 查詢國外最新的 EST 序列資料庫，想要找到更多和小鼠 eIF4G2 相關基因具有高度相似性的 EST，可以(1997 年初)找到 96 個 EST 序列，這要比 UniGene 歸在同一群組的 EST 數目要多出很多。由於這些搜尋的結果之中，有不少的 EST 序列是 NCBI 剛收錄不久的，在中研院或是國家高速電腦中心的序列資料庫都找不到這些較新的序列，因此必須要用 NCBI 的 Batch Entrez 序列查詢系統，才能將一些較新的 EST 序列由 NCBI 的序列資料庫取回，以進行進一步的分析。

將 96 個 EST 根據不同組織來源分類之後，利用 GCG 中的 GelMerge 程式可以將各個組織的 EST 序列分別組合成較長的接合序列(contig)。將所得到的較長的接合序列拿來與小鼠的 eIF4G2 相關基因進行序列排比，經過整理並計算相對位置之後，可以得到如圖 12-2 的位置關係圖，縱軸是相對於小鼠的 eIF4G2 相關基因的序列位置，橫軸是標明組成的接合序列是來自各種不同組織。經過分析之後，可以發現這些 EST 只能包含小鼠基因由 1310 至 3770 bp 之間的序列，不僅缺乏最難以獲得序列的 5' 端，而且如果進一步去組合這些接合序列，會發現其中還少了一些序列片段，無法將各片段完全連接起來，如果沒有小鼠基因序列的輔助，無法只由這些接合序列直接組合成更長、更完整的序列。

圖 12-2 用 GCG 中的 GelMerge 程式來重組 96 個與小鼠 eIF4G2 相關基因間具高相似性的 EST，圖示各接合序列和小鼠 eIF4G2 相關基因之間的序列相似的位置。



3. UniGene 資料庫以驚人的速度不斷更新

在前述時間大約二個月後，再以與上述相同的方式查詢 UniGene，發現被歸類這個群組的 EST 數目由原來的 2 個，遽增為 141 個(圖 12-3)。又過了大約三個星期之後，重新再搜尋 UniGene，同樣歸類於這群組的 EST 序列數目減少為 128 個，而且這個基因群組的說明由原來的"ESTs, highly similar to....."(圖 12-3)改成了"Human p97 mRNA, complete cds."(圖 12-4)。很明顯的，在這段短短的二個月之內，UniGene 不斷的將新進入資料庫的序列加以分類，而在二週之後，又在這個原來未能組成全序列

的 EST 群組之中，加入了可和大多數 EST 序列相容的 p97 基因的全序列，同時將一些差異性較大的 EST 序列由這個群組之中剔除。根據序列排比結果，在肝癌組織之中具有差異表現的序列與 p97 基因互相比較，可以在 499 bp 的範圍內有 91.6% 的相同性。

圖 12-3 第二次以 W39270 為關鍵字搜尋 UniGene 資料庫的結果

ESTs, Highly similar to EUKARYOTIC INITIATION FACTOR 4 GAMMA [Homo sapiens]

FUNCTIONAL INFORMATION

est protein match: Swiss-Prot Q04637 (p=4.4e-151)
Protein name: EUKARYOTIC INITIATION FACTOR 4 GAMMA [Homo sapiens]

...

EXPRESSION INFORMATION

Observed in libraries prepared from: fetal heart, placenta, blood--white cells, thyroid gland, testis, multiple sclerosis lesions, fetal lung, parathyroid gland (tumor), uterus, pancreatic islets

SEQUENCES (141)

圖 12-4 第三次以 W39270 為關鍵字搜尋 UniGene 資料庫的結果(約在圖 12-3 的三星期之後)

Human p97 mRNA, complete cds

FUNCTIONAL INFORMATION

est protein match: Swiss-Prot Q04637 (p=4.4e-151)
Protein name: EUKARYOTIC INITIATION FACTOR 4 GAMMA [Homo sapiens]

...

EXPRESSION INFORMATION

Observed in libraries prepared from: melanocyte, multiple sclerosis lesions, fetal brain, fetal lung, fetal heart, fetal cochlea, placenta, colon, thyroid gland, testis

SEQUENCES (128)...

4. 自行以 Blast 程式找到相關的 EST 序列，可以組合成基因全序列

大約與 p97 基因全序列被納入 UniGene 的這個 EST 群組同一時間左右，用 GCG 中的 Blast 程式可以搜尋到 296 個和小鼠 eIF4G2 相關基因具高度相似性的 EST 序列。以 GelMerge 重組這些 EST 序列，並以序列並列程式整理、並計算出所組合成的較長的接合序列和 p97 基因序列的相對位置關係，如圖 12-5 所示，縱軸是 p97 基因序列由 5' 端至 3' 端的位置，橫軸則是由 GelMerge 程式重組 296 個 EST 所產生出來的 23 個較長的片段序列。仔細觀察圖 12-5 後可以發現，只要將所組成的接合序列連接起來，應該可以產生出如 p97 基因的全序列。

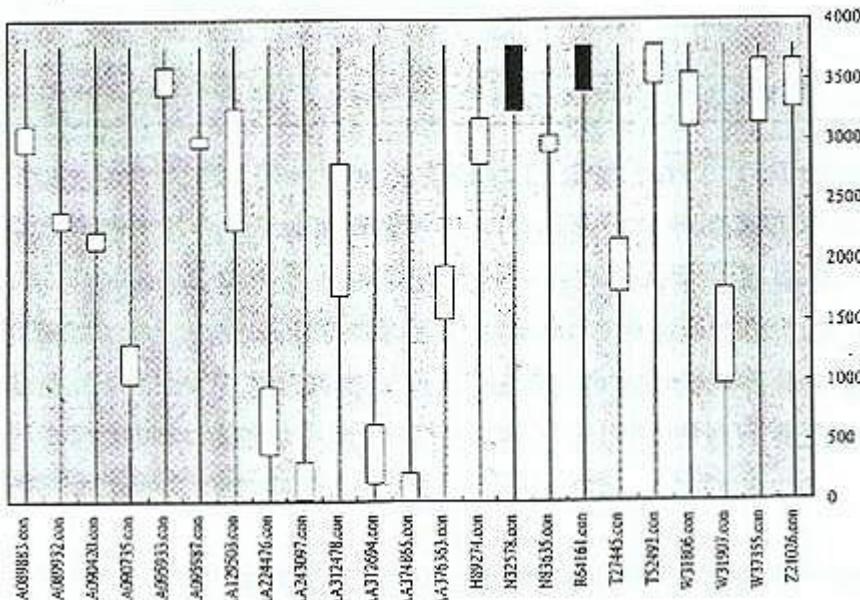
這意味著，如果人類的 p97 基因的全序列尚未被專門研究轉譯因子的實驗室所發表，我們已可藉著電腦程式的輔助，由 EST 的序列自行組合成為相當完整的基因序列。在短短的兩個月內，以同樣方式(Blast 程式)可以搜尋到的相似的 EST 序列數由

96 個增加到 296 個，由原來只能鬆散對應到部分基因序列的狀況(圖 12-2)變成可以包含整個基因全序列。以研究 eIF4G2 的實驗室為例，必須花費較長時間才得相同的結果。

在此處的結果卻令人產生了疑問，為什麼 GelMerge 會無法將這些 EST 直接組合成單一的基因全序列，而是將它們組成了數十個較長的接合序列呢？這裡有數種可能性，基本上都牽涉到 EST 來源的特性：如選殖時發生了錯誤的序列嵌合體(chimera)、或是由自動定序儀所讀出的序列品質不良，所得到的 EST 序列都可能和原來的基因有很大的差異。此外，由於 EST 是用 random priming 的方法由 cDNA 基因庫(library)中所決定的部份 cDNA 序列，因此在 EST 序列的 5' 端常常會多出一段和原來基因的序列差異較大的序列，這種序列的存在同樣也會干擾電腦程式進行序列組合。

但是除了上述原因之外，其實還有一種非常自然的可能性：如果這種基因具有不只一種的基因剪接型式(splicing form)，那麼很自然就會由實驗上得到一些彼此的序列不太相容的 EST，也因此在組合序列之時造成了電腦程式處理時的困擾，無法直接將衍生於各種不同的基因剪接型式的各個 EST 組合成為單一的接合序列，不得不加以分成數個群組。

圖 12-5 將 296 個由 Blast 程式所搜尋到的 EST 序列，以 GelMerge 進行重組及分類的結果，圖示表示各接合序列和 p97 基因全序列的相關位置。

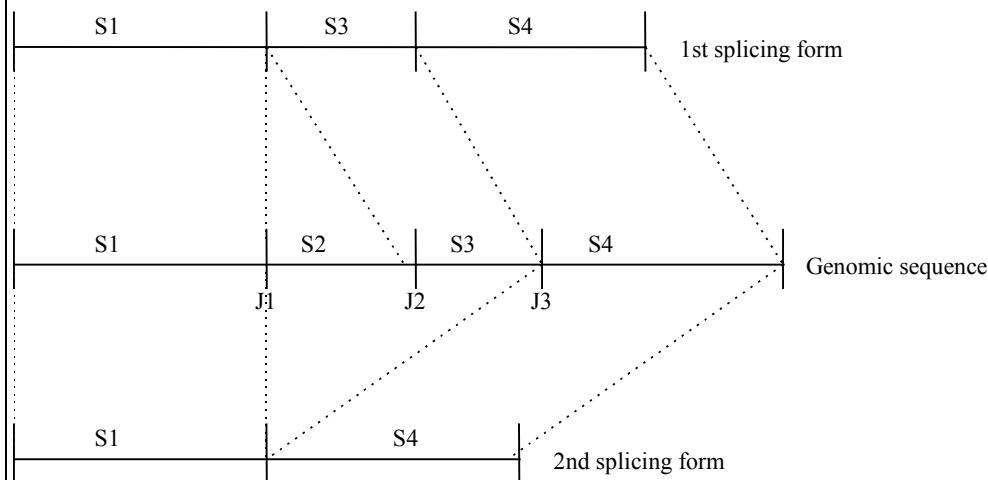


5. 由接合序列與基因全序列的序列排比來預測基因剪接點的位置

如圖 12-6 所示，假設有一基因具有三個基因剪接點：J1、J2，以及 J3，將 genomic sequence 分成 S1、S2、S3，及 S4 四個區段。由於選擇不同的基因剪接點，因而造成二種不同的基因剪接型式：1st 和 2nd。若是有一接合序列 X 源自於 1st splicing form 而且含有 S1、S3、以及 S4，另一接合序列 Y 源自於 2nd splicing form 而且含有 S1 及 S4，則比較 X 與 Y 之間的差別，便可以發現 X 比 Y 多了 S3，同時可以將序列分成 S1、S2、S3，以及 S4 四段區域，而且可以更進一步推估出三個基因剪

接點：J1、J2、以及 J3 的位置。雖然在理論上預測基因剪接點並不難，但實作之時並不是如此直接了當。

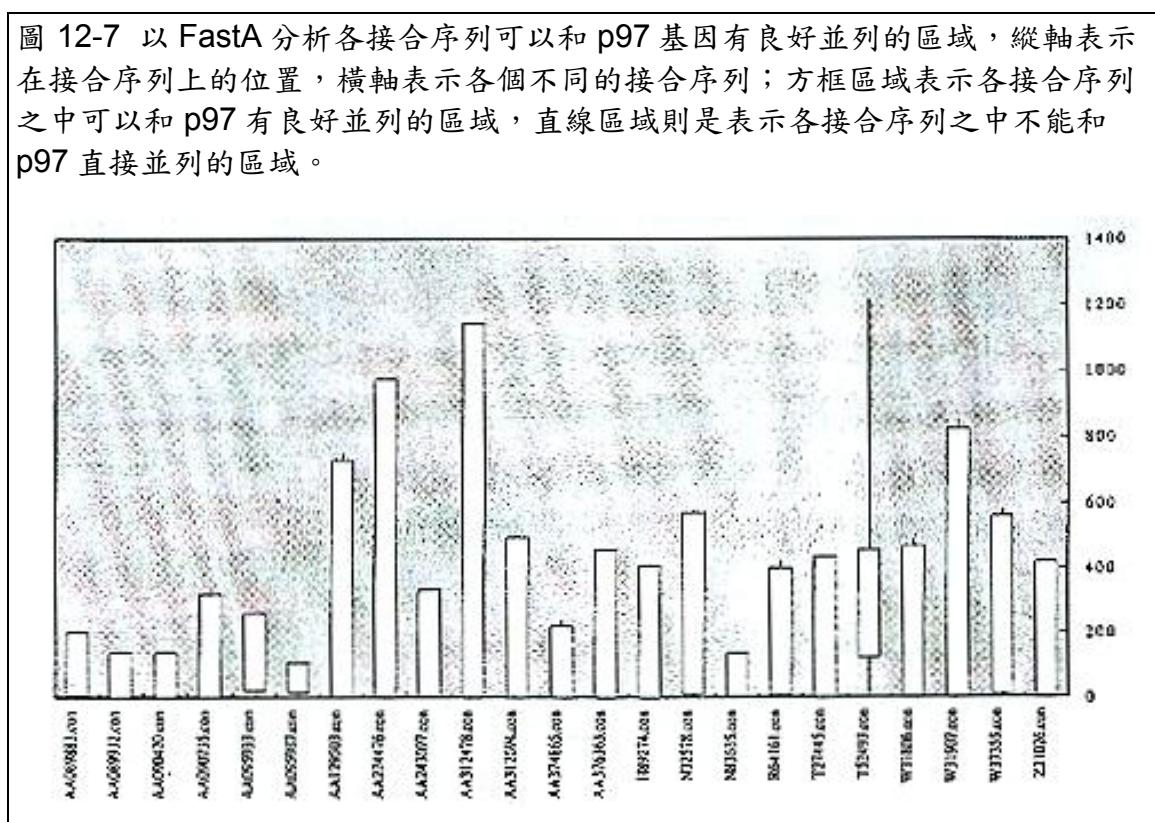
圖 12-6 假設有一基因具有三個基因剪接點: J1、J2，以及 J3，將 genomic sequence 分成 S1、S2、S3，及 S4 四個區段。由於選擇不同的基因剪接點，因而造成二種不同的基因剪接型式: 1st 和 2nd。



由 296 個 EST 可以組合成 23 組較長的接合序列，再用 FastA 針對接合序列進行序列並列位置分析，算出接合序列中實際可以和 p97 基因有良好並列的區域時，可以發現很多接合序列在一端或是兩端會有似乎是多餘的序列(如圖 12-7 中各接合序列的直線部分所示)。大致上而言，幾乎所有的接合序列都會有現象，如果這種多餘序列的長度在 20 bp 以內，很有可能是 random priming 所造成；但如果多餘序列超過 20 bp，額外的序列就可以用來預測基因剪接點的位置。

