目　　錄

壹、國家衛生研究院生物資訊服務簡介 3
一、前言 3

貳、SeqWeb使用說明 5
一、Login SeqWeb 5
二、SeqWeb 主網頁簡介 6
三、Sequence Manager功能 7
四、Job and Result Manager 16
五、Preference Manager 18

參、系統、資料庫及程式簡介 19
一、系統及資料庫簡介 19
二、GCG核酸及蛋白質資料庫 19
三、SeqWeb程式簡介 23

肆、序列格式簡介 27
一、序列格式種類 27
二、SeqWeb所輸出的檔案型式 30

伍、以文字搜尋資料庫 33
一、StringSearch：以字串尋找所要的序列 33
二、LookUp：以keyword尋找所要的序列 34

陸、以序列搜尋比對資料庫 35
一、序列比對分析基本概念 35
二、BLAST程式操作 38

柒、多序列並列分析 45
一、BestFit與GAP：雙序列並列分析 45
二、PileUp：多序列並列分析 47
三、Pretty：找出Consensus sequence 48

捌、尋找ORF及圖譜 51
一、Frames：尋找Open Reading Frame 51
二、Map：尋找圖譜 51
三、Translate+：轉譯序列 53

練　習：透過實例學習SeqWeb 55

參考資料 61

壹、國家衛生研究院生物資訊服務簡介

一、前言

隨著全球基因體計畫的迅速發展，序列分析已經成為生命科學研究領域的基本工具。簡單的說，序列分析就是要透過分析方法得到蘊含在序列中所有的資訊，以尋找出它們在生物學上所扮演的角色與生物意義。雖然不同的序列有著變化多端的排列組合，但是從許多研究發現已經知道序列的一些排列規則，我們因此可以依據這些規則，尋找序列中關鍵的訊息，得到更多的資訊，作為進一步研究的參考。序列分為核酸序列、蛋白質序列二大類。研究者可以分析序列的基本性質，如核酸序列的酵素切割區、尋找 PCR 引子，或是蛋白質序列的親疏水性質及帶電性質等。更進一步，是分析序列的基因資訊，如核酸序列的 ORF (open reading frames)、exon 和 intron 區域、找尋同源性（homology）序列；以及找尋蛋白質序列的 motif 或 domain、對序列的蛋白質二級結構區域的預測、找尋同源性序列等。

序列分析有一個特色，就是要參照比對的資料庫種類很多，資料量非常龐大，使用者必須花工夫去瞭解各種資料庫的性質和內容，且必須使用計算量大且運算速度快的電腦，選擇適當的分析軟體，來協助找尋序列的生物意義。因此，進行序列分析有幾個關鍵的環節，首先是使用者一定要有生物學的基本知識，要知道序列的基本性質，以及確切知道要問什麼問題。然後使用者要了解資料庫的種類和內容，以及各種資料庫的序列格式，並且必須對各種分析工具有基本的認識，進而學習分析程式的操作，得到分析結果。最後，是判讀分析結果，著手進一步的實驗加以驗證。分析結果的判讀，取決於使用者對每一個環節的瞭解是否深入而完整，換句話說，使用者對於資料庫和分析工具的瞭解愈清楚，就愈能瞭解分析結果的意義。

「巨分子序列分析基礎課程」是國家衛生研究院所提供的生物資訊服務中的一個項目，這個課程的目的是介紹資料庫和分析工具的基本概念，以及介紹 GCG (The Wisconsin Package)服務中常用的分析軟體操作方法，以協助國內生命科學的研究工作。

二、國家衛生研究院生物資訊服務 (http://bioinfo.nhri.org.tw)

國家衛生研究院根據國內生命科學研究單位所需，於八十六年開始提供生物資訊服務，包括**巨分子序列分析（GCG）服務、分析工具之提供、資料庫及鏡相站建置與維護**，並提供國內相關**教育訓練課程**，此外國衛院生物資訊服務網頁設有全球重要生物資訊網站及資料庫的超連結，方便研究人員獲取最新資訊。各項服務的最終目標是協助並推廣國內生物資訊研究。國家衛生研究院生物資訊服務簡介如下：

- **巨分子序列分析（GCG）服務**：

The Wisconsin Package 套裝程式組，一般通稱為 GCG (Genetics Computer Group)，包括百餘種相關軟體程式，研究人員可用以進行 DNA 和蛋白質的編輯、比對、比較與連結，以及 RNA 二級結構之預測，DNA fragment 的重組及演化的分析，並可對全球主要基因資料庫如-- GenBank, EMBL, PIR 及 SWISS-PROT 提供資料庫搜尋及相關序列分析功能。本項服務包括 Unix 介面的 GCG Command Mode，以及網路介面的 SeqWeb 軟體，供使用直接在瀏覽器上以點選的方式進行

序列的比對及分析。本院並開辦巨分子序列分析基礎課程，協助使用者熟悉 GCG 基本程式的使用。

■ EMBOSS 服務：

EMBOSS 全名為 European Molecular Biology Open Software Suite，是免費公開的序列分析用程式組，由分子生物研究社群所共同開發，目前則由 SourceForge.net 來維護 (http://emboss.sourceforge.net/credits/)，其程式可和蛋白質、核酸資料庫相連結，能夠用來進行序列搜尋、序列比對、尋找蛋白質motif、domain、二級結構等分析工作。EMBOSS程式碼完全公開，且其核心程式基本設計為與各種開發平台相容，可供研究人員作為開發應用程式的平台，世界各不同機構也因此開發了各種EMBOSS使用介面，目前國衛院安裝的是常用的 JEMBOSS，unix command mode EMBOSS，以及Luke's EMBOSS GUI。

一、 JEMBOSS

JEMBOSS 全名為 Java-EMBOSS，即 Java 介面的 EMBOSS 程式圖形使用介面，為英國的 HGMP-RC (Human Genome Mapping Program Resource Centre) 所開發，採用 user-friendly 的 Java 視窗執行 EMBOSS 程式，並加入檔案管理功能，使用者 download Java 執行程式後，即可在自己的 PC 執行 emboss 程式並儲存分析結果，使用方便。

二、 Command Mode EMBOSS

在 Unix 系統執行 EMBOSS，對熟悉Unix指令的是用者而言，程式執行十分快速。其核心程式亦由英國的 HGMP-RC (Human Genome Mapping Program Resource Centre) 開發，目前由 Sourceforge.net 負責維護。

三、 EMBOSS GUI

由 Luke McCarthy (Plant Biotechnology Institute, National Research Council of Canada，http://bioinfo.pbi.nrc.ca/) 所開發出來的網路介面，並且持續維護更新，將常用的EMBOSS程式整理分類，撰寫簡潔的網路介面，供使用者以網路瀏覽器操作 EMBOSS程式，很受學界歡迎。

■ 資料庫及鏡相站

維護下列鏡相站之系統，並定期更新資料庫，使研究人員可於國內直接使用站中提供之資料庫與分析工具，免於取道國際網路擁塞的限制。

A. Protein Data Bank (PDB) 蛋白質結構資料庫鏡相站（與清大合作建置）

B. ExPASy 蛋白質資料庫鏡相站

■ 國內資料庫的建置與維護

A. Liver EST 肝細胞 EST 資料庫

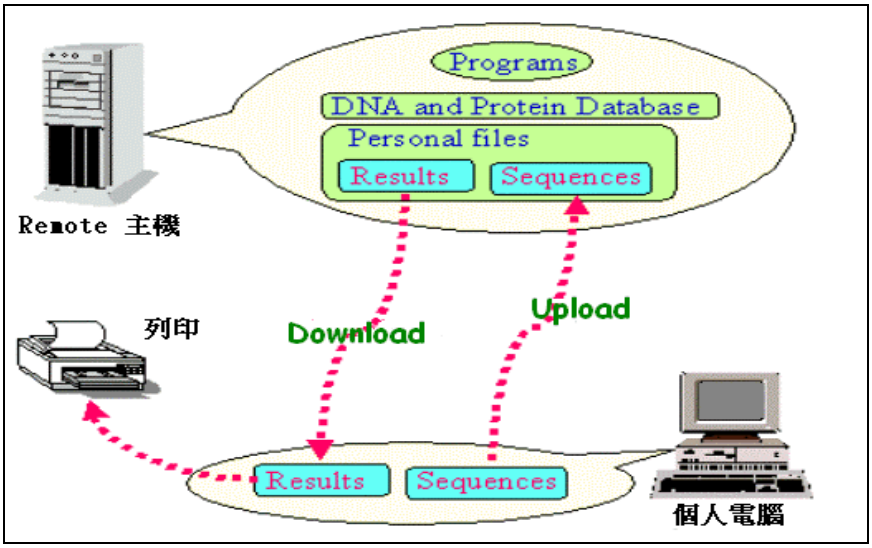
B. Taiwan Polymorphism Marker Database 「台灣多變異性標竿資料庫」

C. Bladder cancer DB 膀胱癌資料庫

貳、SeqWeb 使用說明

SeqWeb 是 GCG (Computer Genetics Group) 簡化後的網頁版本，GCG 正確的名稱應該是 Wisconsin Package，是包含 120 種以上的分析程式的程式套組，它有兩個使用介面，一個是以 Unix 指令操作的 GCG Command Mode，另一個就是以 Web 網頁介面執行程式的 SeqWeb，而 Accelrys 公司於 2005 年底將 SeqWeb 程式更新至 version 3.0 版，除了解決舊版操作上的一些問題之外，也更新部份功能及程式，而目前最新的版本則為 SeqWeb 3.1.2。

SeqWeb 3 可讓使用者利用不同的瀏覽器(如 Internet Explorer、Mozilla、FireFox、Opera 等)來執行序列分析的工作，其 user friendly 的使用者介面，對初學者來說可省卻不少學習 Unix 指令的工夫！唯 SeqWeb 僅包含了多數使用者較為常用的程式，但並未包含 Wisconsin Package 所有的程式，因此在整體功能上，SeqWeb 仍不如 GCG Command Mode 的強大與齊全。



使用SeqWeb / GCG command mode之檔案傳送方式

在使用 SeqWeb 時首先要注意的就是，所有的分析程式和資料庫是存放在 GCG 主機中，序列分析也是由 GCG 主機執行，序列分析結果也是存放在主機中，因此使用時 query 序列檔案必須先上傳至主機中，分析完成後必須把分析結果的檔案下載至個人電腦中。SeqWeb 中檔案傳送透過瀏覽器可以很容易的完成，如果在使用時隨時注意將檔案傳送到正確的地方，使用起來會更感方便。

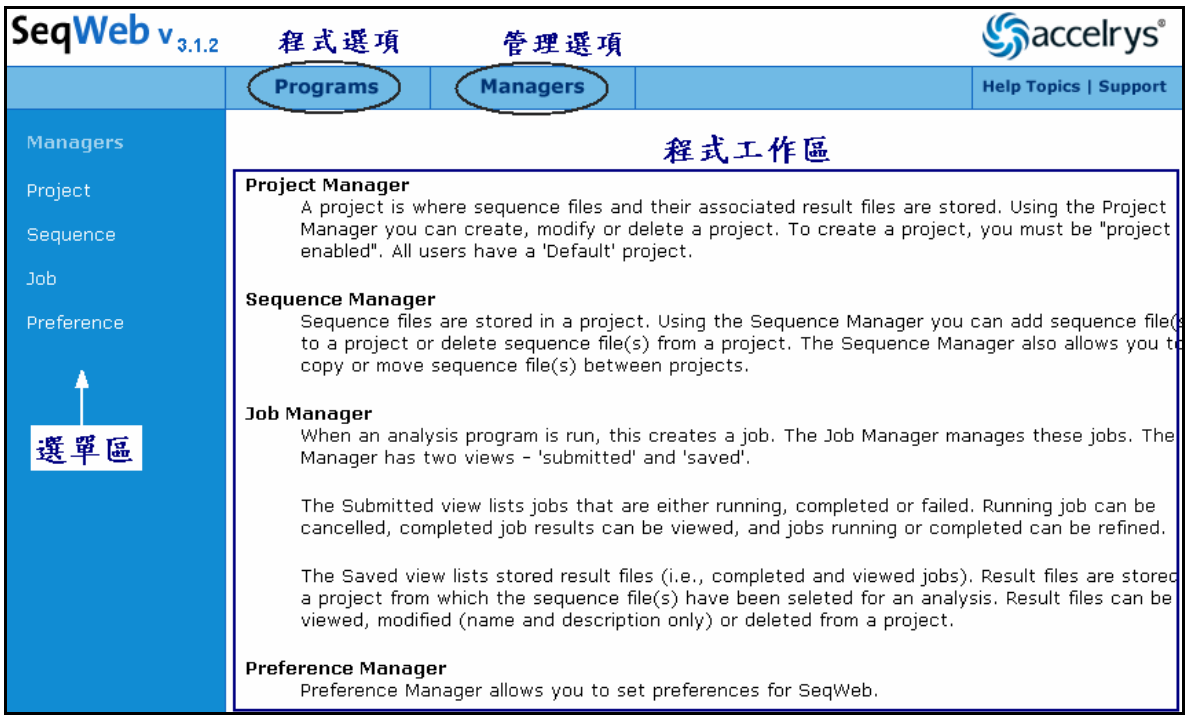
一、Login SeqWeb

SeqWeb 3 支援各種不同的瀏覽器。使用時可透過直接輸入SeqWeb網址的方法 (http://v8803.nhri.org.tw:8003/mgr.shtml) 登入使用，或是由「基因體醫學生技研發生物資訊核心(GMBD Bioinformatics Core)」(http://www.tbi.org.tw) 網站登入SeqWeb。初登入時會呈現login視窗，此時需鍵入使用者帳號及密碼方可進入此系統。

Login 畫面

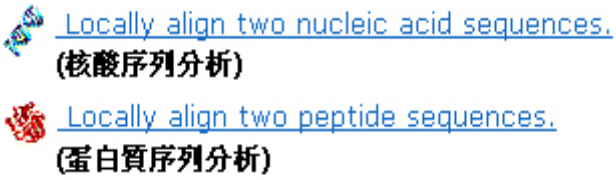
二、SeqWeb 主網頁簡介

Login 之後即可看到如下的 SeqWeb 主網頁。



■ 程式工作區

在主網頁右邊是各程式之主要工作區，當使用者由程式選項中點選了分析程式後，就會在程式工作區呈現該程式的內容及選項。網頁左方有一欄選單區，依照分析程式的功能來分類，能讓使用者很容易明白要作哪一種的分析工作。特別要注意的是：SeqWeb將分析核酸和分析蛋白質的程式做了明確的區分。若選用分析核酸的功能，當使用者進入後，屆時只能 Input核酸序列檔案，蛋白質序列檔案會被隱藏起來，反之亦然。



以左圖為例：許多程式均明確的將核酸序列及蛋白質序列區分為兩個不同連結，供使用者選用。特別需注意的是，若選擇核酸序列分析，Sequence Manager 裡就不會出現蛋白質序列，反之亦然。

■ **使用說明與諮詢**

網頁右上角，Accelrys Logo 的下方提供了“使用說明(Help Topics)”與“諮詢(Support)”兩個選項。

A. Help Topics--包括完整的 SeqWeb 使用手冊及 Data File 說明。

B. Support--有美國 Accelrys 原廠的聯絡電話與 E-mail 信箱等資訊，但建議使用者先與本院諮詢信箱(bioinfo@nhri.org.tw)聯絡，若本院系統管理人無法解決您的問題，則會代您向 Accelrys 原廠詢問解決方法。

三、Sequence Manager 功能

SeqWeb3 的 Sequence Manager 主要透過 LDAP 的方法將序列資料儲存，並透過網頁方式點選便能使用。Sequence Manager 具有序列管理和編輯之功能，當使用者將以 SeqWeb3 來進行序列分析前，所有的個人序列檔案都必須先存放在 Sequence Manager 中，再至程式選項裡選擇欲分析的序列，就能完成分析的結果。

SeqWeb3 的 Sequence Manager 共分為 “Project”、“Sequence”、“Job”、“Preference”等四部分，其中“Project”與個人工作計畫的建立編輯相關；“Sequence”與個人序列之管理、序列增減編輯功能相關；“Job”與序列分析時狀態和分析結果相關；“Preference”則與 SeqWeb 個人化設定功能相關。以下則針對這四大 Sequence Manager 的項目作詳細使用說明及介紹。

注意：在此特別提醒您：使用 SeqWeb 時，應注意隨時將所需要的檔案或分析結果下載至個人電腦中，國衛院對於存放在生物資訊主機中的檔案不負保管責任，如有任何資料或檔案遺失，使用者須自行負責。

■ **進入 Sequence Manager**

步驟：

A. 進入 SeqWeb3.1.2 版主網頁

B. 請將滑鼠指標移至網頁上方選單之 Managers 選項,此時下方會出現四項 manager 選單，請選擇 “Sequence” 進入。

C. 進入後可看到以下畫面，同時提供了一些資訊：

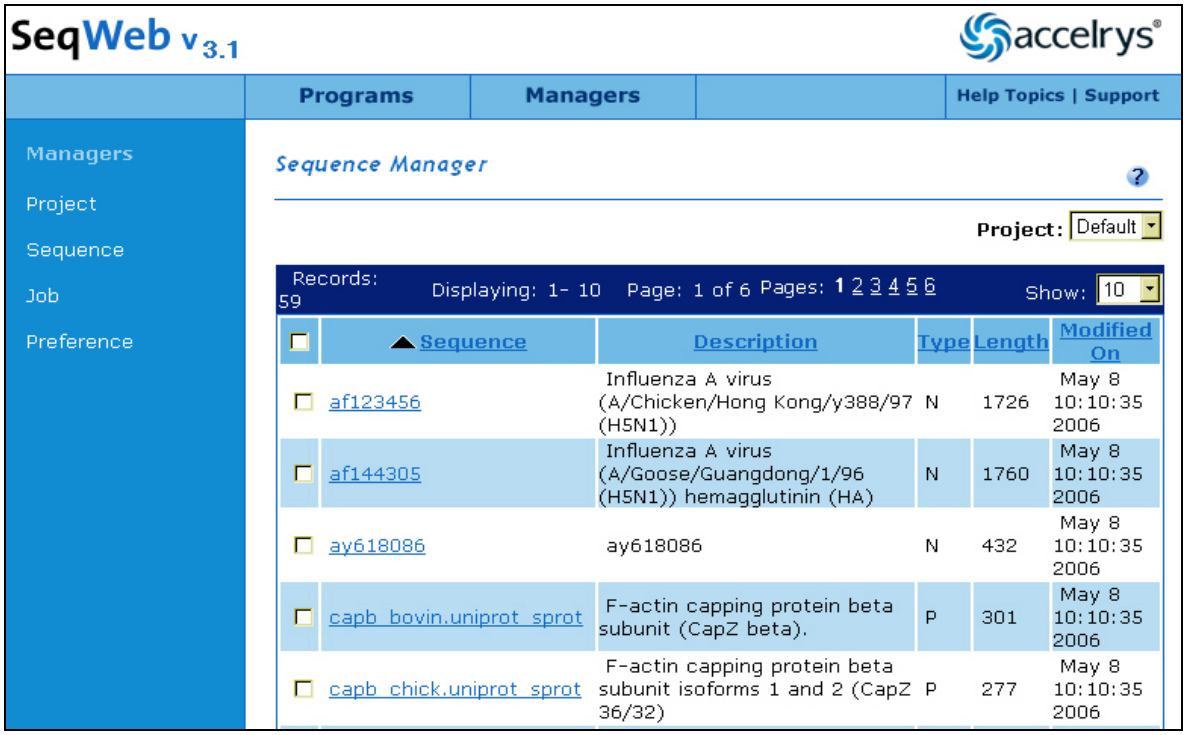
Project: Default-- 此為一下拉式選單，預設值為 Default，可選擇進入個人所建立的 project 目錄內。

Show: 10-- 此為一頁所能呈現的序列筆數，預設值為 10 筆，但建議**可調整成 100 或 200 筆**為佳。

Sequence-- 為儲存在 Sequence Manage 裡的序列名稱。

Description-- 為該序列的描述文字，可作為選擇序列時之參考用。

Type-- 序列類型，N 代表核酸序列、P 代表蛋白質序列。
Length-- 序列長度。
Modified On-- 序列編輯/存檔日期。

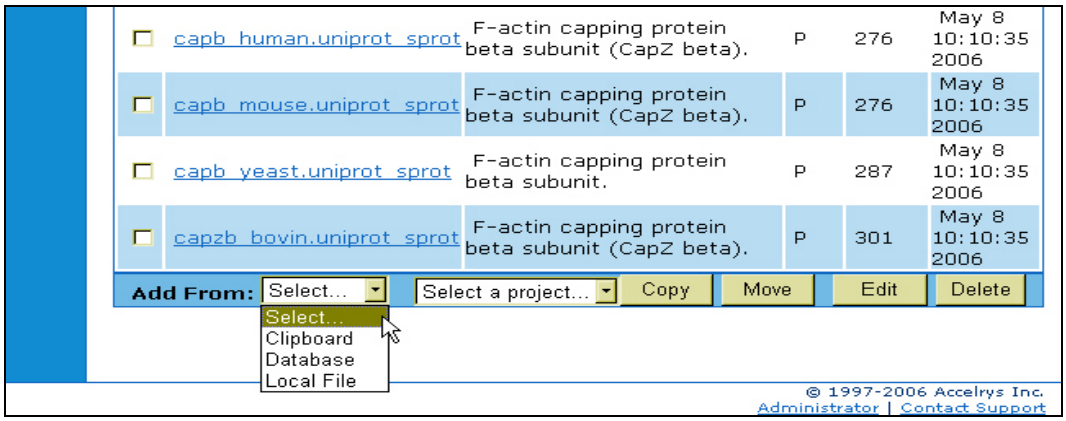


D. 在上述的選項 (Sequence, Description, Type……等)，直接點選任一項目，則序列就會以該類別重新排序。如：點選 “Length” 的項目，序列就將以 “長度” 進行升冪或降冪之重新排序。

■ 加入序列（add sequence）

當使用者已經有一段序列：來源可以是 sequencing 的結果、或是由 searching 到的結果 copy 一段而來，這些存於使用者電腦裡的序列文字檔，都可以利用這項功能加入到 SeqWeb 之中。

Sequence manager 的下方具有選單，它提供了三種不同方法可讓您將序列加入到 SeqWeb 中。加入序列檔案的方式有 Add From Local File、Add From Clipboard、及 Add From Database 三種，分別敘述如下：



A. Add From “Local File” --從 PC 中將序列檔案加入至 SeqWeb

步驟：

1. 於 **Add From** 的下拉選單中選擇 **"Local File"**，將會跳出 Add From Local File 的新視窗
2. 在 Number of Files 的下方欄位中，點"瀏覽"按鈕，就能從自己電腦的儲存媒介(硬碟、隨身碟、光碟)中，將序列檔案上傳至 SeqWeb 裡。使用者可選擇一次加入多條序列，最多可以一次上傳 20 條。
3. 加入序列檔案完成後按 OK。回到剛剛 Sequence 的視窗，並重新整理網頁，就能看到剛剛加入的檔案。

   注意：序列上傳至 Sequence Manager 後，序列名稱未必和存放於個人電腦中的檔名相同，最好是由 "Modified On" 的日期來檢查最為確定！

檔案需為 SeqWeb 所接受的檔案格式方可上載。序列檔案請先利用 NotePad 存成純文字檔 (*.txt)，如果以 Microsoft Word 儲存之序列檔案，將不被接受！。

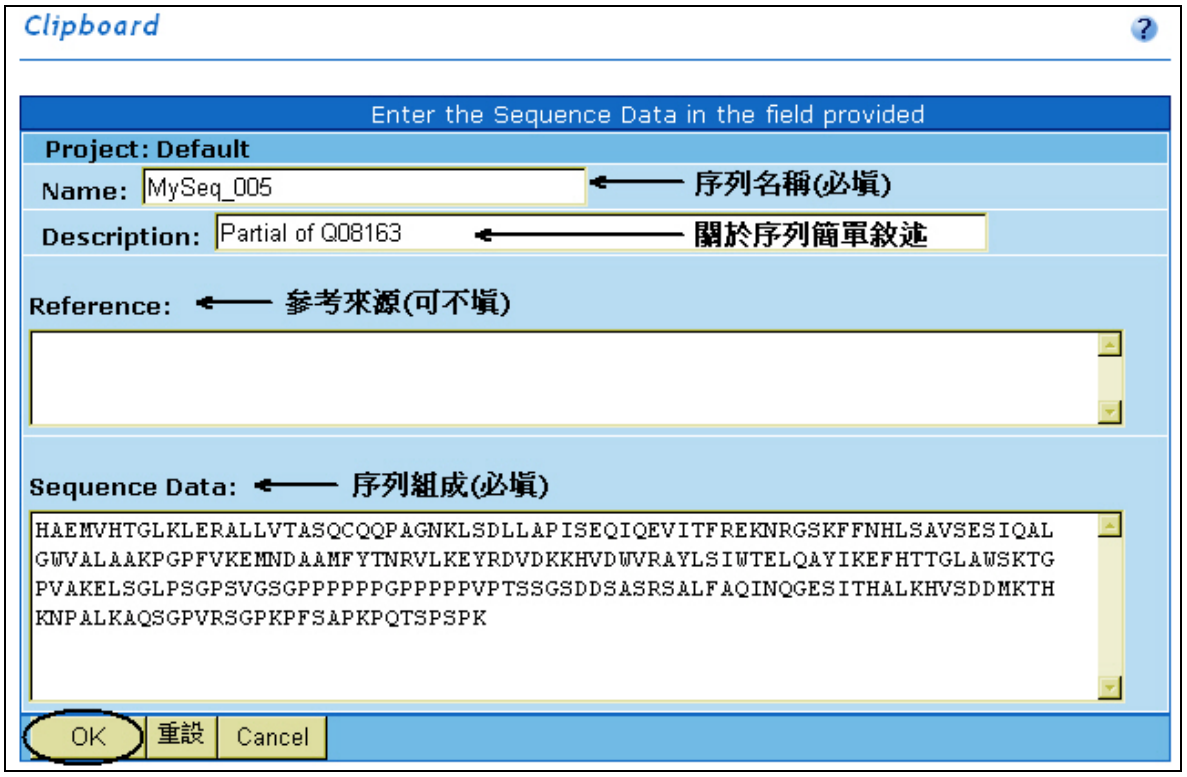
B. Add From "Clipboard" --直接鍵入 sequence，存成序列檔案

步驟：

1. 於 Add From 的下拉選單中選擇 "**Clipboard**"，將會跳出 Add From Clipboard 的新視窗
2. 新視窗內有一些欄位需輸入資訊：
   a. Name：輸入序列名稱，一定要填。
   b. Description line：填寫對序列的描述文字，可不填。
   c. Reference：填寫列來源或參考文獻，可不填。
   d. Sequence Data：序列之組成，可以直接用鍵盤 key in 序列，或是用 copy / paste 的方式輸入序列。
3. 完成後按 OK。回到剛剛 Sequence 的視窗，並重新整理網頁，就能看到剛剛加入的檔案。

請注意：蛋白質序列請以**單一字母**來表示一個胺基酸，如果是以三個字母來表示(例 Gly 或 GLY)，Sequence Manager 會誤認為是三個不同的胺基酸。
Sequence Manager 可辨識的字母及符號如下列，Sequence Manager 也可辨識空格(space)。
字母：A B C D E F G H I K L M N P Q R S T U V W X Y Z (not J or O)

(小寫) a b c d e f g h i k l m n p q r s t u v w x y z (not j or o)
符號： . (period) ~ (tilde) * (asterisk)



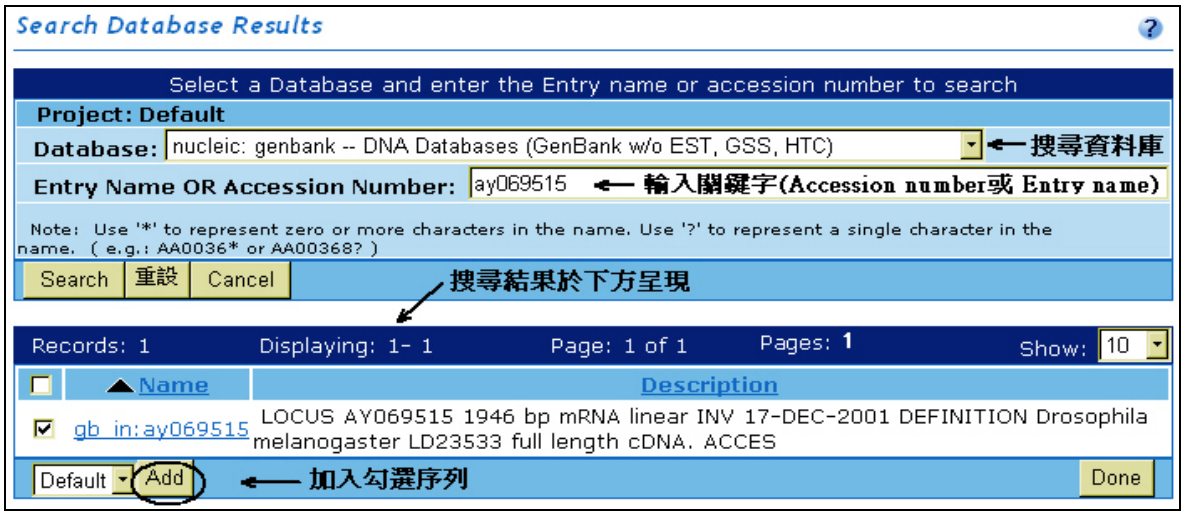
C. Add From “Database” --由資料庫中加入序列檔案

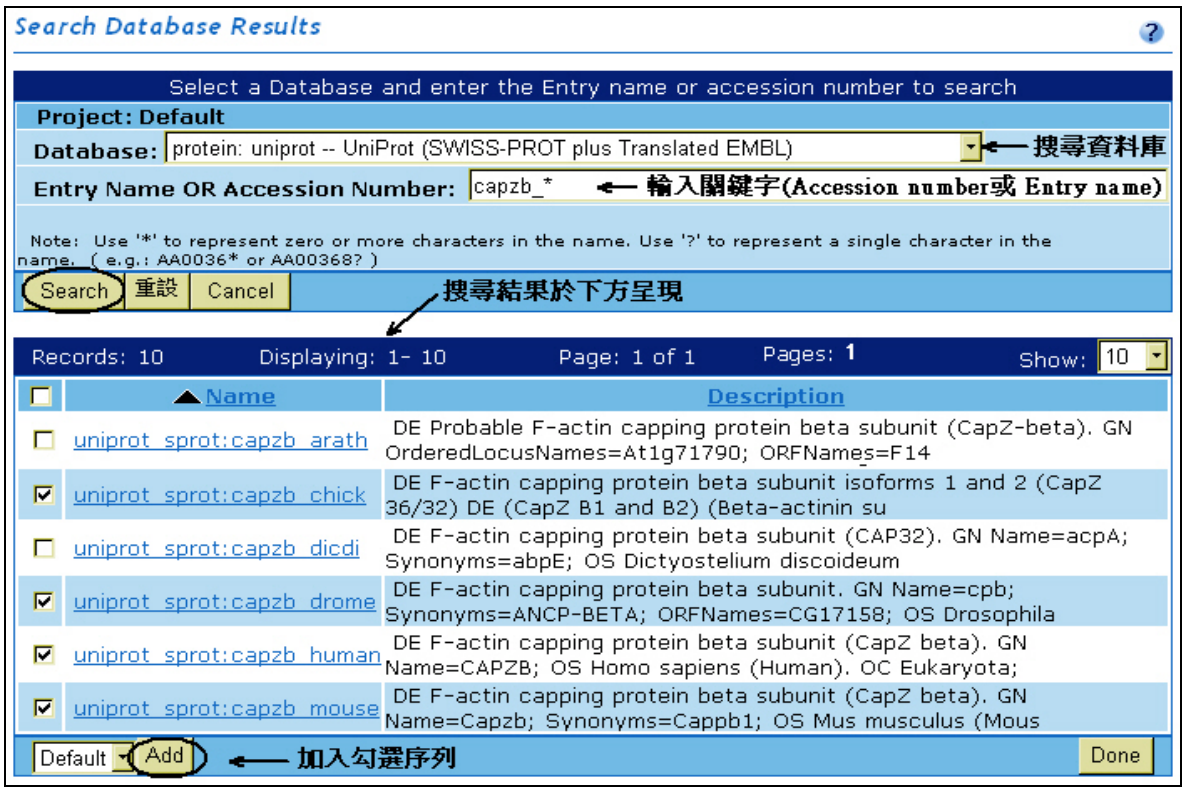
當使用者知道您需要的序列是存在於資料庫中的話，可以利用此功能先行尋找序列，找到後再將它們加入 Sequence Manager 中，以進行分析工作。

使用之前必須已經知道序列的 accession number 或 entry name，如果不知道序列的 entry name 或 accession number，它們可從 paper 中查到，或是利用 SeqWeb 的 Lookup、StringSearch 等程式來搜尋，這兩個程式的使用將在後面章節另作介紹。

步驟：

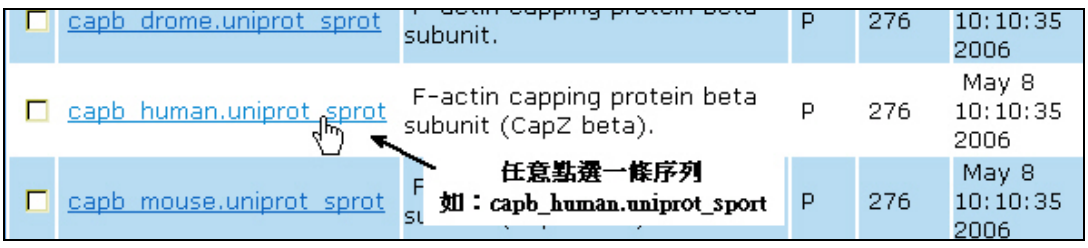
1. 於 Add From 的下拉選單中選擇 “Database”，將會跳出 Add From Database 的新視窗。
2. 於欄位中輸入序列的 entry name 或 accession number。如果不太確定 sequence name 或 accession number 也可以用萬用字元 "*" 號代替以協助搜尋，例如想找 F-actin capping protein beta subunit 相關的蛋白質，可下 “**capzb_***” 的關鍵字來尋找。
3. 按 OK 即進行搜尋。搜尋完成之後結果將直接餘下方呈現。使用者可以在小方格中打勾代表選取該序列，並將它們加入 Sequence Manager 裡。
4. 如果點選序列名稱的超連結，可以直接瀏覽序列的內容。
5. 完成後按 OK。回到剛剛 Sequence 的視窗，並重新整理網頁，就能看到剛剛加入的檔案。