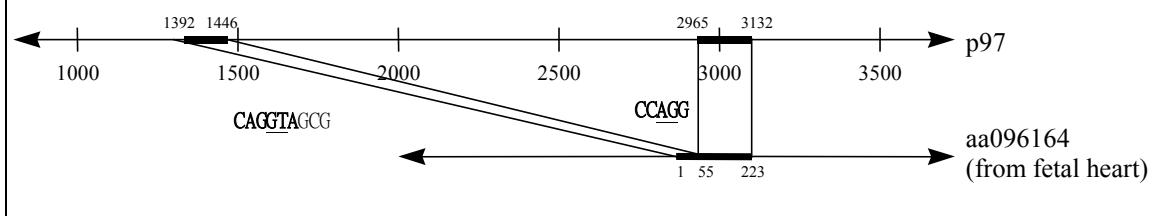
對上述所得到的 23 個接合序列做進一步的序列排比，可以發現有些可能是來自於不同的基因剪接型式，進行分析序列並列之後，其中之一結果經分析後發現，來自於 fetal heart 的 aa096164 這一個接合序列缺乏 p97 上約由 1445 bp 至 2962 bp 的區域，再配合已知的基因剪接點附近的守舊序列(圖 12-8之中以底線標示處)，預測基因剪接點可能位於 p97 的 1446 bp 及 2965 bp 兩個位置上。因此若是這種基因剪接型式存在，則其 mRNA 應比 p97 短 1.5 kb。

圖 12-7 以 FastA 分析各接合序列可以和 p97 基因有良好並列的區域，縱軸表示在接合序列上的位置，橫軸表示各個不同的接合序列；方框區域表示各接合序列之中可以和 p97 有良好並列的區域，直線區域則是表示各接合序列之中不能和 p97 直接並列的區域。



以我們所發現的具差異表現性的序列製成 RNA 探針，對 multiple tissue blot 進行 Northern hybridization，發現在數個組織(包含了心肌、胎盤、及骨骼肌等等)之中，除了分子量為 4.0 kb(符合 p97 mRNA 的大小)的訊號之外，還會有一個分子量約為 2.4 kb 的訊號，兩者的差異大約是 1.6 kb 左右。由這個實驗結果，似乎符合所預測出來另一種基因剪接型式所產生 mRNA 的大小。由實驗結果顯示，這是值得再進一步進行深究的問題。

圖 12-8 由接合序列 aa096164 與 p97 基因的序列排比，所預測出的基因剪接點的位置。圖示 aa096164 與 p97 基因的序列相關位置，並列出在基因剪接點處的序列，以底線標示出守舊的區域。



6. 憲測性基因剪接資料庫的建立

根據最新(2000~2001)的估計，人類約有三萬至四萬個基因，只有線蟲的兩倍，這不禁使人產生疑問，這多出來一倍的基因，可以合理解釋(和簡單的線蟲相較之下)人類所具有複雜的生理現象嗎？不過若利用基因剪接有組合的效果，使同一基因可以產生不盡相同的基因產物，正是一種可以實際數量的有效策略；換言之，單一染色體區

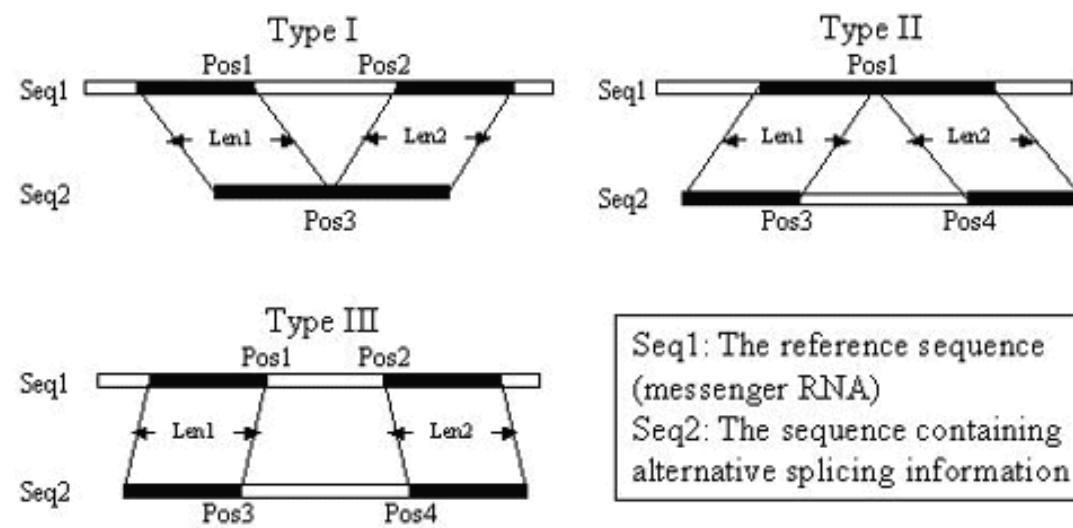
域，可以使用另類基因剪接的方式，在不同的環境及調控下，產生具不同功能的蛋白質。因此，雖然只有三萬多個基因，可以產生的蛋白質數，可以因為另類基因剪接的現象，而產生數倍的變化。依我們的數據，50% 以上的人類已知基因至少有一個另類基因剪接點，這顯示另類基因剪接的現象是在做基因表現研究時不可忽視的因素。

我們並不是第一個做基因剪接點資料庫的實驗室，只是有些網際網路上的基因剪接資訊被保護，無法取用；而少數在網際網路上開放的資料庫，為了強調其正確性，捨棄大量寶貴的資訊，所能提供的資訊大為減少，對於實驗設計的助力不大；此外，相關的資料庫在發表期刊論文後，有部分並沒有繼續加以維護。我們的目的是要建立一個可信、維護良好、經常更新、而且資訊足夠的資料庫做為分析微陣列數據或設計微陣列探針(probe)之用。

PALS db (<http://palsdb.ym.edu.tw/>) 是 Putative ALternative Splicing database 的縮寫，這是一個收錄由 EST 資料與 mRNA 所計算出的可能基因剪接點的資料庫。因為這兩種序列都是來自於實驗，因此正確率相當高。PALS db 的出發點，是認為生物資訊的分析只是提供思考的方向，最終仍需實驗的驗證。因此 PALS 資料庫儘可能地多收集資訊，但對基因剪接點的精確位置並不特別要求，這是因為生物學家必須要先有資訊，才能設計探針驗證想法，而設計探針的工作，只要有基因剪接點的大概位置即可。因此 PALS 資料庫竭力計算最可能的基因剪接位置，可是不會為了其位置可能差一、兩個鹼基就捨棄這筆資料。基於這種設計理念，PALS 資料庫所呈現的資訊量，居各相關資料庫之冠。除此之外，PALS 資料庫更首創以支持給定基因剪接點的 EST 個數來表示臆測的品質。

表現序列標幟 (EST) 序列是目前序列資料庫中，序列數最多的序列，它是部份的 cDNA 序列，有可能跨越另一種基因剪接形式的 cDNA 之剪接點。因此 EST 的序列越多，就有越多的基因剪接資訊。透過序列排比的方法，即可預測基因剪接點的位置 (圖 12-8)。目前(2001-2002 年)基因體序列的初稿品質還不夠好，許多已知的 mRNA 的基因體序列還不夠完整，所以我們捨基因體序列，而採用 mRNA 為排比時的參考序列。這種策略使 PALS db 成為目前收錄最多序列的基因剪接資料庫。除了收錄的基因眾多之外，PALS db 還有提供支持預測的證據的特色。

圖 12-9 預測基因剪接點的原理



理論上來說，EST 是實驗的數據，生物資訊分析只是將實驗的結果用一種更直接的方式呈現出來而已。由這樣的角度看，一個 EST 所支持的基因剪接點就相當可信。不過在選殖 cDNA 的過程中仍有機會引入少許的錯誤，因此若有多個 EST 序列支持同一個基因剪接點就可增加我們對預測的信賴。

提供基因庫來源資訊

EST 序列是來自於不同的 cDNA 基因庫 (library)，因此基因庫的來源表示此 EST 出現於製作基因庫所用的組織或器官，同時也區分正常組織與癌症組織。PALS db 將此資料連接在圖形介面上，只要將滑鼠在相關的排比結果上兩秒鐘，基因庫來源的資訊就會呈現在瀏覽器左下方的訊息區中。當多個 EST 支援同一個基因剪接點時，若不同的 EST 分別出現在代表不同組織的基因庫中，則可信度越高。從另一角度看，此資訊也顯示某一特定的基因剪接型式會表現在哪些組織或器官，這對後續的研究，將有很大的幫助。例如它不但可指引我們到哪一種組織中去選殖特有的 mRNA，亦可提供在不同組織中是否會表現不同基因剪接產物的部份資訊。此外某些基因剪接點大量出現在癌症組織中，在正常組織中卻非常少見，也有基因剪接點只出現在正常組織中。這些資訊不但可以用來探究癌症致病的因子，也能做為分子檢驗的參考。

讓使用者看到所有的相關資訊

在預測的過程中要將數萬個基因與其所有相關的 EST 排比，所以不可能針對個別的排比做最佳化的分析。若使用固定的篩選標準，濾掉某些資訊，則有可能丟掉部份有用的資訊。因此 PALS db 在設計時，採用標示符合選取標準的基因剪接點的方式，讓讀者可看到所有的資訊，而且可選用不同的標準標示可信度高的基因剪接點。有時支持某一基因剪接點的 EST 序列可能只有一個通過設定的標準，卻有其他的 EST 序列排比效果稍差，使用者若能看到，也可增加對預測的信賴程度。

以資訊驅動生物醫學研究

PALS db 在發展的第一階段(2000~2002 年)在發展方法，尚未將相關的訊息整合到資料庫中。因此提供許多相關的超連結，讓使用者可以將序列送到遠方的網站做分析，並將分析的結果和基因剪接資料比對，以發現新的現象。例如某些蛋白質上的訊號(signal)或模組(module)可能出現在一種 mRNA，卻不出現在不同基因剪接形式的 mRNA 中，這兩種 mRNA 所產生的蛋白質可能因此具有不同的功能。再如 SNP 若座落在基因剪接點上，或鄰近的位置，可能引起基因剪接形式的變化，而使個體的體質產生差異等。利用比對資訊的方式去發現新的現象，因而建立假說的研究方式，我們將其稱為「以資訊驅動的生物醫學研究」的模式。

四、 結語

由 eIF4G2 的例子中可以看出，在未來做研究的步調會非常的快，若不能習慣於使用網路的資訊與國際上的資源，將遠遠落在別人之後。在國外，已有一些機構，要求博士研究生要至生物資訊實驗室接受訓練。這充分顯示了未來生物醫學研究的特性，即是「以資訊驅動研究」的特性。對生物醫學研究而言，由實驗桌上所得到的實驗數據，基本上就是一種資訊；確定的實驗結果會更一步成為可信的資訊來源，如 GenBank 就是累積了大量可信的序列資訊的資料庫。因此，沒有必要刻意的去對實驗

數據和來自資料庫的資訊做出清楚的界限。資料探採(data mining)，即由大量資訊之中尋找知識，其實就是將平日小規模的實驗室數據分析，擴大為數百、或數千倍。

當實驗數據累積至人力不易分析的程度，如 EST 序列分析、微陣列數據，就必須要運用適當的資訊工具，將結果和現有資訊及知識作高度整合，以獲得最完整的資料，做出最佳的判斷。因此，「有效率的處理資料、整合為可用的資訊、再進一步理出資訊的脈絡、找出隱藏其中的規則」，我們相信，將是未來研究生命科學必備的核心能力之一。

參考網站

1. <http://www.ncbi.nih.gov/UniGene>
2. <http://www.tigr.org/tdb>
3. <http://palsdb.ym.edu.tw/>

附錄一 PuTTY、X Window 軟體的取得與安裝

PuTTY 程式介紹

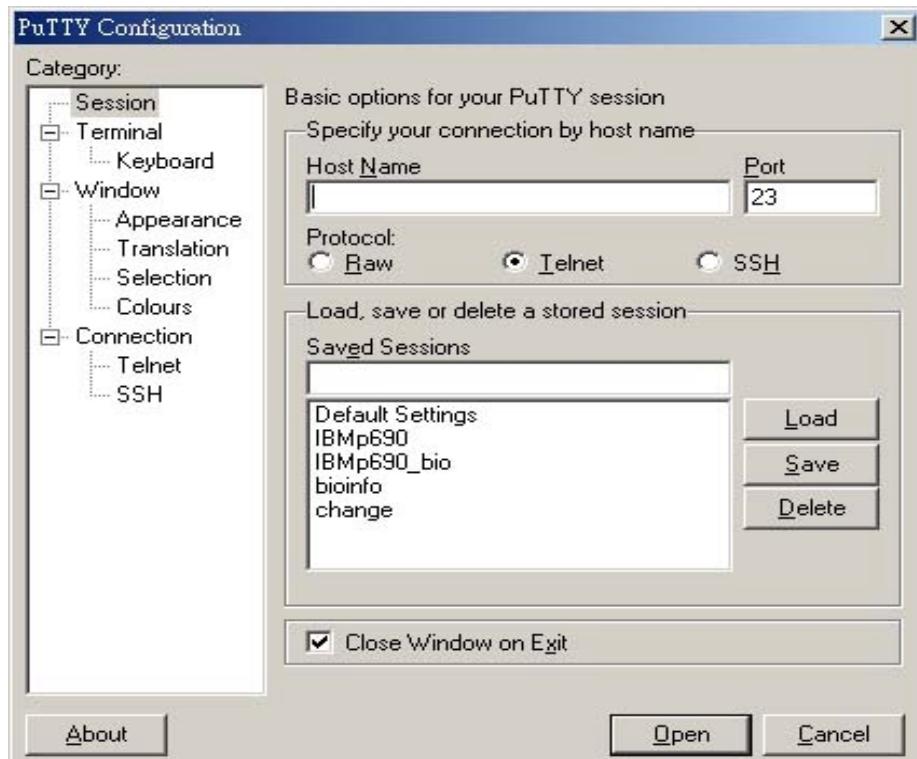
PuTTY 是一套免費提供給 32-bit Windows 系統使用的客戶端簽入軟體，其中利用到 SSH 及 Telnet 通訊協定，這兩個通訊協定都是提供使用者能夠經由網路簽入一台多使用者(multi-user)作業系統的主機。

Telnet 是一個很方便的通訊協定，在 Windows 所有版本也都有提供這樣的客戶端連線服務，但是由於 telnet 採用「明碼」作為傳輸的方式，所以在傳輸的過程中，假設有人在監聽你的網路封包，而你在此時輸入你的帳號和密碼，如此一來，你的帳號和密碼就很容易的被對方解讀，這樣對於系統的安全性會造成很大的威脅。有鑑於此，PuTTY 也提供了另一個連線的協定—SSH，他是 Secure SHell protocol 的簡寫，他可以經由將連線封包加密的技術，來進行資料的傳遞，所以相對的他自然比較安全囉！