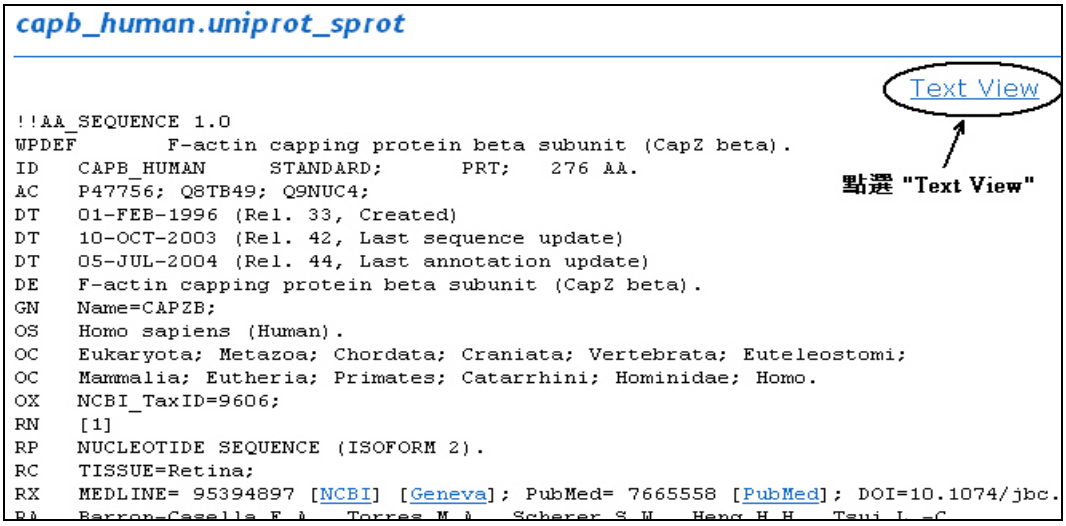


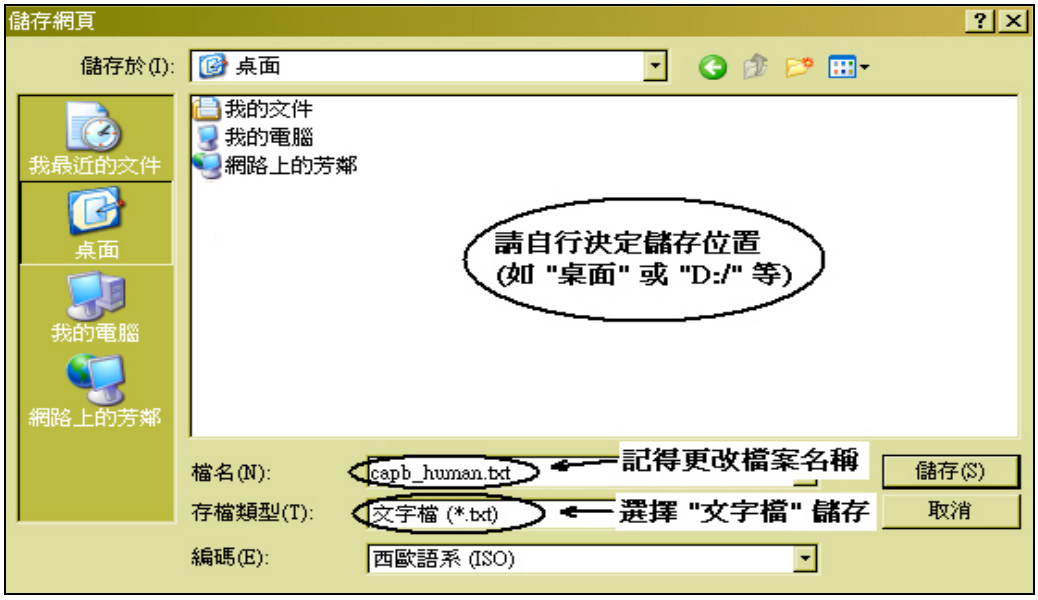


■ 存取序列檔案

SeqWeb v3 提供了比舊版 SeqWeb 2.1 較為方便容易的序列存檔方式。雖然在網頁裡並無任何和存檔相關的按鈕或選項，但是使用者能透過看序列內容的方式，並在跳出序列內容的新視窗中，選擇 “Text View” 的選項，SeqWeb 就能將該序列以 text 的格式呈現，使用者也能在該視窗中直接將序列存檔。







請注意：SeqWeb 存檔後序列為 text 的檔案，但序列內容的格式一律為 “GCG format” 因此若需要在別的生物資訊工具另作分析時，需先確認該工具是否能接受 “GCG format” 如果不行，則需要再另外作格式轉換。

■ 序列檔案管理

A. View：檢視序列內容

步驟：

1. 在 Sequence Manager 網頁裡直接點選序列檔案連結即可。

B. Copying sequence: 複製序列檔案

步驟：

1. Sequence Manager 網頁裡勾選任一條序列檔案
2. 將網頁拉到最下方，選擇要存放複製檔的 project，並按下 Copy 按鈕
3. 此時會 Pop-up 跳出一個指示碼提示的小視窗，並要求你輸入新檔名
4. 輸入複製序列新的檔名後，按下 “確定” 即複製檔案完成

C. Moving / Renamin sequence:移動序列檔案或更改檔名

步驟：

1. Sequence Manager 網頁裡勾選任一條序列檔案
2. 將網頁拉到最下方，選擇要另存的 project，並按下 Move 按鈕
3. 此時會 Pop-up 跳出一個指示碼提示的小視窗，並要求你輸入新檔名
4. 輸入序列新的檔名後，按下“確定”即完成移動或改名的工作

D. Deleting sequence: 刪除檔案

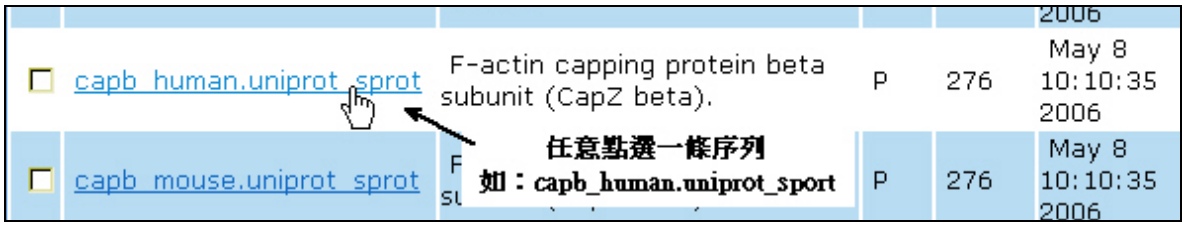
步驟：

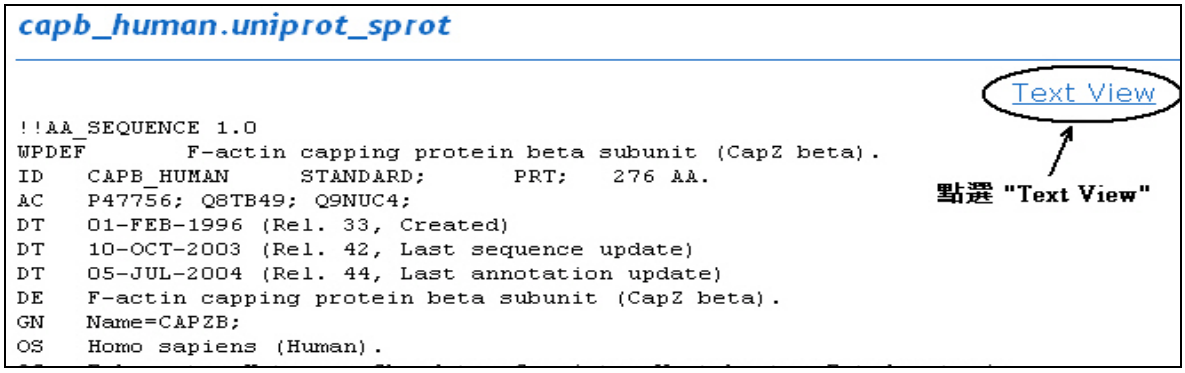
1. Sequence Manager 網頁裡勾選任一條序列檔案
2. 將網頁拉到最下方，按下 Delete 按鈕
3. 在跳出一個確認視窗後，按 OK 即刪除完成

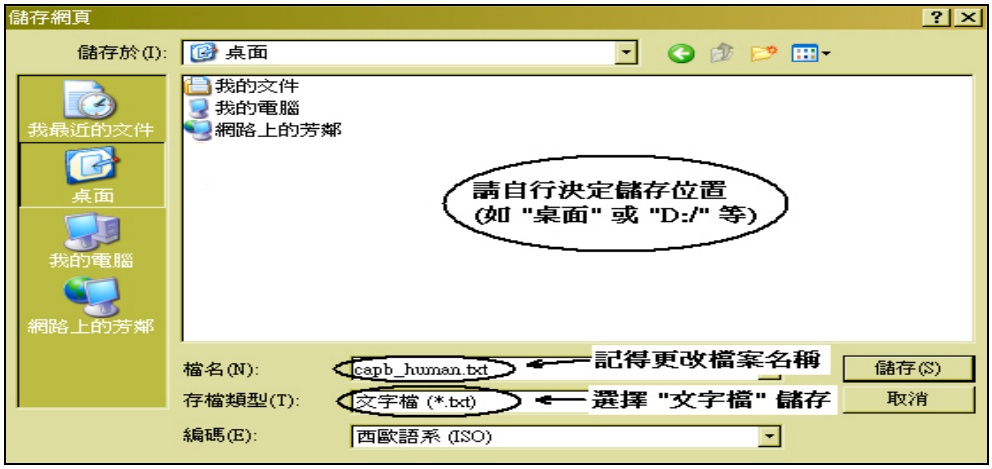
E. Saving sequences: 存檔，將檔案存在個人電腦中

步驟：

1. 在 Sequence Manager 網頁裡直接點選任一條序列檔案
2. 新跳出的序列視窗中，左上方有一“Text View”的連結，點入後序列會以純文字的模式呈現於視窗中
3. 在視窗中選擇“檔案”→“另存新檔”的功能，選擇要儲存的位置及輸入新檔名，並選擇文字檔之存檔類型儲存即可







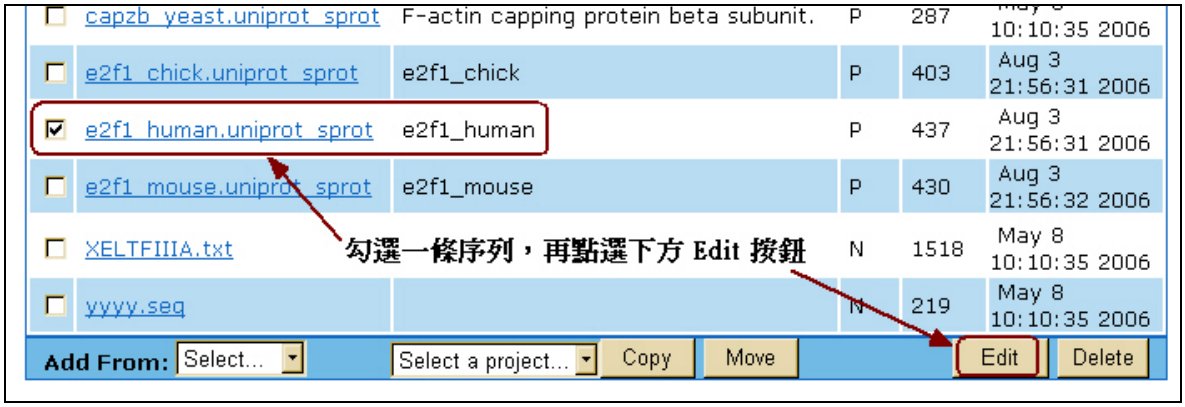
■ 序列編輯

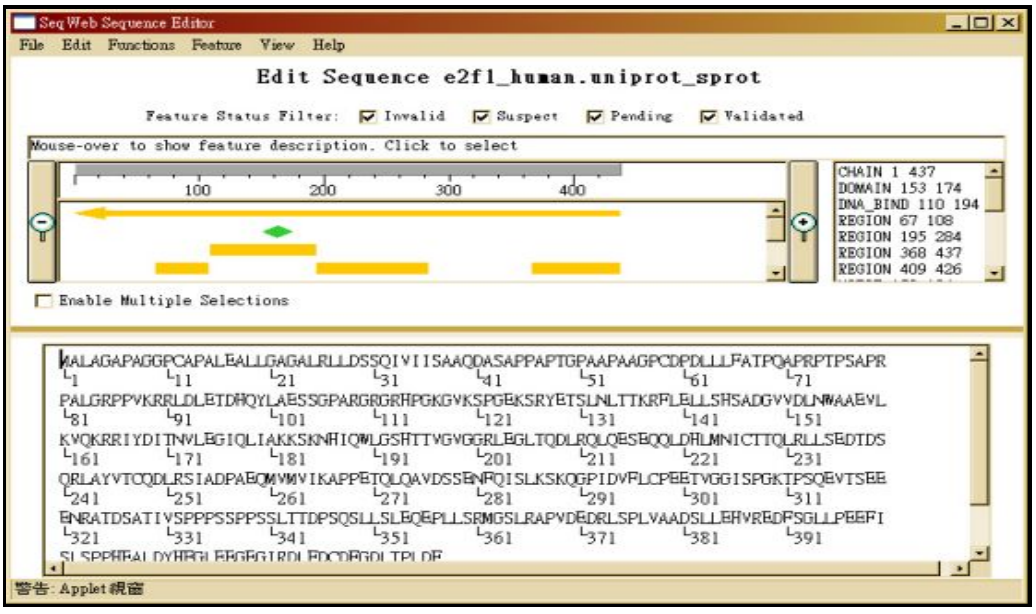
序列編輯主要透過 Sequence Editor 的功能，幫助使用者對序列進行一些序列的特徵(features)或文字描述(description)的編輯，也可對序列本身作些簡單處理，如 Cut、Paste、Translate、Assemble、Reverse Complement 等，並且可以將序列存成圖形檔使用。

A. 開啟 Sequence Editor 功能

步驟：

1. 在 Sequence Manager 網頁裡勾選任一條序列檔案，並點選最下方的 "**Edit**" 按鈕，就會出現 SeqWeb Sequence Editor 視窗





B. Edit Sequence 視窗

1. 視窗分為上下兩部份，上面的框架是序列的圖示區，並具有放大鏡功能，下方的框架則顯示列出全部序列字元(characters)及其說明(comment)；序列字元(characters)和說明(comment)的顯示可使用 View 功能做切換。中間則有 Enable multiple selection、Feature，及 ORF 選項。上下兩個框架的內容是相關聯的，在下方框架選取的序列會即時顯示於上框架對應的圖形中。SeqWeb 在序列的圖示上有美工編輯功能供使用者選擇，對於呈現分析結果的視覺效果有所幫助。

2. 編輯過程中可以從 “View” 選項的 “Font size” 調整字型大小，並且可以隨時使用 “Edit” 選項的 “Undo” 或直接按 **Ctrl+Z** 來回復檔案；此外也可使用 Disable Edit 取消編輯功能來保護序列檔案，或用 Enable Edit 恢復編輯功能。

3. 編輯過程中或編輯結束後皆須存檔，Sequence Editor 有三種存檔方式
   1)把編輯結果存下來，檔案名稱不變：點選 **File** 按 **Save**。
   2)把編輯結果存另存新檔：點選 **File** 按 **Save as**，選擇存檔的 project，輸入新檔名，按 OK。

C. 使用 Sequence Editor 編輯序列

1. Editing a Description：編輯序列文字敘述的內容

步驟：點選**Edit**，按**Edit Description**，出現edit description視窗，輸入description後按OK。

2. Navigating Within Sequences：在序列檔案中瀏覽及搜尋序列片段

瀏覽指定位置(location)的序列

步驟：

i 選擇要編輯的序列，進入 Sequence Editor
ii 點選 Edit，按 Go To，出現 Go 視窗
iii 輸入指定位置的序列 residue 編碼 (例：23) 後按 OK，游標即移至指定的位置 (直接跳至第 23 個 nucleotide 或 amino acid)

3. 搜尋特定序列片段

步驟：

i 選擇要編輯的序列，進入 Sequence Editor
ii 點選 Edit，按 Find，出現 Find 視窗
iii 輸入想要尋找的序列片段 (例：ggatta) 後按 OK。
iv 如果找到相符的序列片段，游標即移至找到的序列最後一個 residue 位置
v 如果沒找到相符的序列片段，即出現 Match not found 視窗，按 OK 結束。

4. Selecting a Range：選取編輯範圍

步驟：

用滑鼠把想要編輯的範圍框起來即可。但是當需要選擇精確的位置時，利用 “**Edit**” ；選項之 “**Select Range**”，輸入編輯範圍的開始及結束的序列編碼 (例 Begin:513, End:2088)，然後按 apply。被選定的序列範圍 Sequence Editor 會以藍色 highlight 起來。

5. Cutting, Copying or Pasting a Range 序列的剪接與剪貼

步驟：點選 **Edit**，按 **Cut**，或按 **Copy**，或按 **Paste** 即可。

6. Performing Functions Specific to Nucleic Acid Sequences：核酸序列的組合、轉譯與互補序列轉換

步驟：

i. 選擇要編輯的**核酸序列**
ii. 在 Edit Sequence 畫面中間勾選 Enable multiple selections
iii. 選取一個或多個序列範圍
iv. 點選 Edit 按 Functions，視需要執行下列功能：
--按 **Assemble** 組合所有選取的範圍成為一個新的序列，

--按 **Translate** 將 nucleotide 序列轉譯為 amino acid 序列，

--按 **Reverse Complement** 將序列中 A-T, C-G 互換得到互補序列，

v. 在各個功能相對應的視窗輸入檔名、description(可省略)後，按 Add to Project 將新的序列存檔。

vi. 或按 Cancel 取消。

7. Working with Sequence Features 序列特徵的美工編輯

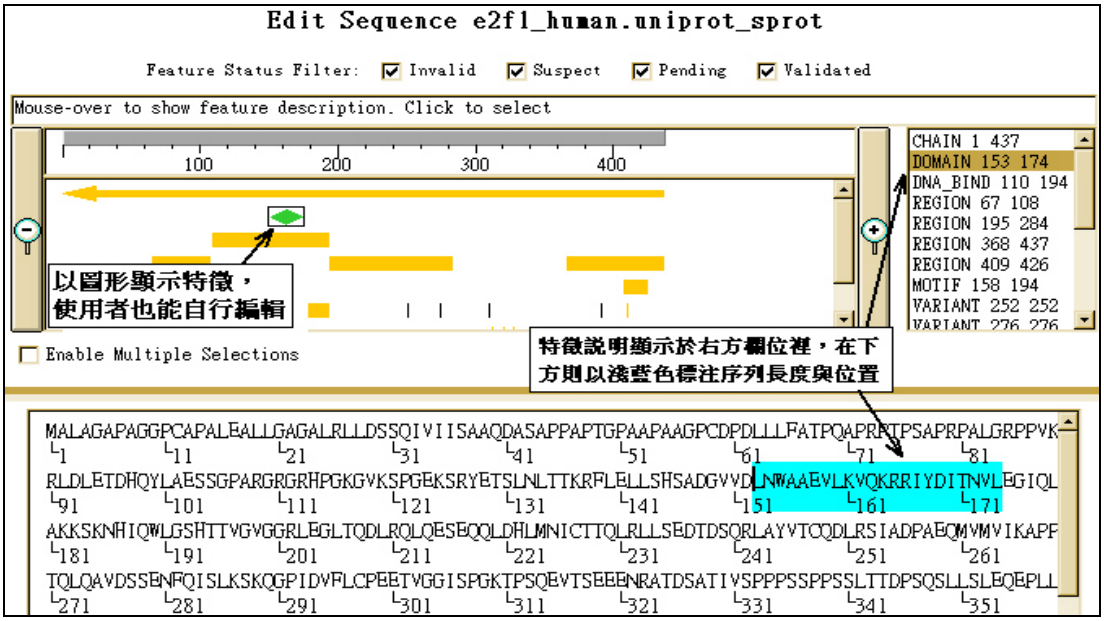
步驟：

i. 加入序列特徵：點選 Feature 再選 Add Feature， 選擇 shape,color 以及說明文字(如 enhancer, TATA Box 等，可省略)，按 Save。此時要 refresh 畫面才會看到加入的 feature。方法是用滑鼠在序列框頁空白處點一下，即可在圖示區看到加入的 feature。

ii. 刪除序列特徵：在圖示區已加入的 feature 圖形上點一下（此時圖的四周出現框線），點選 Feature 再選 Delete Feature。

iii. 編輯序列特徵：在圖示區已加入的 feature 圖形上點一下（此時圖的四周出現框線），點選 Feature 再選 Edit Feature，接著選擇 shape, color 以及說明文字(如 enhancer, TATA Box 等，可省略)，按 Save。同步驟 1 之 refresh 方法，即可在圖示區看到改編後的 feature。

iv. 檢視 ORFs：此功能限於核酸序列。只要在 Edit Sequence 頁面中間勾選 ORF 即可切換至 ORF 的圖示頁面。使用者可以用左、右箭號選 Cutoff 值，按 Set Cutoff 後就只有高於 cutoff 值的 ORF 會顯示在畫面上。ORF 頁面也有放大鏡功能。



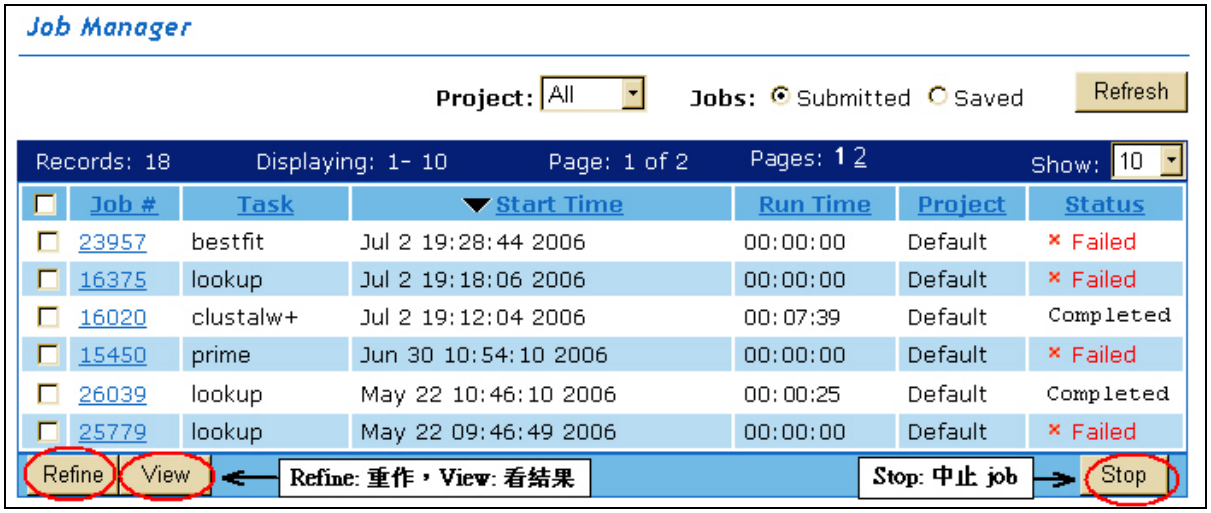
四、Job and Result Manager

■ Job Manager :

Job Manager 能夠記錄使用者曾經用過的程式，並列出使用者的程式運算狀況(工作已完成或是運算中)。透過 Job Manager 中的紀錄，可以察看之前曾經分析的結果，也可以選擇重做工作。

當點選 Manager 中 Job 選項時，視窗的右方欄位，顯示了使用者曾經送出分析的

程式運作的紀錄，使用者可以勾選任一紀錄，選擇 “Refine” 重作或是 “View” 觀看結果。如果送出的 Job 跑很久當掉了，也可以選擇 “Stop” 來中止工作項目。



A. 重新分析工作

當使用者發現序列分析結果出錯、或想以其他的參數再分析一次時，可以利用 Job Manager 中的 Refine 按鈕將結果重新分析一次。若是之前分析的工作已不存在於 Job Manager 之中時，則需要到 “Saved” 選項，才能檢視之前的結果。

B. 中斷正在運算的工作

當程式仍未完成時(顯示 Running 狀態)，若按下右方的 Stop 按鈕，則程式將強制被中斷，工作內容也將由 Job Manager 中移除。

C. 察看完成的工作

當 Job Manager 顯示 Complete 的狀態時，表示工作已完成，此時點選下方的 View 按鈕，將開啟新視窗並展現結果。

D. 回顧已分析完的成果

使用者若需要查看以前曾經作過的分析結果，則先點選 **“Saved”** 之選項，此時視窗將會帶出以前分析過的紀錄。若要看詳細內容，只需直接點選 “File” 之文字連結，就會得到之前的結果。

同樣在 **“Saved”** 選項畫面，下方有數個按鈕可供操作：

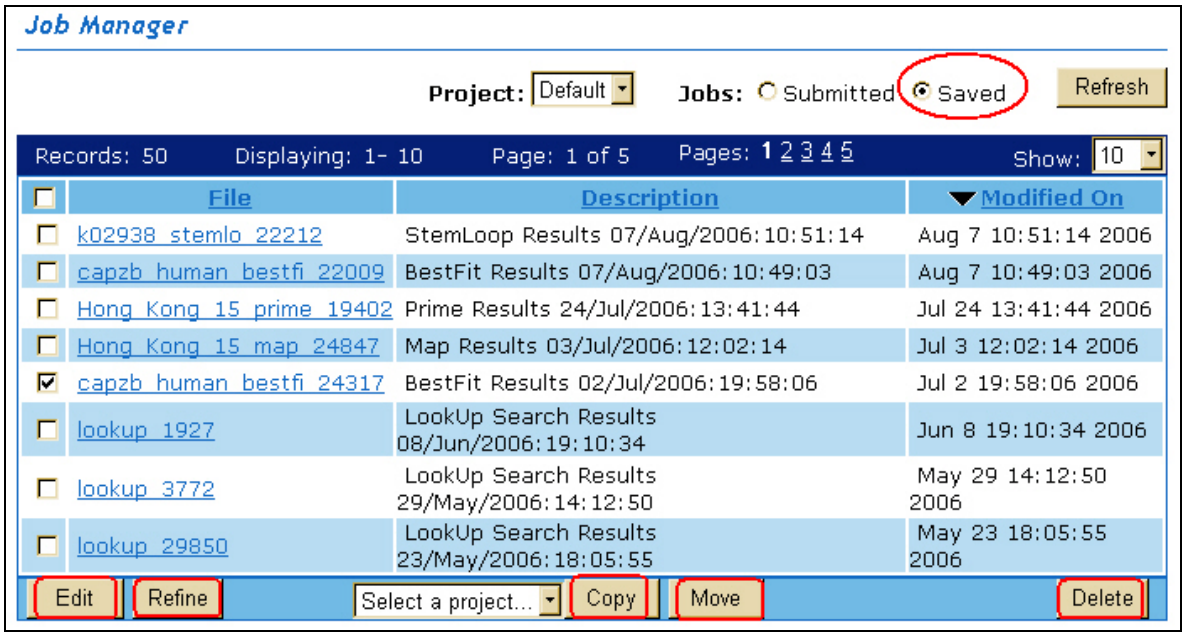
① Edit— 可自行編輯紀錄的 file 名稱及文字說明。

② Refine— 可重新再做一次以前的分析。

③ Copy— 將紀錄複製一份到其他的 project 資料夾中。

④ Move— 將紀錄移動或更名到其他的 project 資料夾中。

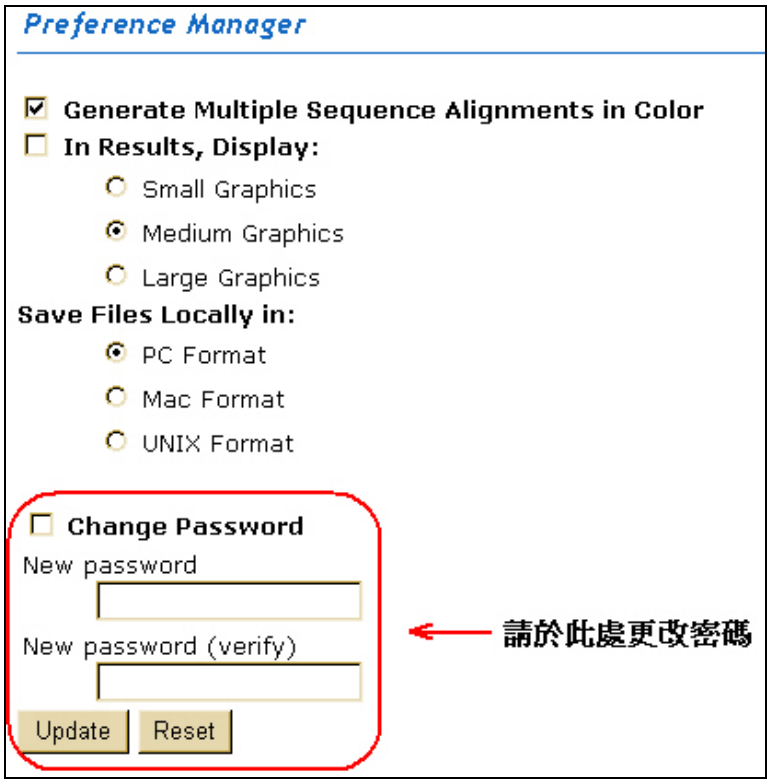
⑤ Delete— 刪除過去的分析紀錄。



五、Preference Manager

Preference 可以讓使用者對於 SeqWeb 的使用介面作一些設定。主要包括了網頁白色背景的選擇、多序列並列分析時圖形顏色的產生、背景工作完成後的郵件發送通知三項選擇，使用者可由滑鼠勾選選項。此外尚可調整圖形視窗的大小，與存檔的格式。較需注意的是，若是麥金塔電腦(mac)、或是 Uuix 系統 (如 Linux 等)的使用者，必須以非 PC 的格式儲存檔案，因此需要至 Preference Manager 中調整。

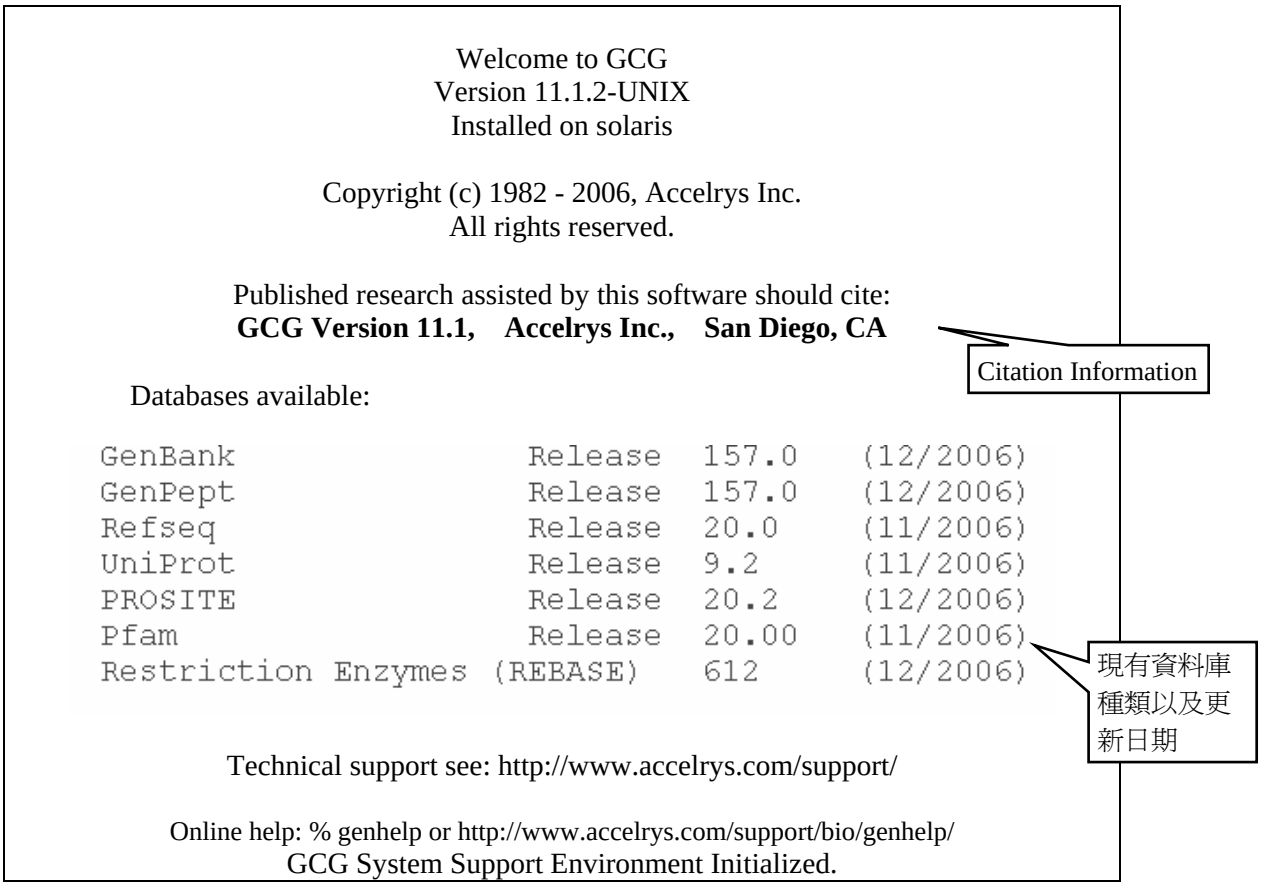
此外，關於使用者於 SeqWeb 中需要**改變或替換密碼**時，也需由 Preference Manager 中進入，輸入新的密碼，並於下方欄位做確認後，下一次再進入 SeqWeb 中，就必須以新密碼登入才行！



參、系統、資料庫及程式簡介

一、系統及資料庫簡介

SeqWeb 與 GCG 的核心程式均為 Wisconsin Package，因此每當程式更新暨資料庫升級時，兩者均獲得最新的資訊。在 Command mode GCG 中登入後就會出現一明確的列表，說明關於 Wisconsin Package 及資料庫版本的說明，以 96 年 3 月所安裝的最新版本而言，將列出以下的資訊：



由上方資訊所見，可知目前使用的 GCG Command mode 為 11.1 版。每隔一段時間，Accelrys 原廠可能對 GCG 及 SeqWeb 進行版本更新，除了修正程式外，也會增減部分功能。新舊不同版本的 Wisconsin Package (GCG / SeqWeb)內含的程式或使用方法可能略有不同，可參考英文線上說明。