由於使用 Linux 作業系統的使用者，已經有 ssh 可以使用，而在 Windows 底下便可以使用 PuTTY 來進行連結，取得的方式可以參考下面的網站：<http://www.chiark.greenend.org.uk/~sgtatham/putty/>。

PuTTY 可直接在 Windows 下執行，執行的介面如圖 A1-1 所示。

圖 A1-1 PuTTY 的執行介面



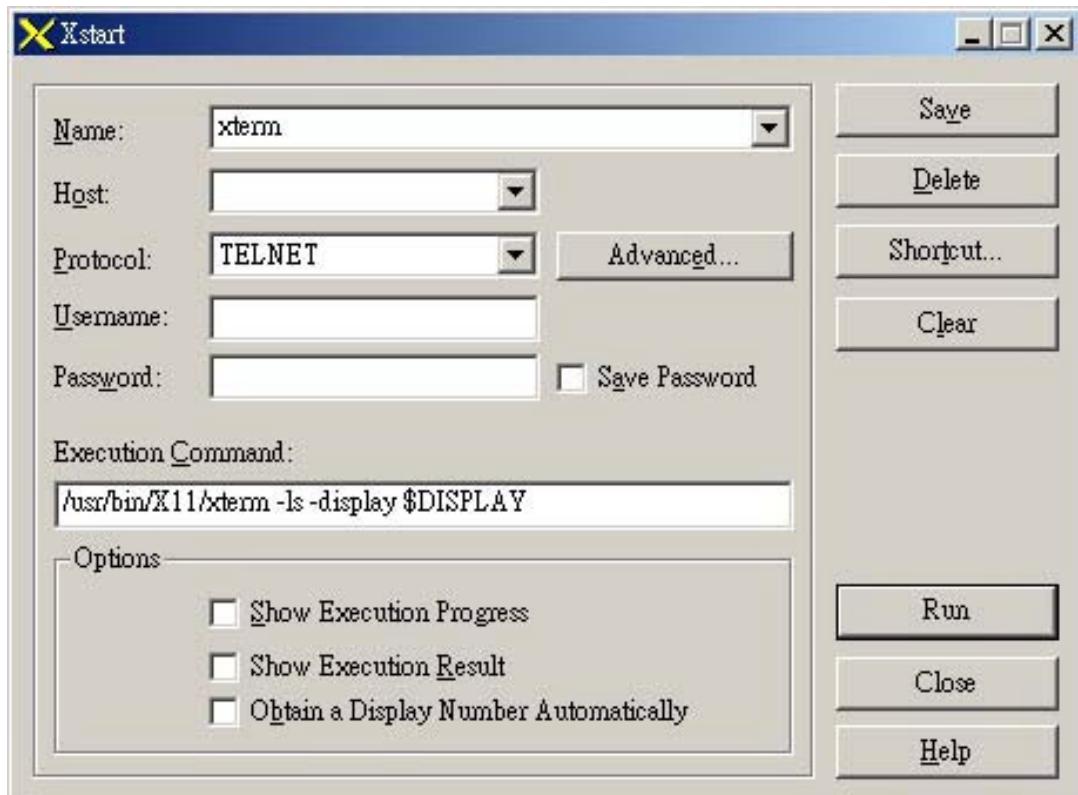
執行出如圖 A1-1 的介面後，便可在 Host Name 的欄位輸入主機名稱或是 IP 位置，並且選擇 Telnet 或 SSH 協定，按下 Open 按鈕便可與主機進行連結工作。

X Window 介紹

X Window 是 UNIX 系統下的一個視窗環境，只要在遠方電腦上的程式支援 X Window，連線到其上的電腦即可透過客戶與伺服器(client-server)的架構，來顯示遠方電腦的螢幕。它在自己的電腦上建立一個 X Server，然後在遠方電腦(客戶端)上利用「xterm -display」指令尋找這一台 X Server。一旦 client 找到 X Server，即可在使用者的電腦螢幕(X Client)上建立一個 X Window，顯示遠方電腦(客戶端)的螢幕。在此，視窗中的畫面，與你用 telnet 方式模擬圖形終端機無異。可是 X Window 可充分運用 UNIX 環境下的多工能力，開啟多個視窗，執行不同的程式。其實在一個視窗內，將工作依序丟到背景中執行也可以，只是在看結果時無法並列比較而已。

而現今有許多在 Win32 架構下的 X Window 模擬器，較著名的有 Xmanager 與 X-Win32，在此介紹 Xmanager 的取得與安裝。Xmanager 可由 <http://www.netsarang.com/> 網站下載，下載後即可執行安裝。安裝完成後執行 Xstart.exe 程式，可得以下視窗介面(圖 A1-2)：

圖 A1-2 Xstart 視窗介面



附錄一 PuTTY、X Window 軟體的取得與安裝

欲建立 X window 連結時，需指定主機名稱(Host)、通訊協定(Protocol)及使用者名稱與密碼，其中的通訊協定包含有 telnet 與 SSH 等，另外就是圖 A1-2 中的 Execution Command 欄位目的是讓 xterm 能夠傳回客戶端螢幕中顯示。設定完畢後，按下 Run 按鈕，即可進行 X Window 連線。

參考網站

1. <http://www.chiark.greenend.org.uk/~sgtatham/putty/>
2. <http://www.netsarang.com/>

附錄二 習題組：定位訊號 RNA 接合蛋白的選殖與分析

本書之目的是希望教各位學員如何使用電腦的工具來分析巨分子的序列。這是一個技術層次的課程，所以做練習是一件非常重要的事。為了讓大家很清楚地知道在什麼場合該用什麼程式，習題的安排是以一個研究計劃做主幹，來說明在做研究時的各種需求。在進入細節之前必須先知道這計劃是如何進行的：在南非水生有爪蛙(*Xenopus*)卵母細胞的發育過程中，細胞中有一種 Vgl mRNA 會集中到植物極⁽¹⁾，而且其蛋白質合成也集中在植物極⁽²⁾。這 mRNA 一旦移動到植物極就會固著在植物極，直到卵母細胞經減數分裂形成卵；卵與精子結合形成受精卵。Vgl mRNA 及其蛋白質產物都始終集中在這區域。受精卵經數次有絲分裂之後，在植物極的 Blastomere 就會繼續合成 Vgl 蛋白質，而另一些沒有 Vgl mRNA 的 Blastomere 則分化出其他的功能。這些能合成並分泌 Vgl 的細胞，在胚胎發育中能誘使鄰近的細胞形成中胚層。因此 mRNA 定位

The Specificity of Translational Control Switched with Transfer RNA Identity Rules

M. GRAFFE, J. DONDON, J. CAILLET, P. ROMBY, C. EHRESMANN,
B. EHRESMANN, M. SPRINGER*

The interaction of *Escherichia coli* threonyl-transfer RNA (tRNA) synthetase with the leader sequence of its own messenger RNA inhibits ribosome binding, resulting in negative translational feedback regulation. The leader sequence resembles the substrate (tRNA^{Thr}) of the enzyme, and the nucleotides that mediate the correct recognition of the leader and the tRNA may be the same. A mutation suggested by tRNA identity rules that switches the resemblance of the leader sequence from tRNA^{Thr} to tRNA^{Met} causes the translation of the threonyl-tRNA synthetase messenger RNA to become regulated by methionyl-tRNA synthetase. This identity swap in the leader messenger RNA indicates that tRNA identity rules may be extended to interactions of synthetases with other RNAs.

Construct	Vgl 3'UTR	Localization
XBG-Vg3'UTR	—	+
XBG-632	—	+
XBG-635	—	-
XBG-718	—	+
XBG-188	—	-
XBG-419	—	-
XBG-366	—	+
XBG-3105'	—	+
XBG-3305'	—	-/(+)
XBG-3405'	—	+
XBG-3045'	—	-/(+)
XBG-2515'	—	+
XBG-340A	—	-
XBG-340A+	—	-
		1108-2378
		1108-1740
		1743-2378
		1441-2159
		1108-1294
		1108-1527
		1441-1809
		1499-1809
		1479-1809
		1441-1783
		1441-1747
		1441-1692
		1441-1531, 1683-1783
		1441-1531, XBG gene, 1683-1783

In situ hybridizations to detect localized RNAs. Oocyte sections are shown with the vegetal pole at the bottom, and the hybridization signal appears as white grains. (A) Oocyte injected with RNA transcribed from the XBG construct and hybridized (4) with an XBG probe. Hybridization is uniform throughout the cytoplasm and the germinal vesicle is evident as a non-hybridizing "hole" at the center. (B) Oocyte injected with RNA transcribed from the chimeric construct (XBG-366) and hybridized with an XBG probe. Hybridization to the injected RNA is localized to the vegetal hemisphere. (C) Oocyte hybridized with a Vgl probe to detect endogenous Vgl RNA. The hybridization signal is restricted to the vegetal hemisphere.

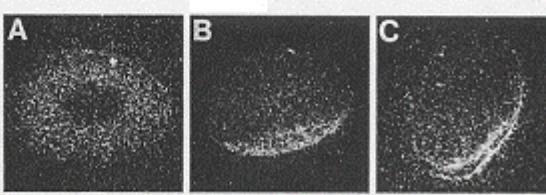


圖 A2-1 定位訊號 RNA 的決定。

(localization)的現象是細胞分化的一種重要機制。在 1992 年，Melton 的實驗室將 Vgl mRNA 的 3'-端未轉譯區中的不同片段接到 globin 基因之後，然後觀察那一種 chimeric RNA 能位移並固著到植物極⁽³⁾。結果發現有一段 340 個核苷酸的片段是產生定位現象的充分必要條件(圖 A2-1)。

為了進一步研究 Vgl mRNA 定位的機制，首先要找出那些蛋白質能與 Vgl 定位訊號 RNA 產生具專一性的結合。由 Melton 的實驗中，我們可推論 Vgl 的定位訊號 RNA 會獨立摺疊而不受 coding sequence 的影響。可是在分離定位訊號 RNA 接合蛋白時，為了避免找到 coding sequence 的接合蛋白；最好是以僅含 340 nt 的定位訊號 RNA 為探針。為確定定位訊號 RNA 是否也能獨立摺疊，我們將以電腦預測 RNA 的二級結構以供參考。此外也希望比較 Vgl 與其他有 mRNA 定位現象的 RNA (例如 Xcat2)的定位訊號，以觀察不同的定位訊號 RNA 是否具有相似的二級結構。在確定定位訊號 RNA 可獨立摺疊後，利用 band-shift assay 協助純化定位訊號 RNA 的接合蛋白。首先先向 Melton 取得含 Vgl mRNA 的質體，在確定自 Melton 處所得質體的圖譜正確後，subclone 出定位訊號 RNA 的 340 nt 的片段。在試管中合成定位訊號的 RNA 做為探針，利用 band-shift assay 來追蹤在定位訊號 RNA 的接合蛋白質在分離過程中的去向。經蛋白質序列分析，得到部份序列；一方面以此序列搜尋 Genbank，以瞭解這接合蛋白是否與已發現的基因相似。在另一方面，也將此 peptide 之序列，反轉錄為 DNA 序列，以便用它在 cDNA 基因庫(library)中選殖定位訊號 RNA 的接合蛋白。結果釣出數個 phage clones，經 DNA 序列分析，須用電腦程式將片段之序列組合起來。一旦組合出全長的 cDNA 後，就可找到蛋白質的位置。經過資料庫的比對，我們將可看出有那些蛋白質與所研究的蛋白質相關，這群蛋白質在並列比較時是否有守舊的區域，值得進一步探究其功能？所研究的蛋白質是否具有已知的特徵？例如 RNA 接合區等。此外這個蛋白質的性質也是研究的重點。以下就是本課程的習題：

Problem 1.為利用 SP6 promoter 轉錄 340nt 的定位訊號 RNA，你將用下列兩個引子夾出定位訊號 RNA 基因：

1436		
	XhoI	Vg1XhoI: gcc <u>ct</u> <u>cga</u> GC AATAT TTCTA CTTTA TTTC
	1812	
	XbaI	Vg1XbaI: gc <u>tc</u> <u>tag</u> AA TGCTC AAGTC ATATG GAC

再將其剪接到 pGEM7Z(+) (圖 A-2) 的 Xhol 與 XbaI 位置上，並命名為 pLS340。請在電腦上將 pLS340 的序列組合出來。

Problem 2.請在 pLS340 的序列上標示出能辨識 6 個核苷酸的限制酵素(6-cutter)之切割位置。

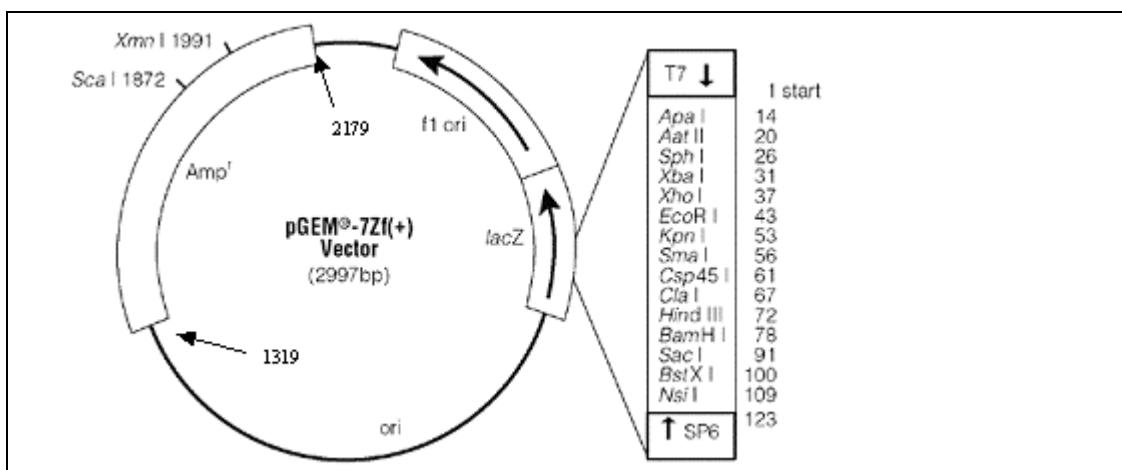


圖 A2-2 pGEM7zf(+)的圖譜

Problem 3.繪出第 2 題之 pLS340 的圓形圖譜，只要標注質體名稱、切一次的酵素切割位置 Amp^r 與定位訊號 RNA 基因的位置即可。

Problem 4.以 pLS340 為模板轉錄出的定位訊號 RNA，可在試管中與 *Xenopus* 的卵母細胞抽出物作用。假設以 band-shift assay 協助蛋白質之純化後，可找到 apparent M.W.，分別為 56、54、42、40kd 等四個蛋白質可與定位訊號 RNA 產生專一性的結合，其中 p56 的部分序列已被決定出來，兩段 peptide 的序列存於「140.129.66.15 或 strive.ym.edu.tw」之「/pub/binfo/problems」目錄下，序列名稱為 p56-1.pep 與 p56-2.pep。要如何確定所找到的蛋白質不是純化時所引入之 artifact ？

Problem 5.請由所找到的 peptide 序列，設計一個選殖 p56 cDNA 的方法。

Problem 6.請找出在組合的 cDNA 序列中，哪一段能轉譯出蛋白質？

Problem 7.在資料庫中有哪些核酸序列與這 cDNA 相似？

Problem 8.在資料庫中有哪些蛋白質序列與這 cDNA 所轉譯出的蛋白質相似？

Problem 9.在問題 9 與 10 中所找到的序列是否相似？如果不相似，為什麼？

Problem 10.請由資料庫中所找出與 p56 相似的蛋白質中選出相似性較高的序列做多序列並列分析。請以“*”表示相同的序列，並寫出共有序列。

Problem 11.試分析這幾個序列在演化上可能的親緣關係。請問它們可分為不同的 subtype 或 subfamily 嗎？

Problem 12.由這 cDNA 所轉譯出的蛋白質上，是否有已知功能的模組？如果有，是位於這個蛋白質的哪一部份？

Problem 13. 請預測這個蛋白質的二級結構，其中是否有兩性的 α -螺旋？如果有，請繪出 helicalwheel，並在序列上標幟其位置。

解答: <http://binfo.ym.edu.tw/bi/books/answers.htm>

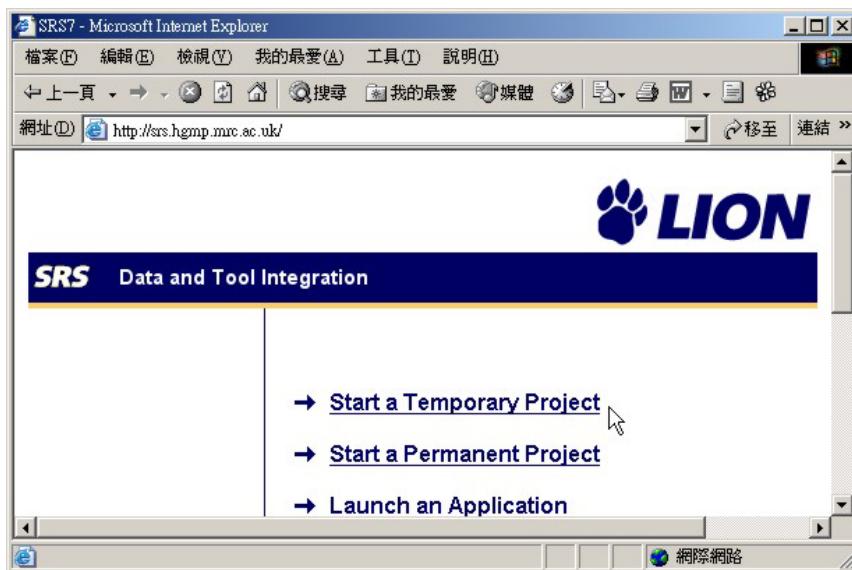
參考文獻

1. Weeks D. L., Melton D. A. (1987). Cell. 51(5):861-7.
2. Tannahill D., Melton D. A. (1989). Development. 106(4):775-85.
3. Mowry K. L., Melton D. A. (1992). Science. 255(5047):991-4.

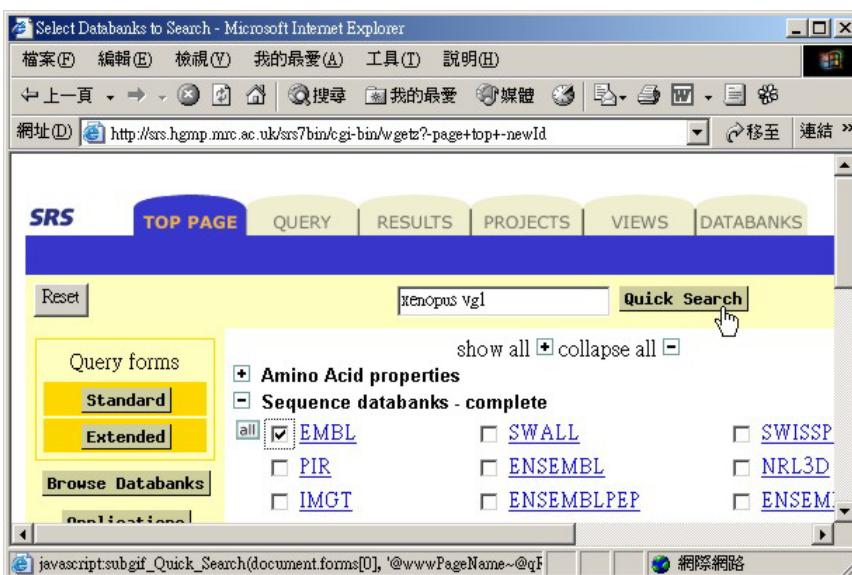
附錄三 習題參考解答

Problem 1.

到 SRS 網站把 Xenopus Vg1 基因和 pGEM7Z Vector 兩條 DNA 序列找出來。SRS 網址為 <http://srs.hgmp.mrc.ac.uk>，請點選 Start a Temporary Project。



點選 EMBL 資料庫，輸入 xenopus vg1 做查詢，按下 Quick Search。



附錄三 習題參考解答

點選 EMBL:XLGFTB(Accession Number 為 M18055)。

<input type="checkbox"/> EMBL:AY028920	AY028920	Xenopus laevis proline-rich Vg1 mRNA-binding protein mRNA, complete cds.	1891
<input type="checkbox"/> EMBL:GGU73003	U73003	Gallus gallus cVg1 mRNA, complete cds.	1283
<input type="checkbox"/> EMBL:XLBMP7	X63427	X.laevis mRNA for bone morphogenetic protein 7 (Xbr41)	1519
<input checked="" type="checkbox"/> EMBL:XLGFTB	M18055	X.laevis transforming growth factor-beta (Vg1) gene, complete cds.	2448

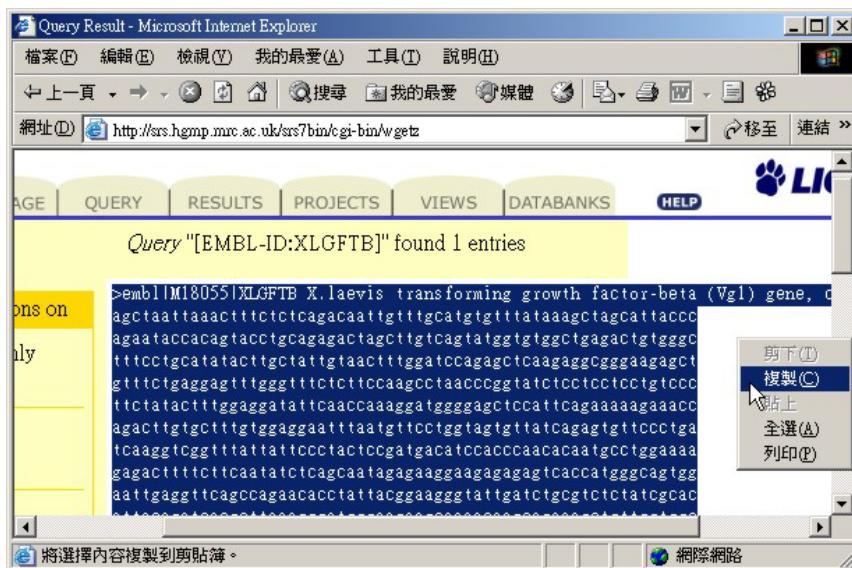
SRS 7.0.1 | [feedback](#)

視窗左邊選”FastaSeqs”然後按”View”。

* selected only		AW147488	XENOPUS_SOURCE
<input type="checkbox"/> Link			5' similar to TR:Q9YI39 PROTEIN ; mRNA se
<input type="checkbox"/> Save			db29e01.y1 Xenopus
<input type="checkbox"/> View			Xenopus laevis cDNA XENOPUS_SOURCE TR:Q9YI39 Q9YI39 mRNA sequence.
<input type="checkbox"/> FastaSeqs			db68g05.y1 Wellcome
<input type="checkbox"/> Sort Results By			Xenopus laevis cDNA IMAGE:3378200 5' si
<input type="checkbox"/> unsorted			gb:gblAF064633.1 AF
<input type="checkbox"/> ascending			laevis Vg1 RNA bindi
<input type="checkbox"/> descending			

將此序列整個複製起來，貼到 PC 的記事本中。

附錄三 習題參考解答

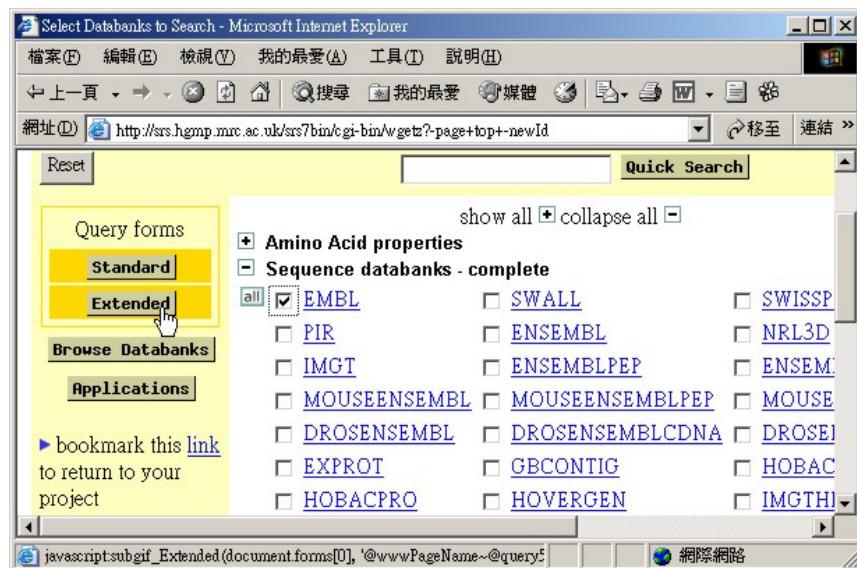


另存新檔成為 vg1.txt。

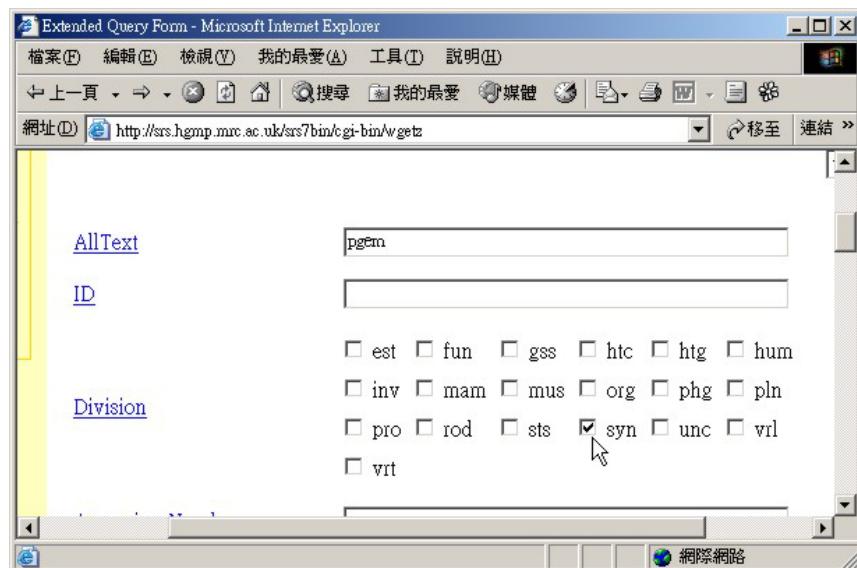


接著找 pGEM7Z Vector 的序列，回到 SRS 網頁點選 EMBL，左邊點選 Extended。

附錄三 習題參考解答



AllText 欄位鍵入 pgem，Division 點選 syn，只要找這一個部分的序列，不然會有太多的序列。



找到 EMBL:CVGEM7ZZFP(Accession number 為 X65310)。

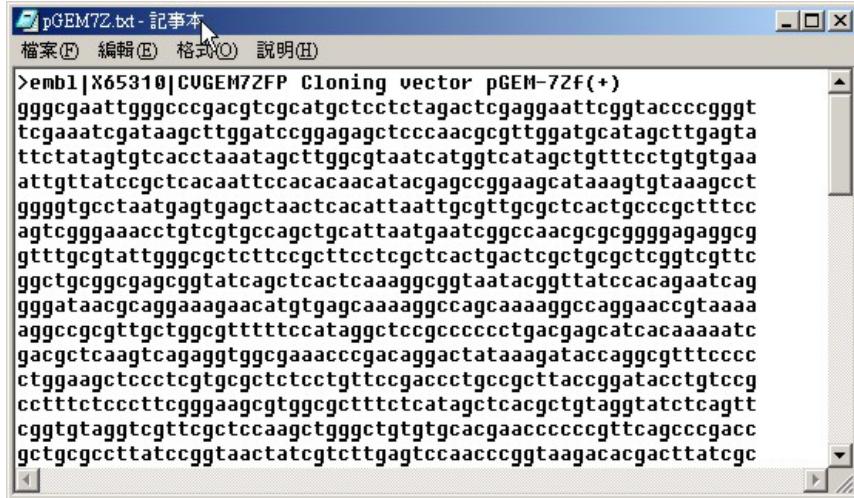
附錄三 習題參考解答

EMBL:CVGEM7ZFP				
	X65309	pGEM-5Zf(-)	3001	
<input type="checkbox"/>	EMBL:CVGEM5ZFP	X65308	Cloning vector pGEM-5Zf(+)	3000
<input type="checkbox"/>	EMBL:CVGEM7ZFM	X65311	Cloning vector pGEM-7Zf(-)	2998
<input checked="" type="checkbox"/>	EMBL:CVGEM7ZFP	X65310	Cloning vector pGEM-7Zf(+)	2997
<input type="checkbox"/>	EMBL:CVGEM9ZFM	X65312	Cloning vector pGEM-9Zf(-)	2912
<input type="checkbox"/>	EMBL:CVPFL59M		single stranded replicative	

複製整個序列貼到記事本存成文字檔。

取名字叫 pGEM7Z.txt，請注意現在 vg1.txt 和 pGEM7Z.txt 兩條 FastA 格式的序列在您的個人電腦。

附錄三 習題參考解答



The screenshot shows a Windows Notepad window with the title 'pGEM7Z.txt - 記事本'. The menu bar includes '檔案(F)', '編輯(E)', '格式(O)', and '說明(H)'. The main content area displays an EMBL sequence file for 'Cloning vector pGEM-7ZF(+)' (emb1|X65310|CUGEM7ZFP). The sequence starts with: >emb1|X65310|CUGEM7ZFP Cloning vector pGEM-7ZF(+) gggcgaattggcccccacgtcgcatgctcctctagactcgaggaaatcggtaccccggtt... The sequence continues for several lines.