序列資料庫部份則是定期作更新，其中 Wisconsin Package 的核酸序列資料庫是以 GenBank 為主。在蛋白質資料庫中則是安裝了 UniProt、PIR 和 GenPept。其中 GenPept 是利用運電腦運算，將 GenBank 的核酸序列轉譯 (translation)成為蛋白質序列，並非完全都是真正存在之蛋白質。Prosite 及 Pfam 是蛋白質 profile 資料庫，可用作預測蛋白質二級結構。REBASE 則是 restriction enzyme 資料庫。

二、GCG 核酸及蛋白質資料庫

在 GCG 中，最主要的資料庫，分別為核酸資料庫 GenBank，及蛋白質資料庫

Uniprot。其中核酸是以 NCBI 之 GenBank 資料為主，其中亦包含短序列資料庫 dbEST, dbSTS, dbGSS 及 Genome project 相關的 dbHTG 資料庫，可謂相當完整。

■ Nucleic Acid Databases

GenBank 中核酸序列之分類，主要分為兩大類：一類是以物種來分類，稱為 Organism Divisions，共有十二類，包括了 BCT (Bacterial)、PRI (Primate)、ROD (Rodent)、MAM (Other mammalian)、VRT (Other vertebrate)、INV (Invertebrate)、PLN (Plant and Fungal)、VRL (Virus)、PHG (Phage)、RNA (Structural RNA sequences)、SYN (Synthetic and chimeric sequences)、UNA (Unannotated sequences)；另一類是以功能作分類，稱為 Functional Divisions，包括了 EST、GSS、STS、HTG、Pattern 等序列。以下針對功能性分類部份簡介：

1. EST (Expressed Sequence Tags)
   EST 序列是指由 cDNA library 的每個 clone 的兩端分別進行一次定序的約 500-800bp 的序列。EST 序列因為是由 cDNA 而來的，所以在進行 Gene Finding 時是相當重要的參考資料庫，但因為僅進行一次定序，所以也包含了很多的錯誤，此外，因為重覆性很高，NCBI 另外將 EST 進行整理，將可能屬於同一個基因的 EST 序列合為一個 cluster，就是 UniGene Database，不過在 GCG 中並不包含 UniGene，但是到 NCBI 的網站或 FTP 站即可查詢或下載 UniGene 的資料。
2. STS (Sequence Tagged Sites)
   STS Database 也是一些短的序列所組成，主要收集一些在 Genome 中確定位置的序列，最重要的用途是用在基因體計畫中各個 Bac Clone 的排列時的 Marker 之用，所以通常每個 STS 都會有配合的 PCR primer。
3. GSS (Genome Survey Sequences)
   是在 Genome 中的一些除了 EST 及 STS 之外的短序列，包括了：
   “BAC/YAC end sequence”, “Exon trapped genomic sequences” 和 “Alu PCR sequemces” 等幾種序列類型。
4. HTG (High Throughput Genomic Sequences)
   收集的是各個 Genome Project 中尚未完成(Finished)的序列。各個基因體中心在進行定序時，必須在序列組合後 24 小時之內即放入 GenBank 中，但這些 Phase 0 至 Phase 3 的序列，大多含有許多的 Gap 及定序的錯誤，為了與 GenBank 一般的序列做區分，所以會將之先放置在 HTG 中。
   在 Phase 3 階段，待錯誤少於百萬分之一，並且做過適當註解之後，就會移入 GenBank 的 Organism Division 中了，例如 Human Genome Project 的序列會由 HTG 移至 Primate；Mouse Genome 的序列則會移至 Rodent。HTG 的序列雖然含有許多錯誤，但若是想早一步查詢 Genome Project 的最新序列，還是必須以 HTG 為主。

> 註：Genome Project 中各個 Phase 的定義:
>
> Phase 0 -- sequences are **single-few pass reads** of a single clone (not contigs usually).
> Phase 1 -- sequences are **unfinished**, **unordered**, and **contain gaps**.
> Phase 2 -- sequences are **unfinished**, **ordered**, and can **contain** one or more **gaps**.
> Phase 3 -- sequences are **high quality finished** sequences which **do not contain gaps**

■ Protein Databases

蛋白質資料庫是由數個單位自行建立收集的，例如 Swiss-Prot 是歐洲 EBI 組織下的 SIB 所建立的蛋白質資料庫，包含了相當完整的蛋白質序列資訊；而 GenPept 則是 NCBI 以 GenBank 核酸序列為基礎，經由電腦進行 translate 後所建立的蛋白質資料庫；PIR 則是美國 PIR 組織所建立的蛋白質序列資料庫。雖然蛋白質序列資料庫有好幾個，且各有特色，但是彼此並不互通，不同的資料庫必須配合不同的 Accession number 才能查到所需的序列，這是在使用時必須格外注意的！

此外，自 2002 年底起，EBI 整合了 SIB 及 PIR，成立了新的組織稱為 UniProt，也一併將 SwissProt 和 PIR 兩個蛋白質序列資料庫進行整合，所以當利用 UniProt 的搜尋介面，可以同時查到二大資料庫之全部序列，這是一大利多！而 Wisconsin Package 的蛋白質資料庫也自 2004 年採用 UniProt，因此使用者欲分析蛋白質序列資料，將不會再為了不知應選擇何種蛋白資料庫較合適而感到困惑，但使用者同時需注意的是：當要擷取蛋白質序列時，也因此須改選擇 uniprot 資料庫才行。

GenBank 的資料中雖然常會有 CDS 的註解，甚至將所 translate 的蛋白質序列都列出來，但若想使用蛋白質序列來進行分析時，就必須再以 GenBank 中的 Cross reference 去找到相對應的 GenPept accession number 才行，若是覺得麻煩，直接將註解中的蛋白質序列以複製/貼上的方式在 SeqWeb 中另存新檔也可以。

Wisconsin Package 中對各資料庫再加以分類而成的子資料庫，使用者在做序列比對或搜尋時，請儘量指定某子資料庫來進行，這樣不但可以加速程式運算的時間，也可以免去得到不需要的序列的結果。至於指定子資料庫的方式，在 BLAST 中有一些選項可選，在 StringSearch 及 FastA 中，就必須依照上表自行鍵入。

這些資料量十分龐大的資料庫，都是直接連同 Wisconsin Package 程式一起安裝在 Server 中的。現在所有的資料庫每二至三個月會更新一次，因此若是要在 Server 中查詢近兩個月的 GenBank 核酸序列，可能會找不到，此時建議使用者不妨去 NBCI 查詢，並將查到的序列另存新檔於個人電腦中，在上傳到主機，如此便能以 SeqWeb 進行分析。

對使用者而言，這個十分龐大的資料庫是 Wisconsin Package 程式與其他個人電腦所使用的序列分析程式最大的不同之處，如果不使用它而又想進行序列搜尋或取得序列資料，建議直接使用美國 NCBI (National Center for Biotechnology Information) 所提供的 Entrez (http://www.ncbi.nlm.nih.gov/Entrez/)，可以查到最新的序列資料。

■ Database Reviews

GenBank:
Benson DA, Boguski MS, Lipman DJ, Ostell J and Francis BF (1998). GenBank. *Nucleic Acids Research.* **26**:1-7.

EMBL:
Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M and Sterk P (1998). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*. **26**:8-15.

DDBJ:
Tateno Y, Fukami-Kobayashi K, Miyazaki S, Sugawara H and Gojobori T (1998). DNA

Data Bank of Japan at work on genome sequence data. *Nucleic Acids Research.* **26**:16-20.

PIR:
Barker WC, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh L-SL, Ledley RS, Mewes H-W, Pfeiffer F and Tsugita A. (1998). The PIR-International Protein Sequence Database. *Nucleic Acids Research.* **26**:27-32.

UniProt:
Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004). UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Research.* **32**:D115-9

Bairoch A and Apweiler R (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Research*. **26**:38-42.

Rashbass J. (1995). Online Mendelian Inheritance in Man (Editorial). *Trends in Genetics*. **11**:291.

Brenner SE (1995). BLAST, Blitz, BLOCKS and BEAUTY: sequence comparison on the Net. (Editorial) *Trends in Genetics*. **11**:330-331.

三、SeqWeb 程式簡介

SeqWeb 中的程式其實僅包含了 GCG Wisconsin package 中最重要的一些程式。而在 3.1 版中又新增了一些程式進來。以下將 SeqWeb3.1 的所有程式列於下方，並將新增者以“星號 (*)”標示。

■ **詳細介紹如下：**

Comparison	
BestFit	BestFit makes an optimal alignment of the best segment of similarity between two sequences.
* ClustalW+	Creates a multiple alignment by progressively adding sequences to an alignment.
Compare	Compare compares two protein or nucleic acid sequences and creates a file of the points of similarity between them for plotting with DotPlot.
FrameAlign	FrameAlign creates an optimal alignment of the best segment of similarity (local alignment) between a protein sequence and the codons in all possible reading frames on a single strand of a nucleotide sequence.
Gap	Gap uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps.
PileUp	PileUp creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments.
PlotSimilarity	PlotSimilarity plots the running average of the similarity among the sequences in a multiple sequence alignment.
Pretty	Pretty displays multiple sequence alignments and calculates a consensus sequence.
Database Searching	
Similarity Searching	
BLAST	BLAST searches one or more nucleic acid or protein databases for sequences similar to one or more query sequences of any type.
FastA	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).
FrameSearch	FrameSearch searches a group of protein sequences for similarity to one or more nucleotide query sequences, or searches a group of nucleotide sequences for similarity to one or more protein query sequences.
MotifSearch	MotifSearch uses a set of profiles search a database for new sequences similar to the original family,
*MotifSearchFrom	Searches a database using a set of MEME profiles. You must first run

Meme	MEME to create the profiles. You run MotifSearch from the MEME result page.
NetBLAST	NetBLAST can search only databases maintained at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA.
ProfileSearch	ProfileSearch uses a profile as a query to search the database for new sequences with similarity to the group.
SSearch	SSearch does a rigorous Smith-Waterman search for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein)
Reference Searching	
LookUp	LookUp identifies sequence database entries by name, accession number, author, organism, keyword, title, reference, feature, definition, length, or date.
StringSearch	StringSearch identifies sequences by searching for character patterns such as "globin" or "human" in the sequence documentation.
Evolution	
GrowTree	GrowTree creates a phylogenetic tree from a distance matrix created by Distances using either the UPGMA or neighbor-joining method.
Mapping	
Map	Map maps a DNA sequence and displays both strands of the mapped sequence with restriction enzyme cut points above the sequence and protein translations below.
MapPlot	MapPlot displays restriction sites graphically. If you don't have a plotter, MapPlot can write a text file that approximates the graph.
Pattern Recognition	
CodonPreference	CodonPreference is a frame-specific gene finder that tries to recognize protein coding sequences by virtue of the similarity of their codon usage to a codon frequency table or by the bias of their composition (usually GC) in the third position of each codon.
FindPatterns	FindPatterns identifies sequences that contain short patterns like GAATTC or YRYRYRYR
Frames	Frames shows open reading frames for the six translation frames of a DNA sequence.
MEME	MEME finds conserved motifs in a group of unaligned sequences.
Motifs	Motifs looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns.
ProfileScan	ProfileScan uses a database of profiles to find structural and sequence motifs in protein sequences.
Primer Selection	

Prime	Prime selects oligonucleotide primers for a template DNA sequence.
Protein Analysis	
CoilScan	CoilScan locates coiled-coil segments in protein sequences.
HelicalWheel	HelicalWheel plots a peptide sequence as a helical wheel to help you recognize amphiphilic regions.
HmmerPfam	HmmerPfam compares one or more sequences to a database of profile hidden Markov models, such as the Pfam library, in order to identify known domains within the sequences.
HTHScan	HTHScan scans protein sequences for the presence of helix-turn-helix motifs, indicative of sequence-specific DNA-binding structures often associated with gene regulation.
Isoelectric	Isoelectric plots the charge as a function of pH for any peptide sequence.
Moment	Moment makes a contour plot of the helical hydrophobic moment of a peptide sequence.
PepPlot	PepPlot plots measures of protein secondary structure and hydrophobicity in parallel panels of the same plot.
PeptideSort	PeptideSort shows the peptide fragments from a digest of an amino acid sequence.
PeptideStructure	PeptideStructure makes secondary structure predictions for a peptide sequence.
SPScan	SPScan scans protein sequences for the presence of secretory signal peptides (SPs).
* TransMem	Scans for likely transmembrane helices in a peptide sequence.
Nucleic Acid Secondary Structure	
MFold	MFold predicts optimal and suboptimal secondary structures for an RNA or DNA molecule using the most recent energy minimization method of Zuker.
StemLoop	StemLoop finds stems (inverted repeats) within a sequence.
Translation	
BackTranslate	Use BackTranslate to translate your peptide sequence into a nucleic acid sequence. Choose either the most probable nucleic acid sequence (utilizing a codon frequency table) or the most ambiguous nucleic acid sequence.
* BackTranslate+	Use BackTranslate+ to translate your peptide sequence into a nucleic acid sequence. Choose either the most probable nucleic acid sequence (utilizing a codon frequency table) or the most ambiguous nucleic acid sequence.
Translate	Use Translate to create a peptide sequence from an nucleic acid sequence.

* Translate+	Use Translate+ to create a peptide sequence from an nucleic acid sequence.
**** Utilities**	
* Extract+	Extract a portion from a sequence.
Reverse	Use Reverse to take to complement or reverse your nucleic acid sequence.
Reverse+	Use Reverse+ to take to complement or reverse your nucleic acid sequence.
SeqConv+	Makes a copy of one or more annotated sequences, saving the new file in one of the following supported file formats; GCG RSF, GCG MSF, GCG SSF, GenBank, EMBL, FastA or BSML.

SeqWeb 程式的數量較原有的 GCG 程式少很多，使用者在利用 SeqWeb 進行序列分析時，可以試著到 GCG 中尋找是否有其他可用的分析程式。使用者日後在碰到序列分析的問題時可以多參考 GCG Manual（http://bioinfo.nhri.org.tw/gcghelp/gcgmanual.html），可能就可以找到適合的程式來進行分析。

肆、序列格式簡介

一、序列格式種類

雖然在 GCG command mode 僅接受 GCG 格式的序列，不過 SeqWeb 可以在 Sequence Manager 中自動轉換序列格式，讓 Wisconsin Package 程式能夠辨識與使用。但使用者仍需要知道不同格式的序列有著什麼樣的特徵，這樣當使用不同的生物資訊分析工具時，就不會 Input 錯誤的格式，而使分析結果無法產出。

生物資訊領域所使用的序列格式相當多種，如 simple text、fasta、genbank、GCG、swiss、msf、clustal、phylip……等，各有特色。但不論是那一種序列格式，都是屬於 **ascii code 的純文字檔類型**。換言之，若是使用 Microsoft® Word 所編輯、處理的序列，則不屬於 ascii code，因此無法被絕大部分的生物資訊工具所認識，所以建議在存檔時，請以純文字的檔案格式進行儲存。以下簡單介紹幾種常見的序列格式：

■ Simple Text 格式

這種格式的序列其實就是沒有包含任何註解，純粹只有連續的序列，這樣的序列檔大多用來做為 Web 界面序列分析程式的 input file，例如 SeqWeb、GenWeb 或 NCBI 的序列分析程式。如果序列來源是使用者自己定序而得的，最好是將序列存成這種格式。若要以 GCG Command Mode 做序列分析時，這種序列檔可以直接上傳至 GCG 主機，再經 reformat 後即可轉換為 GCG 的格式進行序列分析。

```
GATCCTCCATATACAACGGTATCTCCACCTCAGGTTTAGATCTCAACAACGGAACCATTGC
CGACATGAG
ACAGTTAGGTATCGTCGAGAGTTACAAGCTAAAACGAGCAGT……
```

■ FastA 格式

這是相當常見的序列格式，除了序列本身還包含一段對序列功能的簡單敘述，這一行註解必須放在檔案的第一行，前面以" > "區別註解及序列的部份。FastA 格式因為相當簡單明瞭，所以一些序列搜尋的程式的資料庫部份常是以這樣的格式來儲存，例如 GenWeb，在它的 output file 中就可以看到 FastA 格式的序列檔。此外，一些網站上的序列分析程式，有時也會要求使用者以 FastA 格式輸入序列。FastA 格式的優點是序列檔僅含重要註解，在儲存序列資料庫時可節省空間，但相對的可供參考的註解資料就相當有限。

```
>gi|1293613|gb|U49845.1|SCU49845 Saccharomyces cerevisiae TCP1-beta gene, partial cds; and
Axl2p (AXL2) and Rev7p (REV7) genes, complete cds
GATCCTCCATATACAACGGTATCTCCACCTCAGGTTTAGATCTCAACAACGGAACCATTGC
CGACATGAG
ACAGTTAGGTATCGTCGAGAGTTACAAGCTAAAACGAGCAGT……
```

gi|1293613--GI 是指 GenInfo Identifier，這是 Genbank 中每條序列檔的特定編號。
gb|U49845--這是這個序列的 Accession number，每個序列檔都有一個獨一無二的代表編號，gb 是指序列來源為 GenBank。Accsession number 在三大