首先用 EMBOSS 中的 remap 看 pGEM7Z Vector 的核酸限制酵素圖譜，只看 XbaI 和 Xhol，指令如下：

```
#remap -sequence pGEM7Z.txt -enzymes xhol,xbai -notranslation -reverse -outfile pGEM7Z.remap  
-auto
```

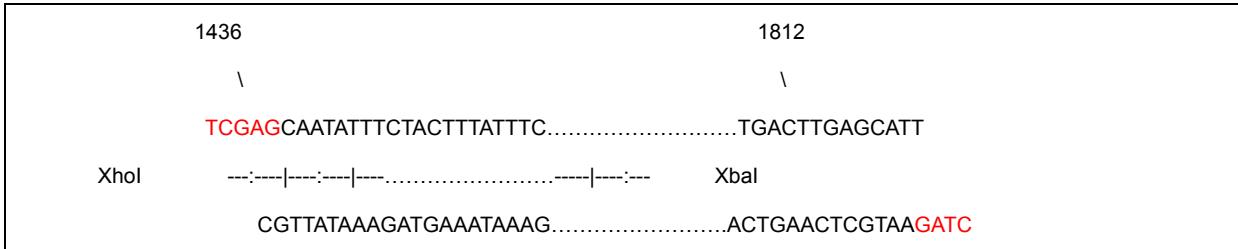
所到結果如下：

```
Cloning vector pGEM-7zf(+) 第 36,37 個 base 要被刪除, vg1-340 要接在第 35 個 base
```

```
XbaI      XhoI  
          \      \  
GGCGAATTGGCCCGACGTCGCATGCTCTT CTAGAC TCGAGGAATTGGTACCCGGGT  
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|  
CCCGCTTAACCCGGGCTGCAGCGTACGAGGA GATC TGAGCT CCTTAAGCCATGGGCCCA  
          /      /  
XbaI      XhoI
```

Vg1 PCR 產物如下，所以要先割下 1436 到 1812 區間的序列，再做一次反轉才可以接到 Vector 上。

附錄三 習題參考解答



使用的指令如下：

```
segret -sequence vq1.txt -sbegin1 1436 -send1 1812 -srev1 -outseq vq1-340.txt -auto
```

產生的新檔案 vg1-340.txt

```
>XLGFTB M18055 X.laevis transforming growth factor-beta (Vg1) gene, complete cds.

aatgctcaagtcatatggactattatatatttttatttcataatggaggattacgcc

tctgtgcagtgaatagtgcacccatctaatttttagtgaatagaaaagtcatcag

gttatataccatgttcttctaagatattattattaacaacatcatatagtgtatataca

ttcacacaagaatcaaagtgaattttagcatatgtgtccttaacagctaacagctaacag

tcaaggcaaataatgatgtatgtccaaacacccttactagcactcaagataactctgt

catgtacaggactgtgaatacgttaagaatgataaaatctatgacataacagtgtagaaa

taaaggtagaaatattgc
```

利用 EMBOSS 中的 MSE，編輯 pGEM7Z.txt，去掉第 36、37 個鹼基，從第 35 個以後插入 vg1-340.txt。答案如下：

```
PLS340 Length: 3372 December 17, 19102 14:36 Check: 8756 ..  
  
1 gggcgaattg ggcccgacgt cgcatgctcc tctagaatgc tcaagtcata  
51 tggactatta tatattttt tattttcata ttggaggatt acgcctctgt  
101 gcagtgaata gtgcaaggcac ctcatctaatttttgtgaaa tagaaaagtc  
151 atcaggttat ataccatgtt cttctaaat attattattt acaacatcat  
201 atagtgtat atacattcac acaagatcaa agtgaatttt agcatatgtg  
251 ctccttaaca gctaacagct aacagtcaag gcaaatcaat atgatgtatgt  
301 cccaaacaccc ttactagcac tcaagataact ctgtgcattgt acaggactgt  
351 gaatacgtta agaatgataa aatctatgac ataacagtgt agaaataaaag  
401 tagaaatatt gctcgaggaa ttccgttaccc cgggttcgaa atcgataaagc  
451 ttggatccgg agagctccca acgcgttgaa tgcatagctt gagtattctaa  
501 tagtgtcacc taaatagctt ggctaaatca tggctcatagc tttttccatgt
```

附錄三 習題參考解答

551 gtgaaattgt tatccgctca caattccaca caacatacga gccggaaagca
601 taaaatgtaa agcctggggt gcctaattgag tgagctact cacattaatt
651 gcgttgcgt cactgcccgc tttccagtcg ggaaacctgt cgtgccagct
701 gcattaatga atcgccaac gcgcggggag aggccgttgc cgtattgggc
751 gcttccgc ttccctcgctc actgactcgc tgccgtcggt cgttccgc
801 cggcgagcgg tatcagctca ctcaaaggcg gtaatacggt tatccacaga
851 atcaggggat aacgcaggaa agaacatgtg agaaaaaggc cagaaaaagg
901 ccaggaaccg taaaaaggcc gcgttgctgg cgttttcca taggctccgc
951 cccccgtacg agcatcacaa aaatcgacgc tcaagtccaga ggtggcgaaa
1001 cccgacagga ctataaagat accaggcggtt tccccctggc agctccctcg
1051 tgcgctctcc tgttccgacc ctgcccgtta ccggataacct gtccgccttt
1101 ctcccttcgg gaagcgtggc gctttctcat agctcacgct gtaggtatct
1151 cagttcggtg taggtcggtc gctccaagct gggctgtgtg cacgaacccc
1201 ccgttcagcc cgaccgctgc gccttatccg gtaactatcg tcttgagtcc
1251 aacccgttaa gacacgactt atcgccactg gcagcagcca ctggtaacag
1301 gattagcaga gcgagggtatg taggcgggtgc tacagagttc ttgaagtgg
1351 ggcttaacta cggctacact agaagaacag tatttggtat ctggctctg
1401 ctgaagccag ttaccttcgg aaaaagagtt ggtagctt gatccggcaa
1451 acaaaccacc gctggtagcg gtggttttt tgttgcaag cagcagattt
1501 cgccgagaaa aaaaggatct caagaagatc ctttgatctt ttctacgggg
1551 tctgacgctc agtggAACGA aaactcacgt taaggattt tggtcatgag
1601 attatcaaaa aggatcttca cctagatcct tttaaattaa aaatgaagtt
1651 ttAAATCAAT ctAAAGTATA tatgagtaaa ctgggtctga cagttaccaa
1701 tgcttaatca gtgaggcacc tatctcagcg atctgtctat ttctttcatt
1751 catagttgcc tgactccccg tcgtgttagat aactacgata cgggaggggct
1801 taccatctgg ccccaagtgc gcaatgatac cgccgagaccc acgttcacc
1851 gctccagatt tatcagcaat aaaccagcca gccggaaaggc cccggccgac
1901 aagtggcct gcaactttat ccgcctccat ccagtctatt aatttgttgc
1951 gggaaagctag agtaagttagt tcggccgttta atagtttgcg caacgttgg
2001 gccattgcta caggcatcgt ggtgtcacgc tcgtcggttgc gtatggcttc
2051 attcagctcc gtttccaaac gatcaaggcg agttacatga tccccatgt
2101 tgtgcaaaaa agcgggttagc tccttcggtc ctccgtatgc tgtcagaagtt
2151 aagttggccg cagtgttatac actcatggtt atggcagcac tgccataatt
2201 tcttactgtc atgcccattcg taagatgctt ttctgtgact ggtgagttact
2251 caaccaagtc attctgagaa tagtgtatgc ggcgaccgag ttgtcttgc
2301 ccggcgtcaa tacgggataa taccgcgcacatgcagaa ctttaaaatgt
2351 gctcatcatt gaaaaacgtt ctccggggcg aaaactctca aggatcttac
2401 cgctgttgag atccagttcg atgttaaccca ctcgtgcacc caactgtatct
2451 tcaagcatctt ttactttcac cagcgtttct qgggtqagcaa aaacagggaa

附錄三 習題參考解答

```
2501 gcaaaatgcc gcaaaaaagg gaataagggc gacacggaaa tggtgaatac
2551 tcatactttt ccttttcaa tattattgaa gcatttatca gggttattgt
2601 ctcatgagcg gatacatatt tgaatgtatt tagaaaaata aacaaatagg
2651 ggttccgcgc acatttcccc gaaaagtgcc acctgtatgcg gtgtgaaata
2701 ccgcacagat gcgttaaggag aaaataccgc atcaggaaat tgtaagcggt
2751 aatattttgt taaaattcgc gttaaatttt tgtaaatca gctcattttt
2801 taaccaatag gccgaaatcg gcaaaatccc ttataaatca aaagaataga
2851 ccgagatagg gttgagtgtt gttccagttt ggaacaagag tccactatta
2901 aagaacgtgg actccaacgt caaaggcga aaaaccgtct atcagggcga
2951 tggcccacta cgtgaaccat caccctaatac aagtttttg gggtcgaggt
3001 gccgtaaagc actaaatcg aaccctaaag ggagcccccg atttagagct
3051 tgacgggaa agccggcga cgtggcgaga aaggaaggga agaaagcga
3101 aggagcgggc gctagggcgc tggcaagtgt agcgtcacg ctgcgcgtaa
3151 ccaccacacc cgccgcgcctt aatgcgccgc tacagggcgc gtccattcgc
3201 cattcaggct gcgcaactgt tgggaaggc gatcggtgcg ggcctttcg
3251 ctattacgcc agctggcga aaaaaaaaaaaaaatgt gctgcaaggc gattaagtt
3301 ggtaacgcca gggtttccc agtcacgacg ttgtaaaacg acggccagtg
3351 aattgtaata cgactcacta ta
```

Problem2.

使用指令如下：

```
#remap -sequence pLS340.txt -sformat1 gcf -enzymes all -sitelen 6 -notranslation -noreverse
-outfile pLS340.remap -auto
```

附錄三 習題參考解答

pLS340.remap 如下：

附錄三 習題參考解答

-----|-----|-----|-----|-----|-----|-----|

AcsI

\

ATTATTATTAACAAACATCATATAGTGTATATACATTACACAAGATCAAAGTGAATTT

190 200 210 220 230 240

-----|-----|-----|-----|-----|-----|-----|

BmyI

FauNDI Alw21I

\ \

AGCATATGTGCTCCTAACAGCTAACAGCTAACAGTCAAGGCAAATCAATATGATGATGT

250 260 270 280 290 300

-----|-----|-----|-----|-----|-----|-----|

AauI

SmlI TatI

BpuEI | Hpy188III | BstNSI

\ \ \ \\

CCCAACACCCTACTAGCACTCAAGATACTCTGTGCATGTACAGGACTGTGAATACGTTA

310 320 330 340 350 360

-----|-----|-----|-----|-----|-----|-----|

SmlI

BssHI

Ama87I

| PsrI

| | AcsI

MslI TspRI SspI | | EcoRI

\ \ \ \ \\

AGAATGATAAAATCTATGACATAACAGTAGAAATAAAGTAGAAATATTGCTCGAGGAA

370 380 390 400 410 420

-----|-----|-----|-----|-----|-----|-----|

BanIII

| HindIII

| | PsrI

| | | BamHI

| | | BstX2I

附錄三 習題參考解答

| | | | BscBI
 | | | | | BsaWI
 | | | | | AccIII
 | | | | | | Hpy188III
 | | | | | | | Ecl136II
 | | | | | | | | BmyI
 Acc65I | | | | | | | | BanII
 AccB1I | | | | | | | | Alw21I
 | BscBI | | | | | | | | Psp124BI
 | | KpnI | | | | | | | | MwoI
 | | BsaJI | | | | | | | | MluI
 | | | BsaJI | | | | | | | | AfI III
 | | | Cfr9I | | | | | | | | BstXI
 | | | Ama87I | | | | | | | | MwoI
 | | | | SmaI AsuII | | | | | | | | | | Ppu10I
 \| \| \| \| \| \| \| \| \| \| \| \| \| \| \| \| \|
 TTCGGTACCCGGGTTCGAAATCGATAAGCTGGATCCGGAGAGCTCCAACGCCTTGA
 430 440 450 460 470 480
 -----:-----|-----:-----|-----:-----|-----:-----|-----:-----|

BfrBI
 | EcoT22I
 | | SmlI BfmI BpuEI
 \| \| \| \| \|
 TGCATAGCTTGAGTATTCTATAGTCACCTAAATAGCTTGGCGTAATCATGGTCATAGC
 490 500 510 520 530 540
 -----:-----|-----:-----|-----:-----|-----:-----|-----:-----|

AccBSI
 \|
 TGTTTCCTGTGTGAAATTGTTATCCGCTCACAAATTCCACACAACATACGAGCCGGAAGCA
 550 560 570 580 590 600
 -----:-----|-----:-----|-----:-----|-----:-----|-----:-----|

BsaJI
 | MwoI
 | AccB1I
 | | BscBI AseI BtsI
 \| \| \| \| \|

附錄三 習題參考解答

TAAAGTGTAAAGCCTGGGTGCCTAATGAGTGAGCTAACATTAATTGCAGTCGCT

610 620 630 640 650 660

-----|-----|-----|-----|-----|-----|-----|-----|

BstC8I

| PvuII

TspRI

| MspAII

MwoI BstC8I Hpy188III | | AseI CfrI BsaXI

\ \ \ \ \ \ \ \

CACTGCCGCTTCCAGTCGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAAC

670 680 690 700 710 720

-----|-----|-----|-----|-----|-----|-----|

MwoI

| BsaXI

| | Bsp143II

| | | SapI TspRI

Bsc4I | | | Bst6I TaqII

\ \ \ \ \ \ \

GCGCGGGGAGAGGGCGGTTGCGTATTGGCGCTCTCCGCTTCCTCGCTCACTGACTCGC

730 740 750 760 770 780

-----|-----|-----|-----|-----|-----|-----|

BstC8I

| AccBSI

BsaOI | | MwoI

\ \ \ \ \

TGCCTCGTCGGCTGGCTGCGCGAGCGGTATCAGCTCACTCAAAGGCAGTAATACGGT

790 800 810 820 830 840

-----|-----|-----|-----|-----|-----|-----|

AfI^{III}

BspLU11I BstC8I

| BstNSI | Bsc4I

\ \ \ \

TATCCACAGAACGAGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGG

850 860 870 880 890 900

-----|-----|-----|-----|-----|-----|-----|

附錄三 習題參考解答

MwoI

BscBI	Bsc4I		BstC8I	EciI	BscBI
\	\		\	\	\
CCAGGAACCGTAAAAAGGCCGTTGCTGGCTTTCCATAGGCTCCGCCCCCTGACG					
910	920	930	940	950	960
-----:----- -----:----- -----:----- -----:----- -----:-----					

SmlI

BpuEI		DrdI			
\		\ \			
AGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAACCCGACAGGACTATAAAGAT					
970	980	990	1000	1010	1020
-----:----- -----:----- -----:----- -----:----- -----:-----					

BciVI

				Bsc4I	
BsaJI	BssSI			BsaWI	
\	\			\ \	
ACCAGGCCTTCCCCCTGGAAGCTCCCTCGTGCCTCTCCGTCCGACCCCTGCCGCTTA					
1030	1040	1050	1060	1070	1080
-----:----- -----:----- -----:----- -----:----- -----:-----					

EciI	Hpy188III	Bsp143II	BfmI		
\	\	\	\		
CCGGATACCTGTCGCCCTTCTCCCTCGGAAGCGTGGCGCTTCATAGCTCACGCT					
1090	1100	1110	1120	1130	1140
-----:----- -----:----- -----:----- -----:----- -----:-----					

Alw44I

				Hpy8I	
				BmyI	
				Alw21I	
				Bme1580I	
				\ \ \	
GTAGGTATCTCAGTCGGTGTAGGTCGTCGCTCCAAGCTGGGCTGTGCACGAACCCC					
1150	1160	1170	1180	1190	1200
-----:----- -----:----- -----:----- -----:----- -----:-----					

BsaOI

SmlI

附錄三 習題參考解答

MspAI | BsaWI \ Hpy188III \

1210 1220 1230 1240 1250 1260

----:----|----:----|----:----|----:----|----:----|----:----|

CCGTTCAGCCCCGACCGCTGCGCCTTATCCGGTAACCTATCGTCTTGAGTCACCCGGTAA

AlwNI

BpuEI TspRI | TspRI

\ \ \ \

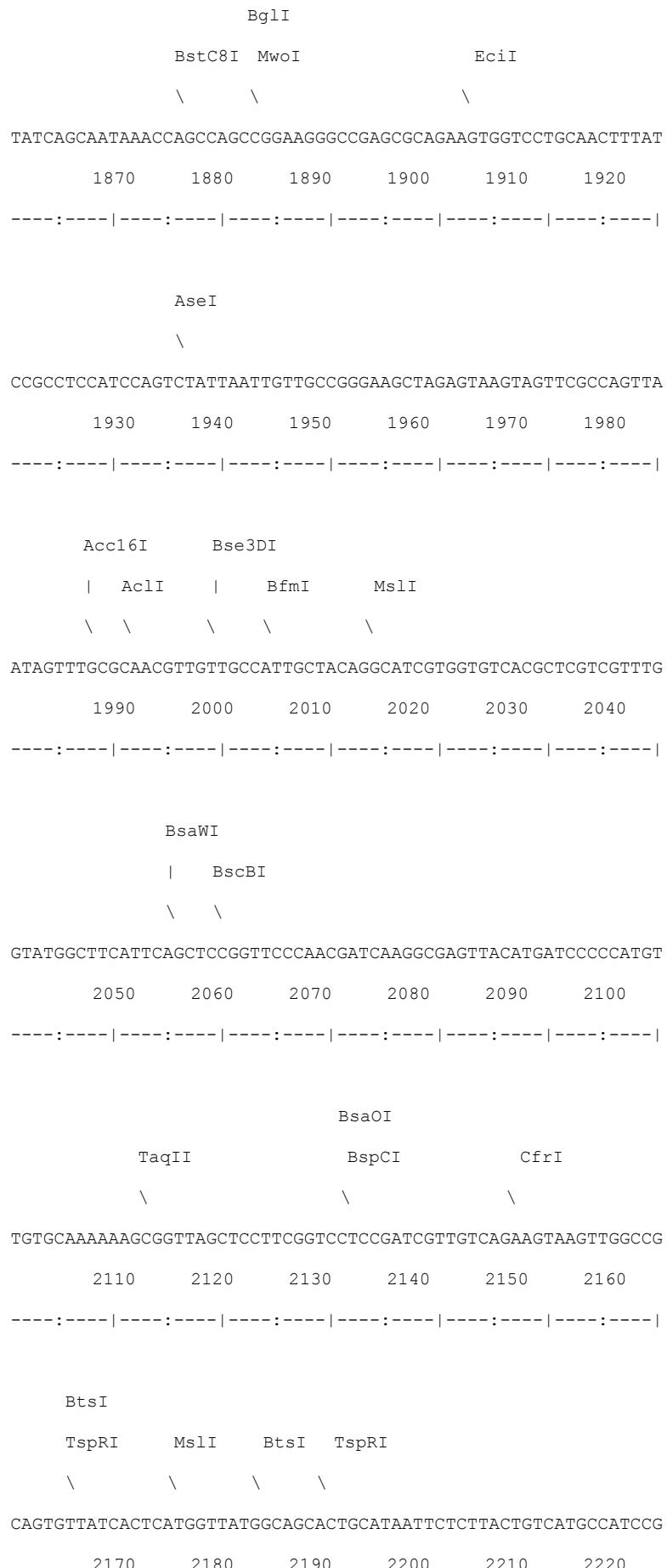
GACACGACTTATGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATG

1270 1280 1290 1300 1310 1320

-----|-----|-----|-----|-----|-----|-----|-----|-----|

BfmI	Hpy188III	Bsc4I			
\	\	\			
1330	1340	1350	1360	1370	1380
----- ----- ----- ----- ----- ----- ----- ----- -----					

附錄三 習題參考解答



附錄三 習題參考解答

附錄三 習題參考解答

PstI Bsc4I TaqII
 \ \ \

GCAAAATCCCTATAAATCAAAGAATAGACCGAGATAGGGTTGAGTGTGTTCCAGTTT
 2830 2840 2850 2860 2870 2880
 -----:----|----:----|----:----|----:----|----:----|----:----|

Hpy8I
 | PpiI BsaXI
 | AloI | PpiI
 | BsaXI Hpy8I DrdI | AloI
 \ \ \ \ \ \ \ \

GGAACAAGAGTCCACTATTAAAGAACGTGGACTCCAACGTCAAAGGGCGAAAACCGTCT
 2890 2900 2910 2920 2930 2940
 -----:----|----:----|----:----|----:----|----:----|----:----|

AdeI
 BsaAI AccB1I
 | Hpy8I | BscBI
 \ \ \ \

ATCAGGGCGATGGCCCCTACGTGAACCATCACCTAATCAAGTTTTGGGTCGAGGT
 2950 2960 2970 2980 2990 3000
 -----:----|----:----|----:----|----:----|----:----|----:----|

BscBI
 | BmyI
 BscBI | BanII
 \ \ \

GCCGTAAGCACTAAATCGAACCTAAAGGGAGCCCCGATTAGAGCTTGACGGGAA
 3010 3020 3030 3040 3050 3060
 -----:----|----:----|----:----|----:----|----:----|----:----|

MwoI
 MroNI | AccBSI
 Bse118I | | BstC8I
 | NaeI | | | Bsp143II
 | BstC8I MwoI | | | | Bsp143II
 \ \ \ \ \ \ \ \ \ \

AGCCGGCGAACGTGGCGAGAAAGGAAGGGAAAGCGAAAGGAGCGGGCGTAGGGCGC
 3070 3080 3090 3100 3110 3120

附錄三 習題參考解答

```

          BstC8I
          |
          | PvuII
          |
          | MspAII
Bst6I      |   | BstC8I
\           \   \
GGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTG
3250      3260      3270      3280      3290      3300
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|

```

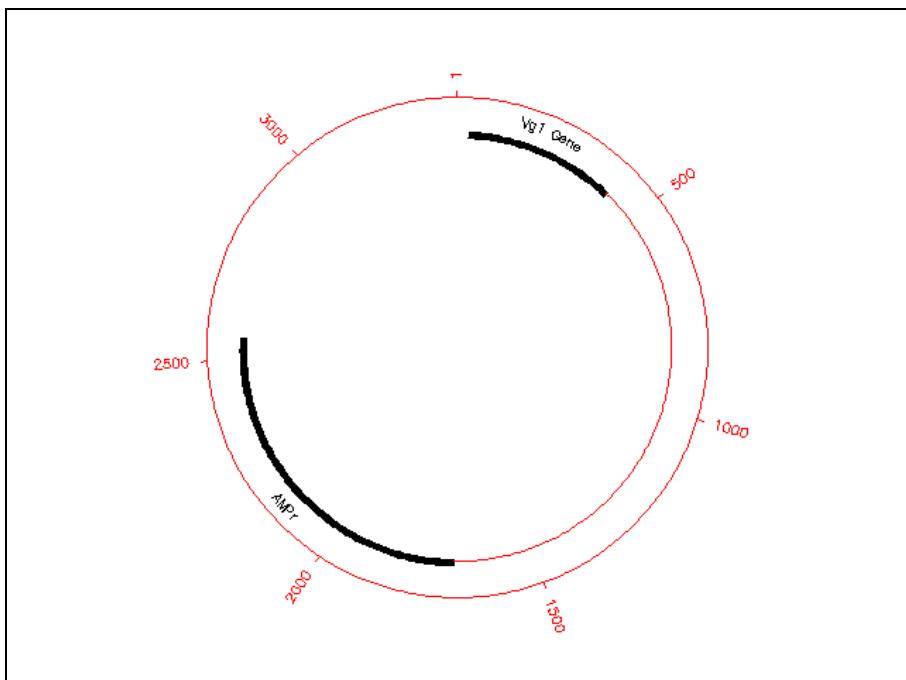
BsaJI

	BfiI	Hpy188III	CfrI	TspRI
\	\	\	\	\
GGTAACGCCAGGGTTTCCAGTCACGACGTTGTAAACGACGGCCAGTGAATTGTAATA				
3310	3320	3330	3340	3350
----- ----- ----- ----- ----- ----- ----- ----- ----- -----				

CGACTCACTATA
3370 3380 3390 3400 3410 3420

Problem3.

```
#cirdna -graphout png -inputfile d_tmp_emboss_cirdna-pLS340 -ruler Y -blocktype Filled  
-originangle 90.0 -posticks Out -posblocks Out -intersymbol Y -intercolor 1 -interticks N  
-gapsize 500 -ticklines N -textheight 10.0 -textlength 2.0 -tickheight 1.0 -blockheight 1.0  
-rangeheight 1.0 -gapgroup 1.0 -postext 1.0 -auto
```

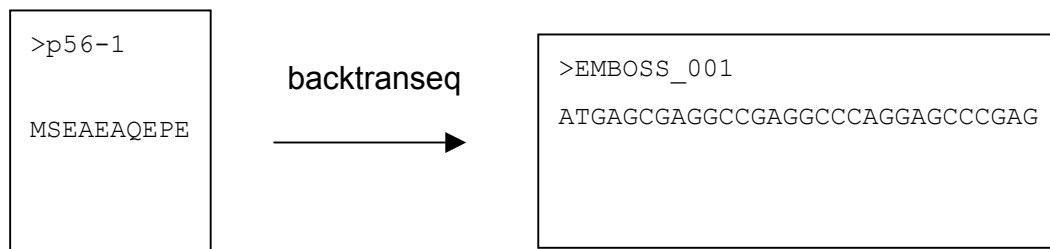


Problem 4.

為確定你所純化出的蛋白質不是 artifact，請你將所得到片段的序列，請至 NCBI BLAST 網站做 BLASTP 與資料庫中的序列比對，以確定它們不是來自於一些已知的，存量豐富的蛋白質。結果發現，並無在分離過程中可能引入的蛋白質，因此判定沒有 artifact。

Problem 5.

使用 EMBOSS 中 backtranseq 將 p56-1 及 p56-2 的蛋白質序列變成 DNA 序列當成 probe 去釣 cDNA Library。

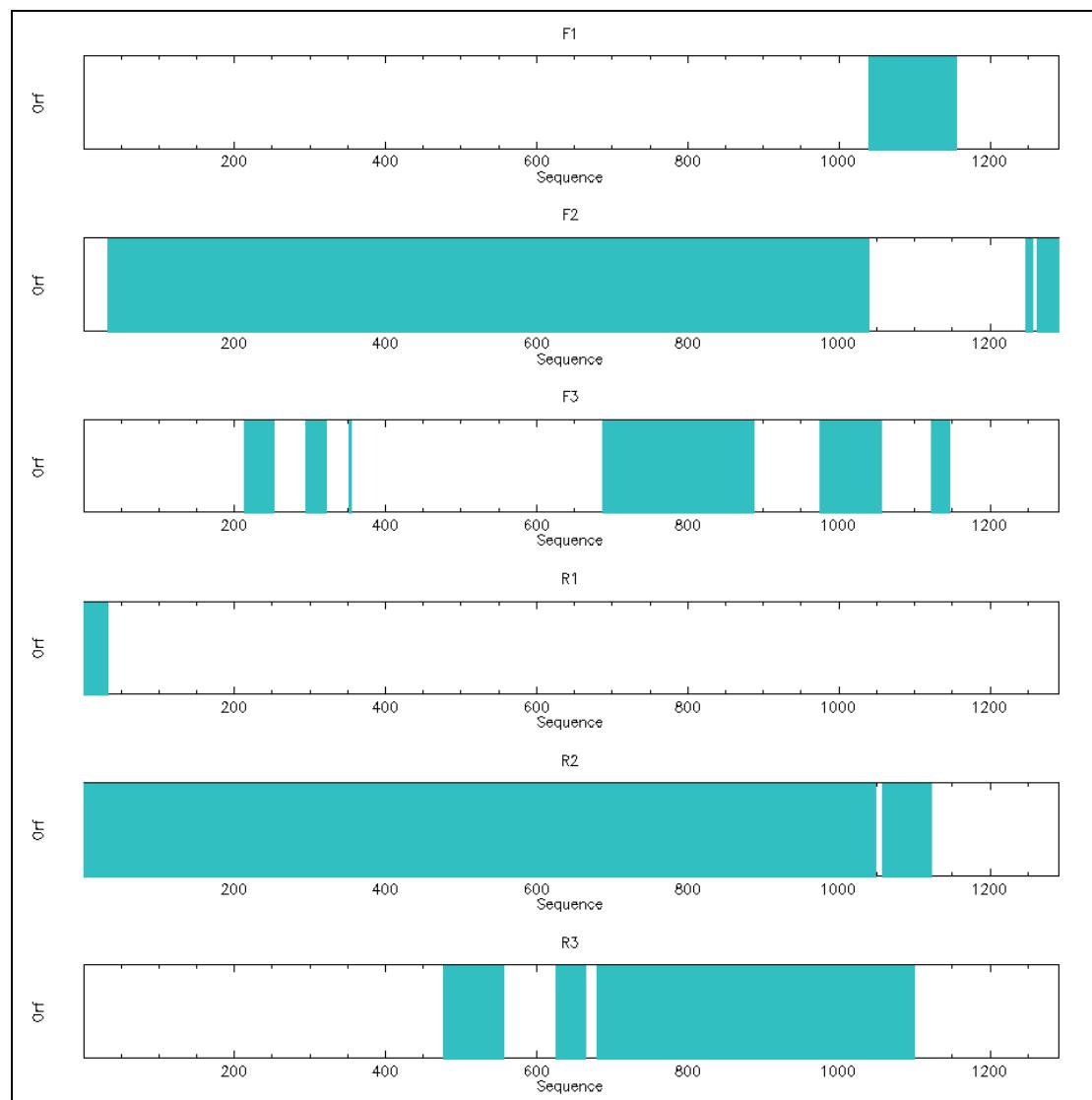


Problem 6.