資料庫 (Genbank, EMBL, DDBJ)是可以通用的，在進行序列分析時指定 Accseeion number 方式來進行搜尋或擷取序列是最方便的！
要注意的是，在 NCBI 裡另外建了稱作 RefSeq 的序列資料庫，那些序列並不屬於 GenBank 和 Uniprot 的資料庫，它們的 Assession number 的格式略有不同，常為 NM_123456 (英文字母和數字間多了底線唷！)。如果使用者以為它們屬於 GenBank，而進行搜尋的話，應該是找不到序列的！

■ GenBank 格式

這是 GenBank 原始的序列格式，這種格式包含相當完整的註解，對使用者而言可以在找到序列的同時，藉由這些資料對這段序列能有相當的了解，以下例說明：

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      baker's yeast.
ORGANISM    Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
            .....
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"

/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"

/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
```

```
                     ……
                     VDFSNKSNVNVGQVKDIHGRIPEML"
                     ……
BASE COUNT      1510 a    1074 c    835 g    1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
      ……
     4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Locus-- 這個欄位分別代表locus name(SCU49845)、長度(5028 bp)、序列類別(DNA)、GenBank Division(PLN)和Mordification Date(21-JUN-1999)。其中Division是GenBank中再細分的子資料庫,現在共分 16 個division,請參考第20頁核酸序列資料庫介紹。

Keywords--Keyword 在過去是重要的資料搜尋依據,但因為它並非以 controlled vocabulary 做成,所以現在 NCBI 新的序列資料這一個欄位大多是空的,所以在進行字串搜尋時,NCBI 也建議不要用 keyword 來搜尋,而最好是以全文來搜尋。

Source--包含物種的相關資訊。

Reference--參考文獻,通常會有一個相對應的 Medline 編號,讓使用者可以很方便的找到參考文獻,有時一段序列可能有不只一篇參考文獻。

Features-- 這部份包含現在已知這段所包含的生物相關訊息,例如 Source、Gene, CDS(coding sequence)等。在 Source 的部份,會有這個物種的 NCBI texon ID,CDS 則包含 transclation 的結果以及一個 protein ID,如果想找出這條蛋白質序列,可利用這個 ID 到 GCG 的 GenPept 資料庫找,或是以 Genpept:AAA96885 的方式指定之。

■ 4. GCG 格式

在 GCG 中所有的序列分析程式都必須為 GCG 格式方能進行,GCG 格式的序列檔基本上會依循其來源的基本格式,例如 GenBank 來源的序列檔,內容和 GenBank 完全相同,但格式略做修改。在 GCG 格式中有兩項重要標記:

其一,為檔案第一行,含雙驚嘆號之序列格式說明。若是屬於核酸序列,則會標註 NA_SEQUENCE;若為蛋白質序列,則會標註 AA_SEQUENCE,程式可針對此說明判斷出序列種類。

其二,為註解文字末端的分節符號"**..**",GCG 格式以**兩個句點**來區分註解及序列。在雙句點符號以下,就屬於序列本身的內容。

```
!!NA_SEQUENCE 1.0
LOCUS          SCU49845      5028 bp      DNA                  PLN        21-JUN-1999
DEFINITION   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
             (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION    U49845
VERSION      U49845.1   GI:1293613
KEYWORDS     .
SOURCE       baker's yeast.
  ORGANISM   Saccharomyces cerevisiae
             Eukaryota; Fungi; Ascomycota; Saccharomycetes; Saccharomycetales;
             Saccharomycetaceae; Saccharomyces.
REFERENCE    1   (bases 1 to 5028)
  AUTHORS    Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
```

```
   TITLE      Cloning and sequence of REV7, a gene whose function is required for
              DNA damage-induced mutagenesis in Saccharomyces cerevisiae
   JOURNAL    Yeast 10 (11), 1503-1509 (1994)
   MEDLINE    95176709
              ..........................................................................
FEATURES                Location/Qualifiers
     source             1. .5028
                        /organism="Saccharomyces cerevisiae"
                        /db_xref="taxon:4932"
                        /chromosome="IX"
                        /map="9"
     CDS                <1. .206
                        /codon_start=3
                        /product="TCP1-beta"
                        /protein_id="AAA98665.1"
                        /db_xref="GI:1293614"

/translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                        AEVLLRVDNIIRARPRTANRQHM"
     gene               687. .3158
                        /gene="AXL2"
     CDS                687. .3158
                        /gene="AXL2"
                        /note="plasma membrane glycoprotein"
                        /codon_start=1
                        /product="Axl2p"
                        /protein_id="AAA98666.1"
                        /db_xref="GI:1293615"

/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                        ..........................................................
BASE COUNT      1510 a    1074 c    835 g    1609 t
ORIGIN

U49845  Length: 5028  November 30, 2000 10:22  Type: N  Check: 3941  ..

        1 GATCCTCCAT ATACAACGGT ATCTCCACCT CAGGTTTAGA TCTCA      AAC
          ..........................................................
     5001 TTTTAAGCTA TTCAATTTCT CTTTGATC
```

以兩個句號區別註解和序列！

在 GCG 資料庫中還有其他格式的蛋白質或核酸序列，大多和 GenBank 類似，僅在註解項目及格式上略有區別。

二、SeqWeb 所輸出的檔案型式

SeqWeb 的檔案可以直接拿到 GCG command mode 來使用(因為已經自動加入了兩個句號)，那些序列檔可以利用 Sequence manager 中的 Save as 的功能，存至個人電腦中，再上傳至 GCG 主機就可以了。

若是自行分析或經定序所得的結果，可將序列存成純文字檔(*.txt)，再將這些檔案上傳至 SeqWeb，不需 reformat 就可直接使用。若要將 GCG command mode 所得到的序列利用 SeqWeb 來分析，則必須以檔案傳輸程式(ftp)先將檔案下傳至個人電腦中，再以 Sequence manager 中的 Add form local file 的功能加入 SeqWeb 中。要注意的是，序列檔案或文字結果的檔案在下載或做任何修改時最好都是存成純文字檔(*.txt)，並以 ASCII mode 來進行傳輸。

SeqWeb 所輸出的檔案類型如下表:

Program	List File	MSF File	Sequence File
BackTranslate			✓
BLAST	✓		
FastA	✓		
GrowTree		✓	
LookUp	✓		
PileUp		✓	
ProfileSearch	✓		
Reverse			✓
SSearch	✓		
StringSearch	✓		
Translate			✓

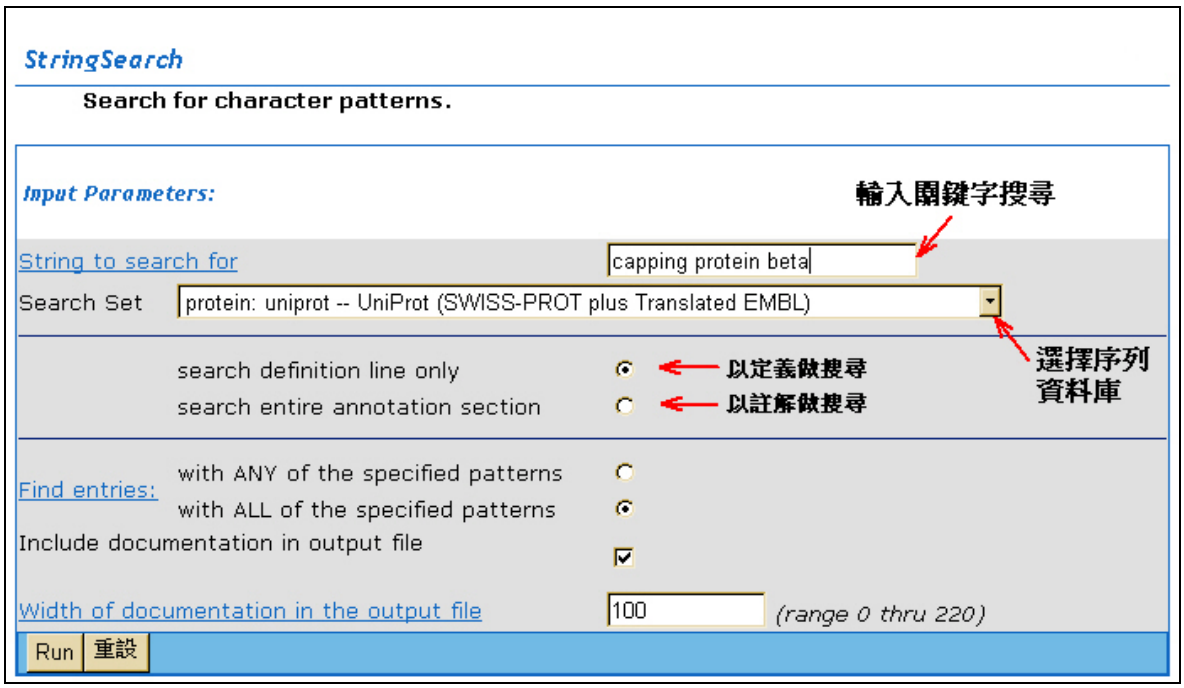
如果 GCG command mode 的 list file 要上傳至 SeqWeb，必須先用 reformat 將 list file 轉成 RSF file 再上傳至 SeqWeb，如果這個 list file 中有 10 個序列檔案，會是以 10 個檔案上傳至 SeqWeb 中，而非單一個 list file。

伍、以文字搜尋資料庫

一、StringSearch：以字串尋找所要的序列

可以用字串直接搜尋序列檔中的定義或註解部份，以找出與其相關的序列檔案，如果已知要找的序列是在某個資料庫中，最好能同時限制所搜尋的資料庫，以更快找到所需的序列。這個程式所輸出的結果在 SeqWeb 中可直接 hyperlink 查看各序列資料。

StringSearch 和接下來介紹的 LookUp 最大的不同，在於可以指定的子資料庫的種類較多，上表所列的資料庫都可以指定，若是要找的序列只限於某一資料庫中時可以省下搜尋所有資料庫及查看結果的時間。

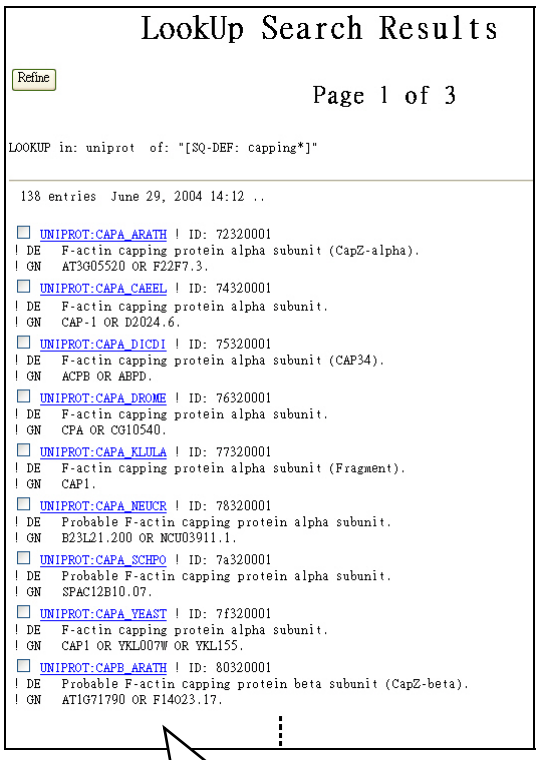


決定要搜尋的部份時，可以先選擇尋找序列檔註解的定義欄(definition)，以節省搜尋的時間，若是這樣找不到，再選擇尋找整個註解欄(annotation)的部份。

StringSearch 的結果可以儲存起來，但建議將結果存成 HTML 之形式 (選擇 Save as HTML)，這樣以後要查看序列詳細內容時，只需點選上面的超連結即可馬上看到。

二、LookUp：以 keyword 尋找所要的序列

LookUp 可以選擇的子資料庫種類較少，但是卻另外列出許多項目，可供指定要搜尋的是序列內容中的那一部份。LookUp 的搜尋方法和 StringSearch 不太相同，但是速度上會較 StringSearch 快得多。特別需注意的是：兩者所得的結果或許會略有不同，因此使用者可以自己喜好挑選合適的程式。



在 LookUp 程式中，於 definition 欄位以 “**capping**” 作爲關鍵字，用以搜尋 capping protein。找到的結果會以列表方式列出，使用者可以點選超連結觀看序列內容，或是勾選序列加入至 Sequence Manager，亦可選擇將結果儲存爲 html 檔

陸、以序列搜尋比對資料庫

一、序列比對分析基本概念

在找到一段新的序列，想知道它可能的功能時，通常會先把這條序列去和資料庫中的序列進行比對，再將比對的結果依序列相似度（similarity）的高低做排列，來判斷是否這條未知序列和現存的序列有 homology（同源性），然後根據已知序列的特性來預測未知序列可能的功能。例如某一條由果蠅得來的未知序列，經過比對，和人類以及老鼠的 Capping Protein 有很高的 homology，我們可以推測這條序列可能就是果蠅的 Capping Protein。序列比對有四個要素，包括比對方式(type of alignment)、計分法(scoring system)、演算法的應用(implementation of algorithm)，以及比對結果在統計上的意義(statistical significance)。

進行序列比對首先要選擇比對方式，目前序列比對方式主要有 Global Alignment 和 Local Alignment 兩種。**Global Alignment** 是將兩條序列頭對頭、尾對尾進行比對，有時會在中間加入空格(gap)，以完成整條序列的比對，用來比對序列整體的相似度，如圖 1。**Local Alignment** 則是用來找到兩條序列中相似度最高的區域(subregion)。它是在兩條序列中各給定一個子序列，將子序列以 Global Alignment 的方式 （頭對頭、尾對尾，必要時在中間加入空格）進行比對，求出相似度得分；Local Alignment 比對的結果，是所有進行比對的子序列對（sequence pair）組合中，Global Alignment 得分最高的一個子序列對（MSP, maximal segment pair）。如圖 2。

圖 1: <u>Global Alignment</u>

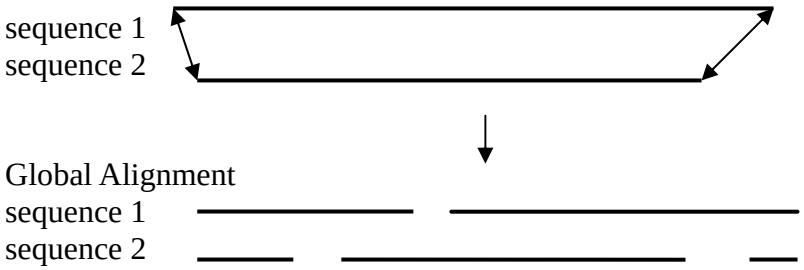
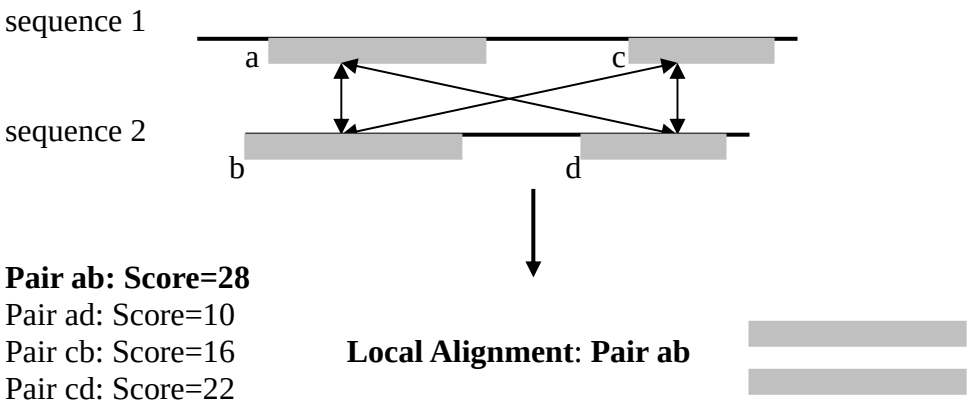


圖 2: <u>Local Alignment</u>



比對方式的選擇取決於進行比對序列的性質以及研究目的。通常使用 Global

Alignment 來做序列整體相似性的比對，適用於相近似的序列。而想知道未知序列中是否含有已知的功能性區域，(例如想知道一個 cell membrane protein 序列中有沒有 GTB binding domain）則可使用 Local Alignment。Local Alignment 的比對方式可以比對出整條序列的相似度，也可以在整條序列相似度低的情況下，比對出相似度較高的區域，因此能夠顯示序列與功能之間的關係，而應用較廣，目前最受歡迎的 BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool)就是 Local Alignment 的搜尋比對工具。

序列比對首先要遇到的就是計分的問題。DNA 或蛋白質序列都是由一個個的 residue 所組成，序列比對就是要比較兩條序列在相對應位置上 residue 的相似度。用數字表達這個相似度，才能方便數學運算，得出客觀的比對結果。在一個位置上，兩條序列具有相同的 residue 給幾分、相異的 residue 給幾分，必須有一個計分的標準，這個計分標準就是計分法(scoring system)。DNA 序列因為只有 A、T、G、C 四個 nucleotide，通常計分時 match 則給 1 分, mismatch 則扣 3 分，再加上空格扣分（gap penalty）。但是蛋白質序列的計分標準就要考慮 aminio acid（胺基酸）基本物化特性，以及這個 amino acid 對整條序列的重要性。例如兩條蛋白質序列在同一位置的 residue 分別是 Serine、Threonine，會因為這兩個 amino acid 物化特性相近給較高分，相對的，如果兩條蛋白質序列在同一位置的 residue 分別是 Serine、Asparagine，則給分較低。因此蛋白質序列計分法較為複雜，需要作成計分表(Scoring Matrix) 來計分，如 BLAST 所使用的 BLOSUM 62，就是一個常用的 Scoring Matrix。就比對結果而言，若比對核酸序列，序列本身僅由 A、T、G、C 組成，得到相似序列的可能性較高，比對結果的資料筆數會較多；若是比對蛋白質序列時，因為胺基酸至少有二十種，而各胺基酸彼此間還可細分為不同性質的 group，所得的結果在生物意義上，會比核酸的比對結果要來得豐富。