在 EMBOSS 中 plotorf 能將核酸序列正反向三個 Frames 做 translation，指令如下：

```
#plotorf -sequence p56.txt -start ATG -stop TAA,TAG,TGA -graph png -auto
```

由下圖可知，第二的 Frame 應為可能的 open reading frame



Problem 7.

請至 NCBI BLAST 網站做 blastn 的搜尋。發現與序號 M59454 *X. laevis* sequence -specific binding protein (FRGY2) mRNA 有很高的相似性。

Problem 8.

請至 NCBI BLAST 網站做 blastn 的搜尋。發現與序號 B38274 非洲爪蛙 Y box-binding protein 2 有很高的相似性。

Problem 9.

使用 EMBOSS 中的指令 Needle 做比對，發現完全相似。

Problem 10.

使用 EMBOSS 中的 emma 做多條序列比對

```
!!NA_MULTIPLE_ALIGNMENT 1.0  
  
p56.aln MSF: 1577 Type: N 17/12/02 CompCheck: 1408 ..
```

```
Name: D26328 Len: 1577 Check: 4229 Weight: 22.50  
Name: L13032 Len: 1577 Check: 4351 Weight: 22.50  
Name: p56 Len: 1577 Check: 6787 Weight: 30.00  
Name: M80257 Len: 1577 Check: 6041 Weight: 25.00
```

//

1	50
D26328	CGGGAGCGGAGAGCGGAGCTCAGCGAGCCCCGGAGCAGGGAGCAGCCCC
L13032	~~~GAGCGGGAGAGCGGAGCTCAGCGAGCCCCGGAGCAGGGAGCAGCCCC
p56	~~~~~
M80257	~~~~~

51	100
D26328	CACGGCCGGCCTCAGTCACCATCACACCGCGGGAGGCAGCGCAAACAGCC
L13032	CACGGCCGGCCTCAGTCACCATCACACCGCGGGAGGCAGCGCAAACAGCC
p56	~~~~~
M80257	~~~~~

101	150
D26328	AGTCACCACCACCGTCCGCACCGAACAGCCGCCATGAGCAGCGAGGC

附錄三 習題參考解答

L13032	AGTCACCACCGTCCGCACCGAACAGCCGCCATGAGCAGCGAGGC	
p56	ACACGCCGGTACCGTC.ATCTCGAGGAGGAGC....ATGAG...TGAGGC	
M80257	~~~~~GAGGAGGAGC....ATGAG...TGAGGC	
	151	200
D26328	CGAGACCCAGCCGCCGCCGCCGCCGTCCCCGGGGCCCCGCCGCC	
L13032	CGAGACCCAGCCGCCGCCGCCGTCCCCGCC...CCCCGCCGCC	
p56	GGAAGCCCAGGAGCCGA.....ACCGGT....GCCACAACCG	
M80257	GGAACCCGAGAGACTGA.....AGCGGT....TACACAGCCG	
	201	250
D26328	CCCCCGCCGACTCAAACCTAACGGCGCAGCGGGAACGGGAGCAGCGC	
L13032	CCCCCGCCGACTCAAACCTAACGGCGCAGCGGGAACGGGAGCAGCGC	
p56GAGTCGAACC.....CGAGATC.....CAGAAGC	
M80257GAACCCGGACC.....CGAGATC.....CACAAGC	
	251	300
D26328	CTGGCCTCGCGCGCCTCCGCCGGGGACAAGAAGGTATCGAAC	
L13032	CTGGCCTCGCGCGCCTCCGCCGGGGACAAGAAGGTATCGAAC	
p56	CCGGAATTGCCG.CGGCTCGAAACCAGGCCAACAGAAAGTGCTGCCAC	
M80257	CGGACATTGTCC.CGCCTCGAACAGATCAACAAGAAGCTTCTGCCAC	
	301	350
D26328	GAAGGTTTGGAACAGTGAATGGTCAACGTACGGAACGGTTACGGCT	
L13032	GAAGGTTTGGAACAGTGAATGGTCAACGTACGGAACGGTTACGGCT	
p56	TCAAGTGCAGGAAACAGTGAAGTGGTTAACGTTCGAACGGCTACGGAT	
M80257	CCAAGTGCAGGAAACAGTCAAGTGGTTAACGTCCGAAACGGTTACGGAT	
	351	400
D26328	TCATCACAGGAATGACACCAAGGAAGATGTGTTGTCCATCAGACTGCC	
L13032	TCATCACAGGAATGACACCAAGGAAGATGTGTTGTCCATCAGACTGCC	
p56	TTATCACAGAAATGACACCAAAGAAGATGTGTTGTGCATCAGACTGCA	
M80257	TTATCACAGAAATGACAGCAAAGAAGATGTGTTGTGCATCAGACTGCA	
	401	450
D26328	ATAAAGAAGATAACCCAGGAAGTACCTCCGCAGCGTAGGAGATGGAGA	
L13032	ATAAAGAAGATAACCCAGGAAGTACCTCCGCAGCGTAGGAGATGGAGA	
p56	ATAAAAAGAACAAATCCACGGAAGTTCTGCGCAGTGTGGGTGATGGTGA	
M80257	ATAAAAAGAACAAATCCACGGAAGTTCTGCGCAGTGTGGGTGATGGTGA	

附錄三 習題參考解答

	451	500	
D26328	GACCGTGGAGTTGATGTGGTTGAAGGAGAGAAGGGTGCGGAGGCAGCGA		
L13032	GACCGTGGAGTTGATGTGGTTGAAGGAGAGAAGGGTGCGGAGGCAGCGA		
p56	GACGGTGGAGTTGATGTCGTAGAAGGAGAGAAGGGAGCAGAGGCCA		
M80257	GACGGTGGAGTTGATGTCGTAGAAGGAGAGAAGGGAGCAGAGGCCA		
	501	550	
D26328	ATGTGACAGGGCCTGGTGGCGTCCAGTGCAAGGCAGCAAATACGCAGCA		
L13032	ATGTGACAGGGCCTGGTGGCGTCCAGTGCAAGGCAGCAAATACGCAGCA		
p56	ATGTGACGGGCCAGGAGGGTCCCAGTTAAAGGGAGTCGCTTGCCCCA		
M80257	ATGTGACGGGCCAGGAGGGTCCCAGTTAAAGGGAGTCGCTTGCCCCA		
	551	600	
D26328	GACCGTAACCATTACAGACGATATCCCGTGC.G.AGGGTCCCTCACGCA		
L13032	GACCGTAACCATTACAGACGATATCCCGTGC.G.AGGGTCCCCACGCA		
p56	AACAG.A.....CGCAGGTTCGCCGGCTTCTACAGGCCCCGTGCGGA		
M80257	AACAGCA.....CG.AGGTTCGCCGGCAATTCTACAGGCCCCGTGCGGA		
	601	650	
D26328	.ACTACCAGCAAAACTACCAGAACAGTGAG.AGT...GGGGAGAAGAACG		
L13032	.ACTACCAGCAAAACTACCAGAACAGTGAG.AGT...GGGGAGAAGAACG		
p56	TACTGCGGGAGAGTCTGGGGTGAAGGGTTAGTCCTGAACAGATGAGTG		
M80257	TACTGCGGGAGAGTCTGGGGTGAAGGGTTAGTCCTGAGCAGATGAGTG		
	651	700	
D26328	AAGGAGCAGAGAACATCCCGAAGGCCAAGCCCAGCAGCGCGTCCCTAT		
L13032	AAGGAGCAGAGAACATCCCGAAGGCCAAGCCCAGCAGCGCGTCCCTAT		
p56	AAGGGG.AGAGAGGG...GAGGAGACTT..CCCCACAGCAGAGACCCAG		
M80257	AAGGGG.AGAAAGGG...GAGGAGACTT..CCCCACAGCAGCGACCCAG		
	701	750	
D26328	CGCAGGCGCGGTACCCACCTTACTACATGCGTAGGCCCTACGGCGTC.		
L13032	CGCAGGCGCGGTACCCACCTTACTACATGCGTAGGCCCTACGGCGTC.		
p56	CGTAGGCGACCCCTCCATTCTTCTACAGAAGGCGCTTCAGAAGGGTCC		
M80257	CGTAGGCGACCCCTCCATTCTTCTACAGAAGGCGCTTCAGAAGGGTCC		
	751	800	
D26328	..GACCACAATATTCCAACCTCCGTGCAGGGAGAGATA GTGGAGGGTG		

附錄三 習題參考解答

L13032	..GACCACAATATTCCAACCCCTCCCGTGCAGGGAGAGATA GTGGAGGGTG	
p56	CAGACC.CAATAA.CCAGCAGAAC CAGGGAGC.AGAGGTTACAGAGCAGT	
M80257	CAGACC.CAATAA.CCAACAGAAC CAGGGAGC.AGAGGTTACAGACCAGT	
	801	850
D26328	CTGACAACCAAGGTGCAGGAGA....ACA....AGGCA..GACCAGT...	
L13032	CTGACAACCAAGGTGCAGGAGA....ACA....AGGCA..GACCAGT...	
p56	CTGAGAATAAGGACCCAGTTGCC CACATCGGAAGCC TGCTAGTGGT	
M80257	CTGAGAATAAAAGACCCTGCTGCC CACATCGAAGCC TGCTAGTGGA	
	851	900
D26328CAGGCAGAACATGTA...TCGAGGTTACAGACCACGATTCCGCAG	
L13032CAGGCAGAACATGTA...TCGAGGTTACAGACCACGATTCCGCAG	
p56	GATGATCCGCAGAGACCGCCCCCTCGCAGGTTCCGACAAAGATTCCGCAG	
M80257	GACGGCCAGCAGAGACCACCCCTCGCAGGTTCCAGCAAAGATTCCGCAG	
	901	950
D26328	CTTGACTT CAGGGT CCTCGTCAA.AGACAGCCTAGAGAGGATGGA	
L13032GGGT CCTCGTCAA.AGACAGCCTAGAGAGGATGGA	
p56	GCCTT CCGT CCTCGCCCTGCACCTCAGCAGAC..CCCAGAAGGA..GGT	
M80257	GCCTT CCGT CCTCGCC CACCTCCACAGAC..CCCAGAAGGG..GGA	
	951	1000
D26328	AACGAAGAAGATAAAGAGAACCAAGGAGATGAGACCCAAGGT CAGCAGCC	
L13032	AACGAAGAAGATAAAGAGAACCAAGGAGATGAGACCCAAGGT CAGCAGCC	
p56	GACGGAGAGACCAAGCTGAATCTGGGA..AGATCCTCGGCCGG.AGCC	
M80257	GATGGAGAACCAAAGCTGAA...GGTGA..A.....CC	
	1001	1050
D26328	ACCTCACGTCGGTACCG.TCGTA ACTTCAACTACAGACGCAGACGCCA	
L13032	ACCTCACGTCGGTACCG.TCGTA ACTTCAACTACAGACGCAGACGCCA	
p56	ACAGAGGCAACGTAACCGACCGTATGTCAA C....GGCGCAGGCCAA	
M80257	ACAGAGGCAACGTAACCGACCGTATGTCAA C....GGCGCAGG.GCCCA	
	1051	1100
D26328	GAGAAC.CCTAAACCACAAG.ATGGCAAAGAGACGAAGACGCCAACCA	
L13032	GAGAAC.CCTAAACCACAAG.ATGGCAAAGAGACGAAGACGCCAACCA	
p56	GGGCCACCCAAAGTAGCGGCCACAGCCCAG.GGTGAGGGCAA.GCAGAA	
M80257	ACAACCACCCA.....CAGTCCAG.GGTGAGAGCAA.GCAGAA	

附錄三 習題參考解答

	1101	1150
D26328	CCAGCTGAGAACACGTCCGCTCCCGAGGCCGAGCAGGGCGGGCTGAGTA	
L13032	CCAGCTGAGAACACGTCCGCTCCCGAGGCCGAGCAGGGCGGGCTGAGTA	
p56	CCAACTCAGCAC...CCTGCTTCTGA.....AGAAGGG.....	
M80257	CCAAGTGAGCAC...CCTGCTTCTGA.....AGAAGGA.....	
	1151	1200
D26328	AATACCGGCTTAACATCTTACCATCATCCGGTTAGTCATCAAAGAAAA	
L13032	AATACCGGCTTAACATCTTACCATCATCCGGTTAGTCATCAAAGAAAA	
p56	...ACC..CCCAGTGATTCTCCCA.CAGATGACGGAG.CACCTGTTCAGA	
M80257	...ACC..CCTAGTGATGCACCCA.CAGATGATGGAG.CACCCGTTGAGA	
	1201	1250
D26328	GACTTAAGAAATGAAGAAGAAAAGAAAATGAAATTCCAGCAATAAGAAAT	
L13032	GACTTAAGAAATGAAGAAGAAAAGAAAATGAAATTCCAGCAATAAGAAAT	
p56	GCTCTGCCCGGATCCAGGGATTGCAGATACACCTGCCAG.....AAT	
M80257CATCAGAACGAGGAGTGGAGGATACAACGTCCCCAG.....AAT	
	1251	1300
D26328	GAACAAAAGAATTGGAACTGAAGACCTTAAGTGCTTGCTTTTGCTGTTG	
L13032	GAACAAAAGAATTGGAACTGAAGACCTTAAGTGCTTGCTTTTGCTGTTG	
p56	GAACTCATCCGCTG...CTAGAAC.....GTGCCA.TTTTCAGGGGC..	
M80257	GAACTCAGTCGCTG...GTAGTCAT.....GGGCAT.CTATCAGGGC..	
	1301	1350
D26328	ACCAGATTACTAGAACTATCTGCATTATCTATGCAGCATGGGTTTTAT	
L13032	ACCAGATTACTAGAACTATCTGCATTATCTATGCAGCATGGGTTTTAT	
p56	A.CAGGT.GC.ACGCCTTCCG.GCTGCACTGGATT.....	
M80257	AACAGGT.GCCAGCGCC.TTCCATCTTC.G.GCTGCACTGGATT.....	
	1351	1400
D26328	TATTTTACCTAAAGAAGTCTCCTTTGGAAACAATAAACACGTTTTTA	
L13032	TATTTTACCTAAAGAAGTCTCCTTTGGAAACAATAAACACGTTTTTA	
p56CCATGGGAATT...TTGGGACACTGCAATAGTCACTTTA	
M80257CCA.GGAAATT...TCGGGACACGGCAATATTCACTTTA	
	1401	1450
D26328	AAAGTCTGTT.TTTCTCAATACACCTTAAAGGTTTAAATTGTTCA	

附錄三 習題參考解答

L13032 AAAGTCTGTT.TTTCTCAATACACCTTAAAGGTTTAAATTGTTCA
p56 ATTGGGG...TGTCCTGCACTGTTGTATTGGTGGAGGGCT..
M80257 AATAGGGGTGTTCCTCTGCACTGTTGCATTGGTGGAGGGCT..