計分法之外，還要加上演算法（algorithm），才能完全解決以數學運算的方式比對序列相似度的問題。由於序列分析的演算法牽涉到複雜的數學與統計原理，我們不多作介紹。要強調的是，序列分析程式的演算法必須要在基本理論上考慮到 DNA 和 Protein 的基本生物學原理，比對結果才可能會有生物意義。此外，演算法也必須考慮到電腦 CPU 運算時間以及 storage 的限制，來加快比對的速度。通常 Global Alignment 的程式採用 Needleman & Wunsch algorithm，而 Local Alignment 則常用 Smith & Waterman algorithm，這兩種 algorithm 皆為 dynamic programming 之特例。

序列比對時，電腦程式會依計分法計分，然後用演算法運算，得出比對結果。以下分別以簡單的實例說明 Global Alignment 和 Local Alignment 的比對步驟。

例：比對下列二條序列的相似度
sequence 1: HEAGAWGHEE
sequence 2: PAWHEAE

■ Global Alignment

1. 選定計分表，本例採用 BLOSUM 50，同時設定空格扣分（gap penalty） 為 -8。

BLOSUM 50 計分表如下：

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	**7**	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	**7**	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	**8**	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4

C	-1	-4	-2	-4	**13**	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	**7**	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	**6**	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	**8**	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	**10**	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	**5**	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	**5**	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	**6**	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	**7**	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	**8**	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	**10**	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	**5**	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	**5**	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	**15**	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	**8**	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	**5**

2. 根據計分表，以 Dynamic Progamming 的 algorithm （Needleman Wunsch algorithm）算出相似度最高的比對結果為：

sequence 1:	**H**	**E**	**A**	**G**	**A**	**W**	**G**	**H**	**E**	**--**	**E**
sequence 2:	**--**	**--**	**P**	**--**	**A**	**W**	**--**	**H**	**E**	**A**	**E**
Score=1	-8	-8	-1	-8	5	15	-8	10	6	-8	6

■ Local Alignment

1. 選定計分表，本例採用 BLOSUM 50，同時設定空格扣分（gap penalty） 為 -8。同樣以 BLOSUM 50 計分表作為計分標準。

2. 以 Dynamic Progamming 的 algorithm（Smith ＆ Waterman）找到一段相似度得分高的區域為：

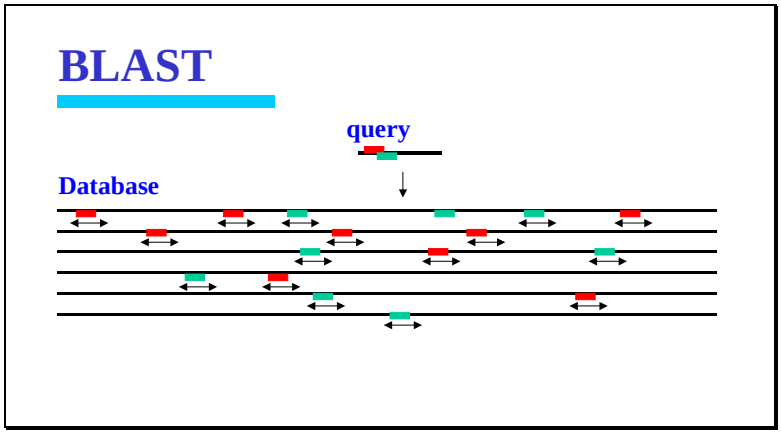
sequence 1:	**A**	**W**	**G**	**H**	**E**
sequence 2:	**A**	**W**	**--**	**H**	**E**
Score=28	5	15	-8	10	6

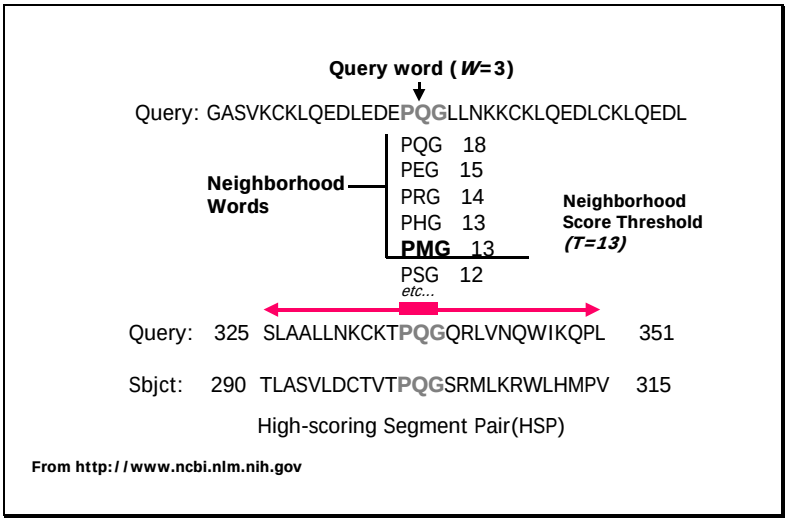
比對結果在統計上的意義(statistical significance)，對於序列比對結果的判讀非常重要。之所以要對序列比對結果做分析，在學理上的出發點是：到底比對出來的相似度是真的，還是碰巧（by chance）得到的？因此以隨機發生的機率來檢視比對結果，可以幫助我們判讀相似度的意義。目前 Global Alignment 的比對結果在統計上的意義較少理論，但是 Local Alignment 的比對結果分析，則已經建立了一些統計理論，可以算出 E-value 作為判讀的參考。以 BLAST 程式為例，序列比對完成後，所有比對出來的序列都會有一個相似度的得分 S 值 (即 Score，是依計分法計分所得出來的值)，接著，BLAST 程式會算出對應於每個 S 值的 E-value。E-value 是一個期望值，是指假設序列的排列是隨機（random）的，在給定序列長度下，碰巧得到相似度得分大於或等於 S 的片段對個數（HSP, high score pairs）的期望值。所以 **E 值愈小**，表示比對出來的相似度**愈不是隨機排列**偶然發生的，而是可能具有生物意義的。

二、BLAST 程式操作

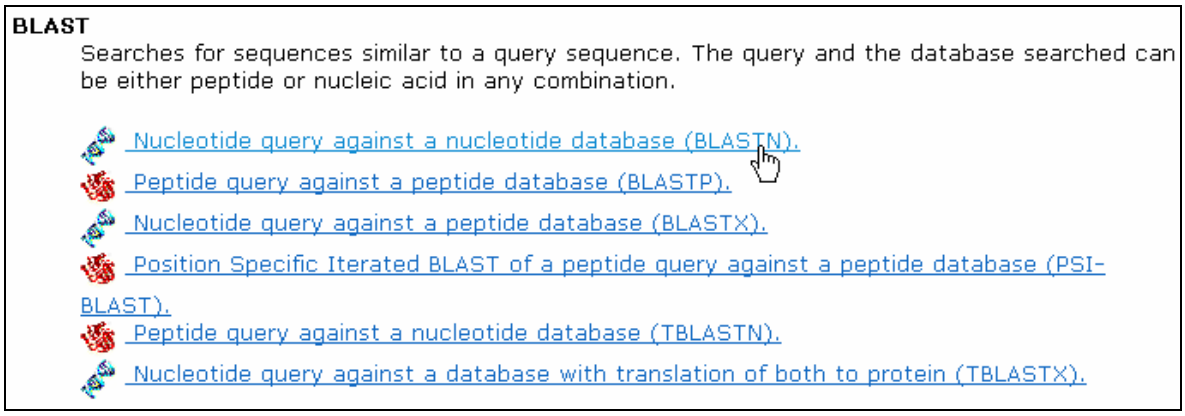
現在較常用的、針對 Local Alignment 開發出來的資料庫搜尋比對程式，有 Smith-Waterman （GenWeb）、FASTA 及 BLAST 三種。Smith-Waterman 應用 dynamic programming 的方法，FASTA 和 BLAST 則是用 Heuristic Algorithm (逼近法)。三者之中 Smith-Waterman 是最先發展出來的，它的靈敏度（specificity）最高，達 70%~88%，但計算量最大，所花的的時間最長，而後發展的 FASTA 速度較 Smith-Waterman 快，靈敏度略差；速度最快的是 BLAST，但靈敏度在三者中較低，約為 53%~67%。BLAST 因為搜尋速度最快，是現在最受歡迎的序列搜尋方法，以下以 BLAST 及 NCBI 的 NetBLAST 為例加以說明。BLAST 的完整的比對的步驟包括：

1. Start：設定比對起始序列 nucleotide 數目(word length, w 值)
2. Scaning phase：依照 word length 選取 query 序列，至資料庫搜尋具有相同序列的片段。
3. Extension phase：找到具有相同序列的片段後，由這片段的一端或兩端開始延伸，每延長一個 nucleotide 就計算一次相似度得分（根據計分表，採用 dynamic programming），一旦延長後相似度得分降低到一定的程度，即不再延長，得到 maximal segment pair （MSP）score, 把這一段序列為比對結果。
4. 重複步驟 2. 至 3. 直到完成所有可能的比對組合為止。
5. 把得分較高的比對結果列出，並給出其 E-value。





■ BLAST



BLAST 程式針對各個 query 及 database 屬於 DNA（N）或蛋白質（P）的不同性質，分別寫成了不同的程式：

BLASTN：以核酸序列搜尋核酸序列資料庫，最常用。

BLASTP：以蛋白質序列搜尋蛋白質序列資料庫

BLASTX：以核酸序列搜尋蛋白質列資料庫，將 query 的核酸序列轉譯為六個 reading frame 的蛋白質序列，再與蛋白質序列資料庫比對。

PSI-BLAST：以蛋白質序列查詢蛋白質資料庫，並利用搜索的結果重新構建 protein profile，找出屬於相同 protein family 的序列。

TBLASTN：以蛋白質序列搜尋核酸序列資料庫，先將資料庫中的所有核酸序列轉譯為六個 reading frame 的蛋白質序列後，再與 query 的蛋白質序列比對。

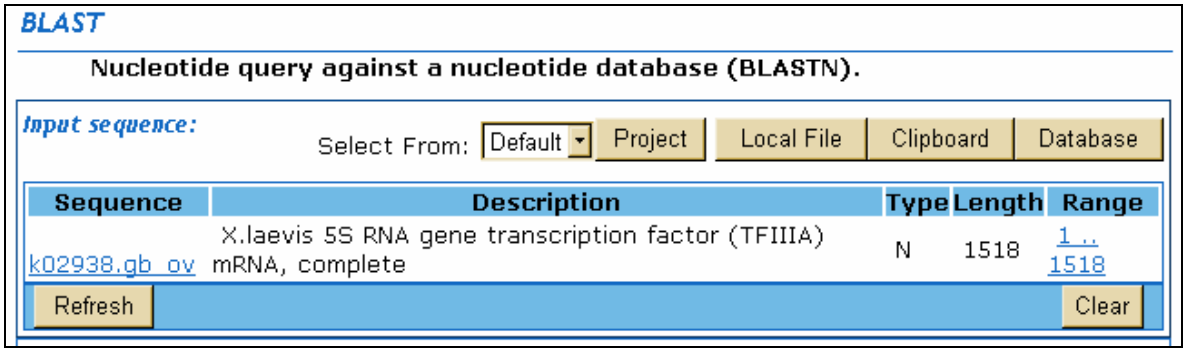
TBLASTX：將核酸序列及核酸序列資料庫都轉譯為六個蛋白質 reading frame 再進行比對。

TBLASTN 和 BLASTX 所需的計算量甚大，有時會運用來找尋具生物意義的結果。至於 TBLASTX 則是計算量最重的方法，若非有特殊需要，否則請不要在 SeqWeb 中進行這類的比對。

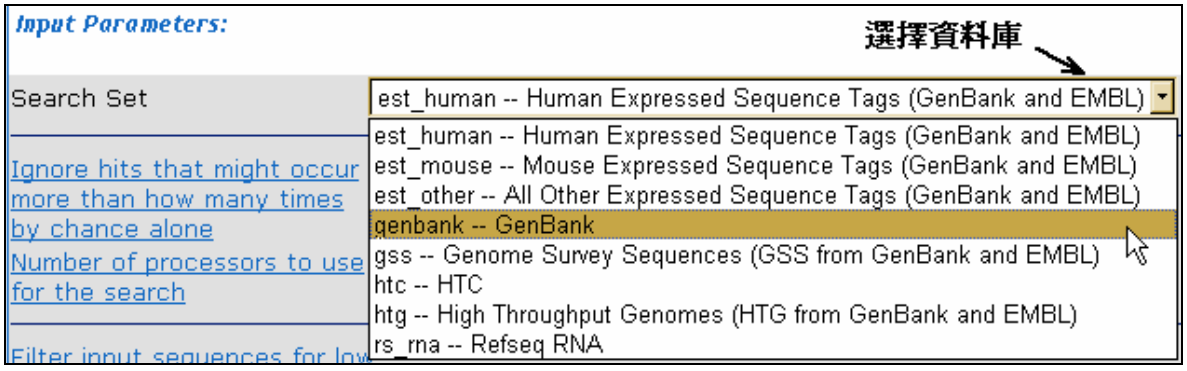
■ BLASTN 操作步驟

A. 由 SeqWeb 的 Database Searching 類別的 Similarity 進入 BLASTN 程式。因為 BLASTN 是分析核酸用的程式，query 及 database 都必須是核酸序列，所以在這個畫面中看不到蛋白質序列及蛋白質資料庫，如果您在選單中找不到比對的序列，除了可以回 sequence manager 中再加入核酸序列外，也可以在這個畫面加入核酸序列，加入的方法和 sequence manager 相同。同樣的，如果採用的是分析蛋白質的 BLASTP，則必須使用蛋白質序列及蛋白質資料庫，核酸序列及核酸資料庫就不會顯示。

B. 選取 project, 並選取 input sequence。如果要 key in 自己的序列，可以選擇 Clipboard。



C. 選取資料庫。



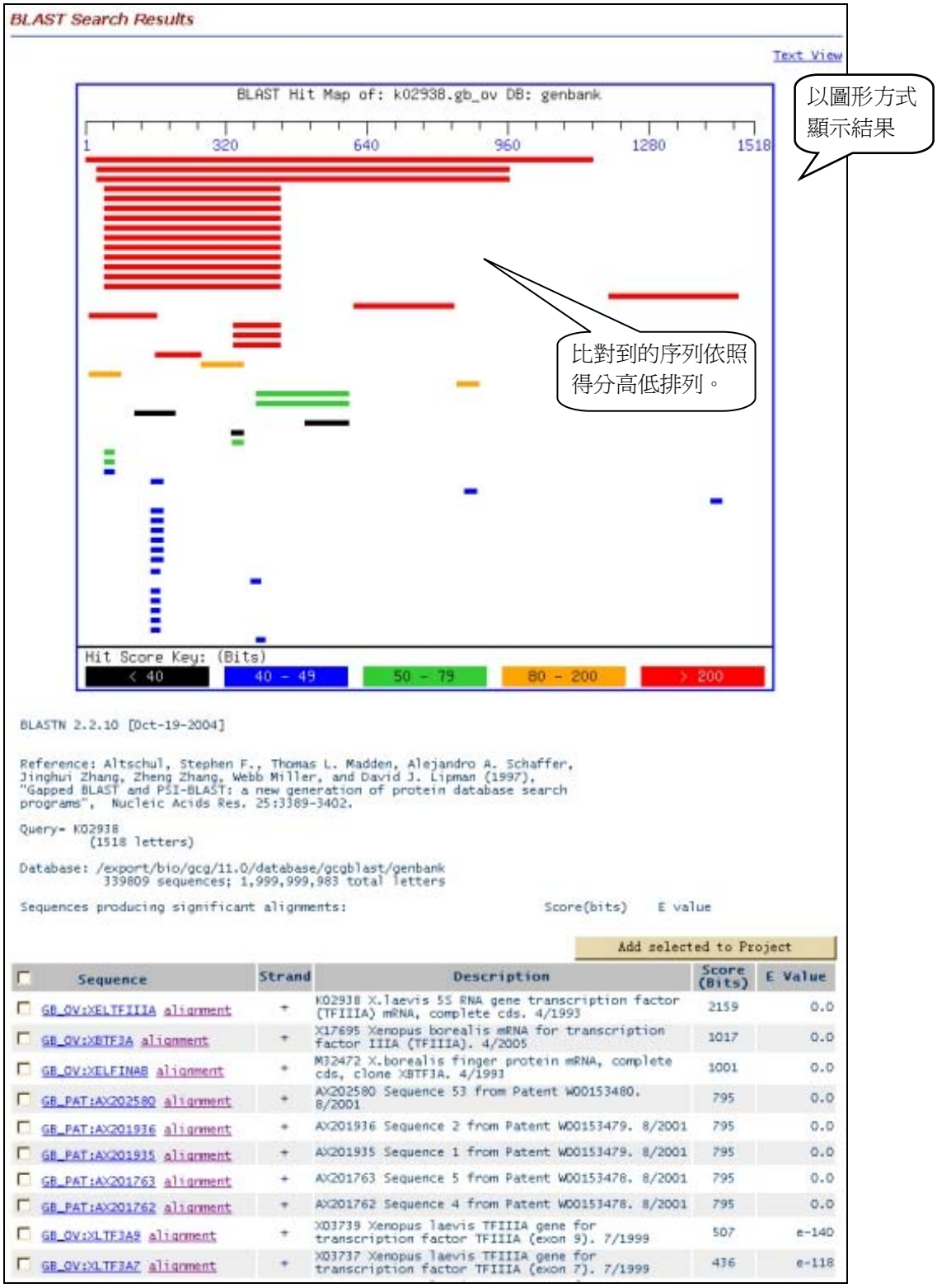
在選擇資料庫的選單中，可以看到不同的資料庫選項，建議選擇 GenBank (此為所有的核酸序列)，這樣比對的資料庫是完整並且具有詳細註解的內容，對未知序列的功能的預測很有幫助。