1451 1500
D26328 TATCTGGTCAAGTTGAGATTTTAAGAACCTCATTAAATTGTATGAA
L13032 TATCTGGTCAAGTTGAGATTTTAAGAACCTCATTAAATTGTATGAA
p56 TGTTT.G.CACTTG.....TTTGGGCGTTCTGTTTATTGCACA.GAG
M80257 TGTTTG.CACTTG.....TTTGGGGGTCAGTTTATTGACA.G..

1501 1550
D26328 AAGTACGCCTGATTTTCAAGTCAAACTGCAAGCATCTGTTAATAAAGG
L13032 AAGTACGCCTGATTTTCAAGTCAAACTGCAAGCATCTGTTAATAAAGG
p56 TGATGCAGCAG....TAAAAGATGGAAGTGCAAAAAAAAAAAAAAA
M80257 TGGTGCAGCAG....TAAA.GATGGAAGT.CGAAAAA~~~~~

1551 1577
D26328 TCTTAAAGTT~~~~~
L13032 TCTTAAAGTTAAAAAAAAAAAAAAA
p56 G~~~~~
M80257 ~~~~~

使用 EMBOSS 中的 cons 來找 consensus sequence

```
!!NA_SEQUENCE 1.0
```

```
EMBOSS_001 Length: 1577 Type: N Check: 1685 ..
```

```
1 cgggagcga gagcggagct cagcgagccc ggagcaggga gcagccccc
51 cacggccgc ctcagtcacc atcacaccgc gggaggcggc gcaaacagcc
101 aNncnccnn accgtccNnn ncgaGGagGa gcccgcattga gcagtggggc
151 GgaAaccag Gagcccgacg cccccgtccc agcggtggcc cccAcagccg
201 ccccccggca ctccGaacct aacggcggca gcgAgatcgg gagcagaAgc
251 ctggcattgg cggcgccctG cAacCAGggC Aacaagaagg tcctcgccac
301 gcaAgtGcag ggaacagtga aGtggttaa cgtacgCaac ggttacggat
351 ttatcaacag Aaatgacacc aaAgaagatg tgtttgtGca tcagactgca
401 ataaaaAaaga acaatccacg gaagttctG cgcaatgtGg gtgtatggta
451 gacGgtggag tttgatgtCg tagaaggaga gaagggagcA gaggcagcA
```

附錄三 習題參考解答

501 atgtgacGgg CccaggaggG gtcccagtTa aaggGagtcG atttgccCca
551 Aacagtaacc atcGcagGtT tcGccGgcgt tTctaCAggc cccGtGcgGa
601 tactGcGGga GaGtctGGGG gtGaagGgGT tagtcctgAg CagatgaGtg
651 aaggGgcaga gaGGGtccGA gGagActtag ccccAcagcA GgcaccccaG
701 ctaggcgAc CCcctccatt cttctacaGA agGcgCttca GaAggGgtcc
751 cagaccacaa taatccaaca GaaccAGgGa gCgagagGtt AcAgagCAGT
801 ctgaGaataa agAcCcaggt gcccccacat cnGaAgccct gGctagtGGn
851 GanGncggc agaGaccgcc ccctcgcAgG ttccgacaaa gattccgcag
901 GcctttccGt ccTCgcctc cacCtcaaca gacagccag aAGggatgga
951 GacgGagaag ccaaagcTGa annnggagat gagancnng gncngcagcc
1001 acaGaGGcAa cgTaaccgac cgtatGtcca actacagGcg cagGcgc
1051 gagaccaccc aannnnncnnN nacAgcccag agGtgaGgGc aAacgcaGaa
1101 ccaActgagc acacgcctgc ttctgaggcc gagaaggcgc gggctgagta
1151 aataccggcc taGtGattct cccatcaGat gAtGGagtc ccaGtTaaGa
1201 gnnntncgac aGAAA CaGga atagaaGatA caactGccCc aGtaagaaat
1251 gaactcaagc GctggaactA Gagaccttaa gtgcttgctt tcagGGgctg
1301 accagGttGc tagcGctatt tCcatcttcn GtgcgcacT ggAttttat
1351 tatttttacc canGaaAtt tcctttGgga cacaGcaata atTcacttta
1401 aatgGGGgtt ntGttccac tGcactGttt GaATTGGtGG GaGGgGctca
1451 tGtttgtca cTttgagatt tttGGgGaGt tcaGtttat tttAcatgan
1501 tGgtGcAGca gatTTaaaa GAtGGaactg caaAaaaanN nnaanaaaNN
1551 ncttaaagtt nnnnnnnnnn nnnnnnn

Problem 12.

首先先產生蛋白質序列 p56.pep

```
#transeq -sequence p56.txt -regions 32-1042 -outseq p56.pep -auto
```

```
#patmatmotifs -sequence p56.pep -full -prune -rformat dbmotif -auto
```

帶有‘Cold-shock’ domain signature

```
#####
# Program: patmatmotifs
```

```
# Rundate: Tue Dec 17 23:31:13 2002
```

```
# Report_format: dbmotif
```

```
# Report_file: p56_1.patmatmotifs
```

```
#=====
#
# Sequence: p56_1      from: 1      to: 336
# HitCount: 1
#
# Full: Yes
# Prune: Yes
# Data_file: /export/bio1/emboss/data/PROSITE/prosite.lines
#
#=====
Length = 20
Start = position 55 of sequence
End = position 74 of sequence

Motif = COLD_SHOCK
```

NVRNGYGFINRNDTKEDVFVHQTAIKKNNP
| |
55 74

```
#-----
#
# Motif: COLD_SHOCK
# Count: 1
#
# *****
# * 'Cold-shock' domain signature *
# *****
#
# A conserved domain of about 70 amino acids has been found in prokaryotic and
# eukaryotic single-strand nucleic-acid binding proteins [1,2,3,E1]. This
# domain, which is known as the 'cold-shock domain' (CSD) is present in the
# proteins listed below.
#
# - Escherichia coli protein CS7.4 (gene cspA) which is induced in response to
#   low temperature (cold-shock protein) and which binds to and stimulates the
```

附錄三 習題參考解答

```
# transcription of the CCAAT-containing promoters of the HN-S protein and of  
# gyrA.  
# - Mammalian Y box binding protein 1 (YB1). A protein that binds to the CCAAT-  
# containing Y box of mammalian HLA class II genes.  
# - Xenopus Y box binding proteins -1 and -2 (Y1 and Y2). Proteins that bind to  
# the CCAAT-containing Y box of Xenopus hsp70 genes.  
# - Xenopus B box binding protein (YB3). YB3 binds the B box promoter element  
# of genes transcribed by RNA polymerase III.  
# - Enhancer factor I subunit A (EFI-A) (dbpB). A protein that also bind to  
# CCAAT-motif in various gene promoters.  
# - DbpA, a Human DNA-binding protein of unknown specificity.  
# - Bacillus subtilis cold-shock proteins cspB and cspC.  
# - Streptomyces clavuligerus protein SC 7.0.  
# - Escherichia coli proteins cspB, cspC, cspD, cspE and cspF.  
# - Unr, a mammalian gene encoded upstream of the N-ras gene. Unr contains nine  
# repeats that are similar to the CSD domain. The function of Unr is not yet  
# known but it could be a multivalent DNA-binding protein.  
#  
# As a signature pattern for the CSD domain we selected its most conserved  
# region which is located in its N-terminal section. It must be noted that the  
# beginning of this region is highly similar [4] to the RNP-1 RNA-binding motif.  
#  
# -Consensus pattern: [FY]-G-F-I-x(6,7)-[DER]-[LIVM]-F-x-H-x-[STKR]-x-[LIVMFY]  
# -Sequences known to belong to this class detected by the pattern: ALL, except  
# for E.coli cspF. This pattern finds 4 out of the 9 repeats of the CSD domain  
# in Unr.  
# -Other sequence(s) detected in SWISS-PROT: NONE.  
#  
# -Expert(s) to contact by email:  
# Landsman D.; landsman@ncbi.nlm.nih.gov  
#  
# -Last update: December 2001 / Text revised.  
#  
# [ 1] Doniger J., Landsman D., Gonda M.A., Wistow G.  
# New Biol. 4:389-395(1992).  
# [ 2] Wistow G.  
# Nature 344:823-824(1990).  
# [ 3] Jones P.G., Inouye M.  
# Mol. Microbiol. 11:811-818(1994).
```

```
# [ 4] Landsman D.  
#       Nucleic Acids Res. 20:2861-2864(1992).  
# [E1] http://transfac.gbf-braunschweig.de/cgi-bin/qt/getEntry.pl?C0019  
#  
# +-----+  
# | This PROSITE entry is copyright by the Swiss Institute of Bioinformatics |  
# | (SIB). There are no restrictions on its use by non-profit institutions as |  
# | long as its content is in no way modified and this statement is not |  
# | removed. Usage by and for commercial entities requires a license agreement |  
# | (See http://www.isb-sib.ch/announce/ or email to license@isb-sib.ch). |  
# +-----+
```

Problem 13.

```
#garnier -sequencea p56.pep -rformat tagseq -outfile=p56.struc -auto
```

```
#####
# Program: garnier
# Rundate: Tue Dec 17 21:45:00 2002
# Report_format: tagseq
# Report_file: p56_1.garnier
#####

=====
#
# Sequence: p56_1      from: 1      to: 336
# HitCount: 96
#
# DCH = 0, DCS = 0
#
# Please cite:
# Garnier, Osguthorpe and Robson (1978) J. Mol. Biol. 120:97-120
#
=====
#=====
```

附錄三 習題參考解答

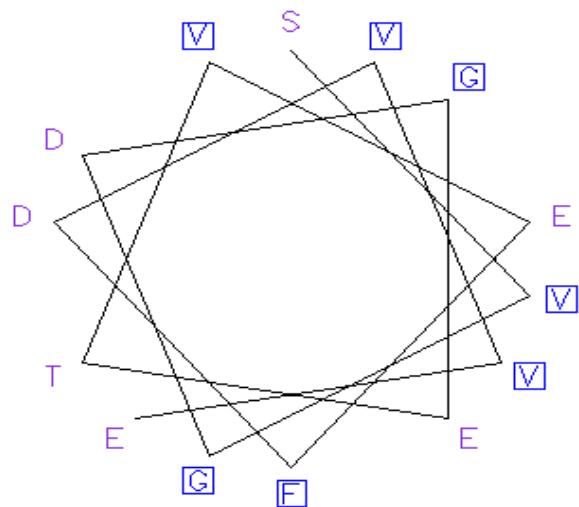
附錄三 習題參考解答

```
sheet          EEE      EE  
turns        TTT    T    TTT           T  
coil  CCCC   CC CCC   CCC   CCCCCC   CCCCCC
```

```
#-----  
#  
# Residue totals: H: 63   E: 53   T: 85   C:135  
#       percent: H: 19.7 E: 16.6 T: 26.6 C: 42.2  
#  
#-----
```

```
#pepwheel -sequence p56.pep -sbegin1 85 -send1 98 -sprotein1 -wheel -steps 18 -turns 5  
-amphipathic -graph png -auto
```

PEPWHEEL of p56.1 from 85 to 98



索引

Adenovirus	腺病毒	1-2, 12-1	Independent folding motif, IFM	獨立摺疊模組	1-3, 9-2
Algorithm	運算法	9-1, 10-1	Interaction map	交互作用圖譜	1-5
Alignment	序列排比	6-1	Internet	網際網路	1-1, 1-7, 2-1
Amphiphilic	兩性的	6-11	Kinase	磷酸激酵素	4-3, 4-21, 6-3
Assumption	假設	10-11	Ligand	配位子	9-7
Bioinformatics	生物資訊學	1-1	Local sequence alignment	區域性的序列排比	6-3
Biology workbench	生物工作台	1-8	Major groove	主溝槽	9-11
Bulge	突出	11-1, 11-20	Motif	蛋白質模組	1-3
Chimera	嵌合體	12-9	Mutation	基因突變	1-3, 9-2
Client	客戶端	2-3	Nucleic acid	核酸	3-1, 6-1
Cluster	群組	5-17, 12-5	Open reading frame, ORF	開放讀架	6-1, 7-3
Command line arguments	命令列參數	3-4	Output file format	輸出檔案格式	3-5
Command mode	指令模式	2-14	Polymerase chain reaction, PCR	聚合酵素連鎖反應	
Consensus sequence	共有序列	6-3, 6-5			1-6, 6-14, 12-3
Curve graph	曲線圖	8-6	Positional cloning	定位選殖	12-2
Cystic fibrosis	纖維囊腫	12-2	Primer	引子	3-4
Data mining	資料探採	12-14	Probe	探針	12-12
Default	參數預設值	1-8	Prompt	提示符號	3-1
Differential display	差異顯示	12-2, 12-3	Protein	蛋白質	6-1
Display	顯示	6-1	Query	查詢	4-22, 5-2
Diversity	多樣性	1-3, 9-2	Recombination	重組	1-3, 9-2
Dot matrix	點矩陣	9-1	Recursive	遞迴法	10-10
Duplication	複製	1-3, 9-2	Refine	微調	11-16, 11-20
Dynamic programming	動態線性規劃	5-11	Rules of thumb	經驗法則	5-11
Edit	編輯	6-1, 6-13	Selectivity	選擇性	5-1
Execution file	程式檔	2-5	Sensitivity	靈敏度	5-1
Expressed sequence tag, EST	表現序列標幟	5-16, 6-14	Squiggles	折曲圖	11-3
Gap separation distance	空隙分離距離	8-11	String search	字串搜尋	3-7, 9-7
Genome project	基因序列分析計畫	1-1, 1-2, 6-14	Super-helix	超螺旋	6-11
Hierarchy	層級	1-5	Terminal	終端機	2-3
Histogram	柱狀圖	5-2	Upstream sequence	基因上游序列	10-1
Hydropathy	親水性行為	6-9	Xenopus	南非水生有爪蛙	3-4, A2-1
Hydrophobic moment	疏水性矩	8-6			