D. 設定下列參數：

1). Ignore hits that might occur more than how many times by chance alone（Default 值為 10）：設定 E-value 值，大於設定值的序列就不列在分析結果中。建議設為 0.01。
2). Number of processors to use for the search（Default 值為 1）：設定本次搜尋所需要的電腦主機運算空間。請使用 Default 值。
3). Filter input sequences for low complex/repeat regions（Default 值為忽略 repeat sequences）：是否忽略 repeat sequences，如勾選，則程式在比對時遇到 query sequence 中的 repeat sequence 即略過不做比對。
4). Reward for all nucleotide matches（Default 值為 1）：同一位置比對得到相同 nucleotide 的得分。
5). Penalty for all nucleotide mismatches（Default 值為-3）：同一位置比對得到不同 nucleotide 的扣分。
6). Word size（Default 值為 11）：起始 nucleotide 長度。BLAST 程式會先以一小段序列作為比對的起始進行比對。設定起始 nucleotide 長度愈短，則較不易漏失相似序列，但相對而言需要做的運算就愈多，且需時愈長。
7). Create gapped alignments（Default 值為允許在比對時加入 gap）：是否允許在比對時加入 gap（空格）。
8). Gap creation penalty（Default 值為 5）：比對時加入 gap 的扣分。
9). Gap extension penalty（Default 值為 2）：比對時延長 gap 的扣分。

10). Maximum number of sequences listed in the output（Default 值為 500）：列出分析結果序列的數目，可視需要增加或減少，最多可以列出 1000 條。

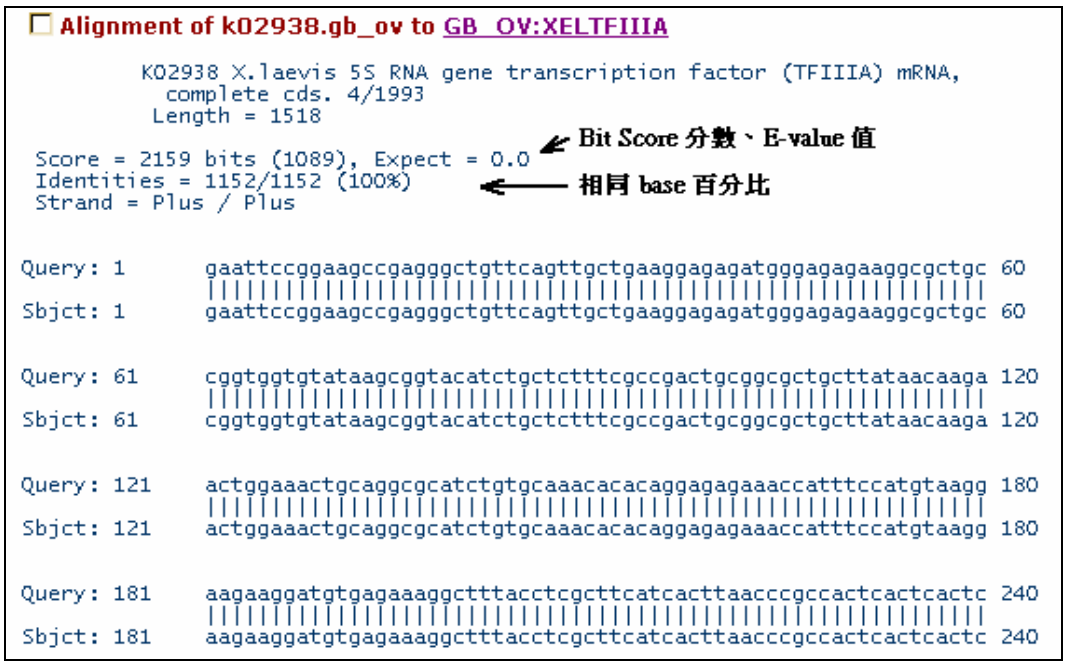
E. 按 Run 執行。跑完分析的結果如下圖，結果網頁的上方先以圖形呈現，再將文字的及找到的序列連結至於下方。



在這個結果中，每條序列是依相似度得分(Score)由高至低排序。Score是程式以計分法(如protein sequemce選用BLOSSUM 或PAM)的計分結果，每一條序

列都有一個Score，接著，BLASTN程式對每一個Score作統計分析，得出每一條序列的Bits Score以及E-value。Bits Score 是機率值，如果Bits Score為 2597，表示至少要做 2^{2597}次比對才能得到如此高分的結果。E value是期望值，表示在隨機狀態下，可以得到相同score的同樣長度的序列的個數，所以E值越小，表示比對出來的相似度愈可能有生物意義。

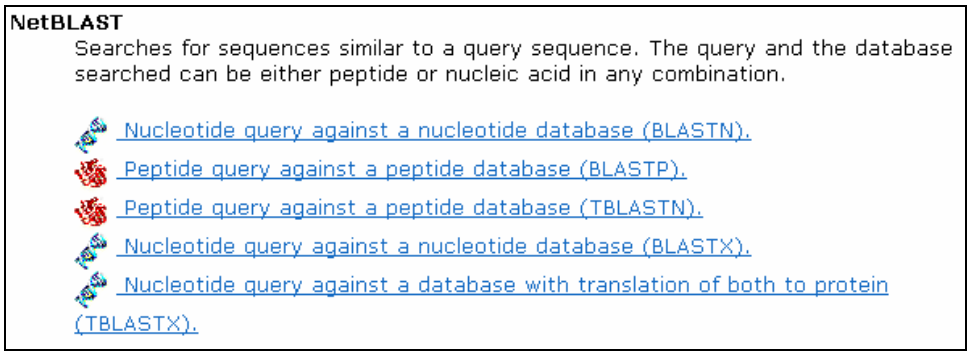
分析結果的每一條序列皆設有超連結，點選序列的超連結即可看到這條序列的詳細資料。若是按 alignment 就可看到 query 和這條序列最相似的區域的並列分析結果。(如下圖)

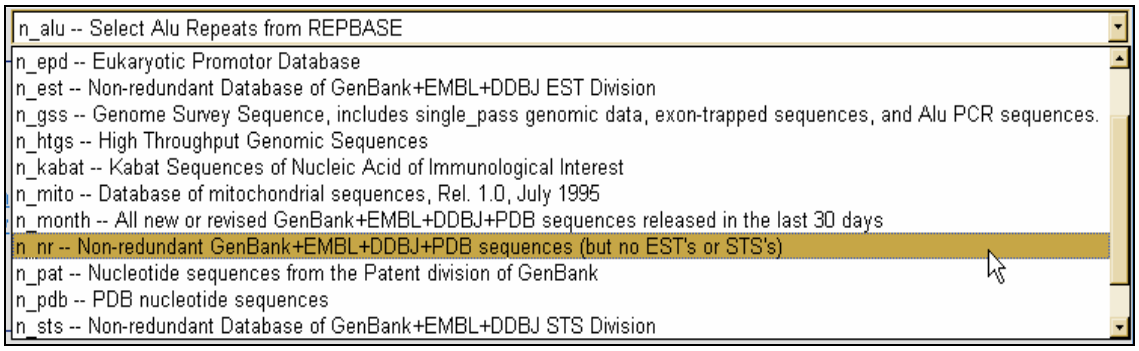


因為 BLAST 是做 Local Alignment，所以兩條序列其實並非從頭到尾都可以成功的並列在一起，而是比對出整條序列的這一小段可以並列得最好，而這裏所寫的 identity 也只是指這一小段的 identity，並不是指整條序列的 identity。BLAST 會將兩條序列間所有相似性高的片段都會列出，有時會看到同樣的兩條序列在不同的區域都有相似性高的片段。BLAST 所得的結果在 SeqWeb 最好也存成 HTML 的格式，日後在看結果時若想查看每條序列的詳細敘述，就可以直接按序列上的超連結就可以了。

■ NetBLAST

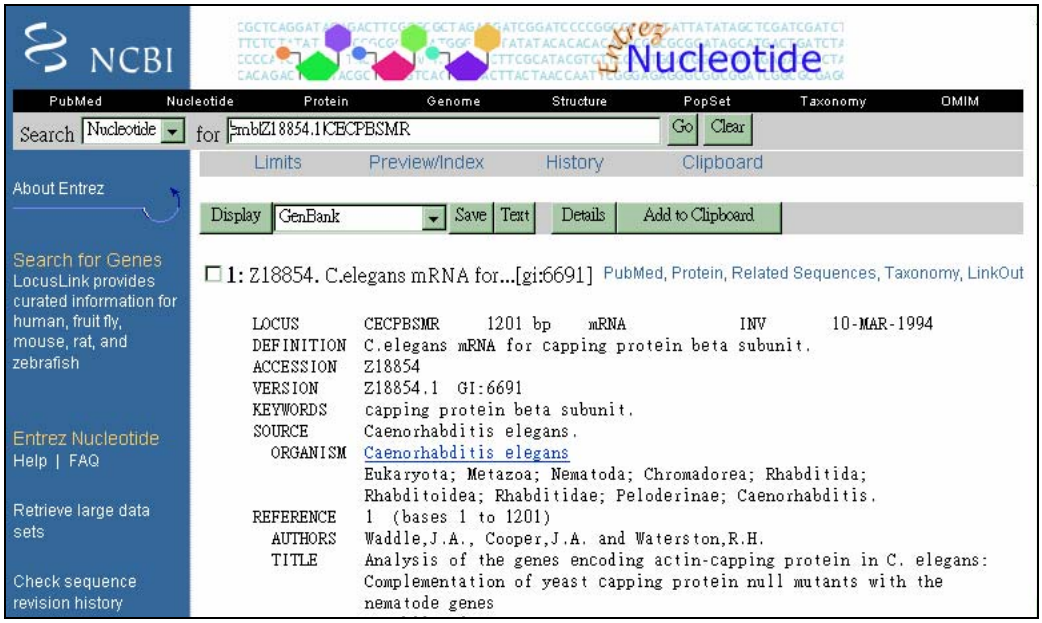
NetBLAST 和 BLAST 唯一的不同是：NetBLAST 是將序列直接送到美國 NCBI 去進行 BLAST，再將結果送回來，這樣使用者就可以直接在 SeqWeb 的介面下執行 NCBI 的 BLAST 程式，比對最新的資料庫，並可以直接透過這個介面將 NCBI 的序列加到 SeqWeb 的 sequence manager 中以進行進一步分析。





NetBLAST 的資料庫的選擇較 GCG 的 BLAST 多，分類也不太相同，還包含一些 GCG 中沒有的資料庫或特殊的分類。

比對結果與 GCG 中的 BLAST 一樣，不過僅有文字顯而無圖形，但是點選序列的超連結時，就會直接連到 NCBI(如下圖)。所以若在 NetBLAST 中選擇 Add Selected Sequence 時，加到 Sequence Manager 中的就會是 NCBI 的序列。雖然 GCG 的 GenBank 資料庫和 NCBI 的時間差僅有幾天，但因為 NCBI 中有些資料庫 GCG 沒有，所以可以這樣直接比對並加入序列還是很方便的。



■ 參考資料

A. BLAST

1. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) A basic local alignment search tool. *Journal of Molecular Biology* **215**:403-410.

2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

3. *EXPECT option (stochastic model for assessing chance alignments):*
Karlin S and Altschul SF (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Science. 87:2264-2268.

B. FASTA

1. Pearson WR and Lipman DJ (1988). Improved tools for biological sequence analysis. Proceedings of the National Academy of Science. 85:2777-2448.

2. Pearson (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in Enzymology. 183:63-98.

C. Matricies

1. *BLOSUM:*
Henikoff S and Henikoff JG (1992). Amino acid substitution matricies from protein blocks. *Proceedings of the National Academy of Science*. **89**:10915-10919.
2. *PAM:*
Altschul SF (1991). Amino acid substitution matricies from an information theoretic perspective. *Journal of Molecular Biology*. **219**:555-565.
3. *Smith-Waterman:*
Smith TF and Waterman MS (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*. **147**:195-197.

D. Match Scoring

1. *Scoring:*
Karlin S and Altschul SF (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Science*. **90**:5873-5877.

柒、多序列並列分析

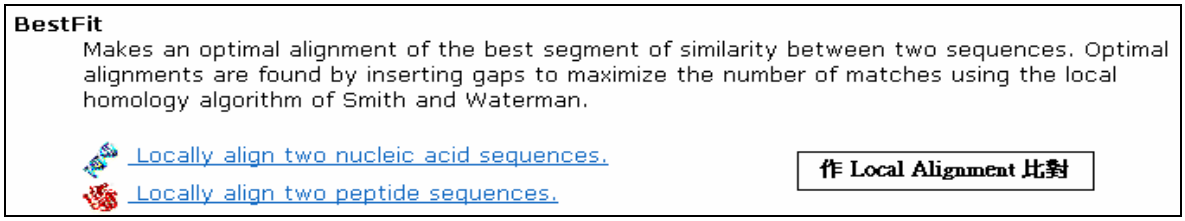
多序列並列分析通常是用在直接比對兩條或多條序列彼此間相似程度，並且可以將最好的排列方式列出。這些程式可以協助使用者比對兩條序列整體的相似度或是尋找序列間最相似的 Conserved Region，或是變化較大的區域。所得的結果可以用來判斷一群蛋白質序列最相似的重要 motif 或是用來做為進一步演化分析的基礎。

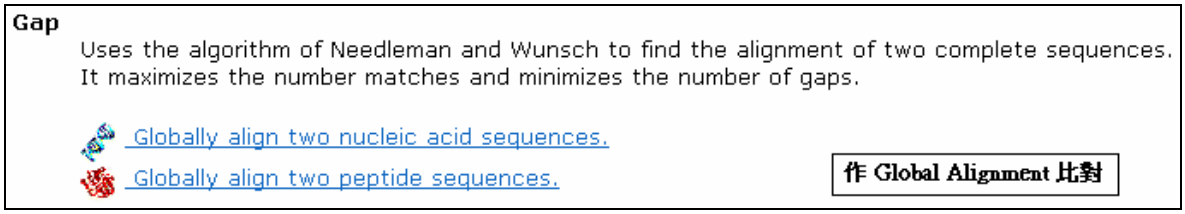
在進行多序列並列分析時，如果比對的序列恰好就是整個檔案中的序列時，就可以直接使用 SeqWeb，但若是需要指定序列中的某一段來進行比對時，使用 GCG Command Mode 會是比 SeqWeb 更方便的選擇。因為如果使用 SeqWeb，就必須一一修改所要比對的序列成為要比對的長度，並一一另存新檔，再用這些新檔案來比對才能得到較好的結果。但在 GCG command mode 中，則程式本身可供使用者指定要比對的序列部份以及正反股(如 BestFit)，或是可以利用 list file，直接指定每個序列中想要 input 的段落來進行比對(如 PileUP)。

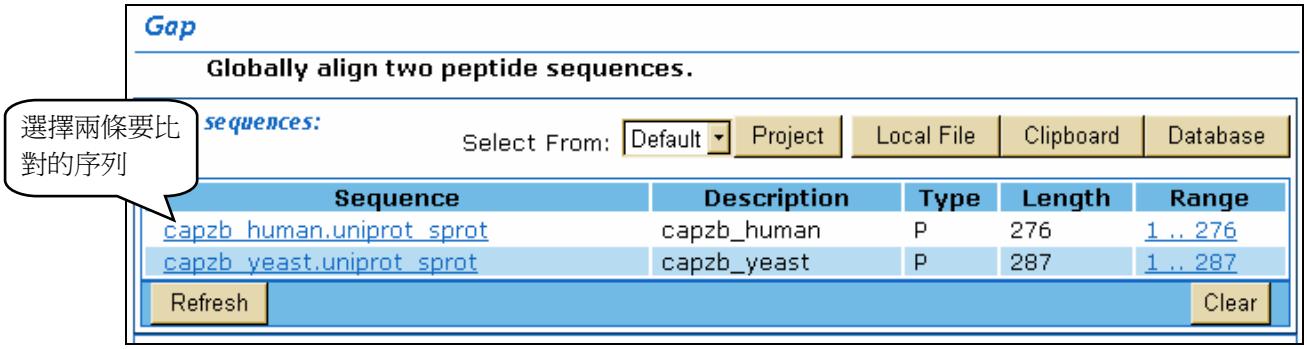
一、BestFit 與 GAP：雙序列並列分析

BestFit 和 Gap 是最常用來比對兩個序列間相似性的程式，BestFit 是用來尋找兩段序列間最佳排列區域(local alignment)；而 Gap 則是用來尋找兩條序列整體的最佳排列方式之用(global alignment)。這兩者的 algorithm 並不相同，BestFit 只列出相連的最相似的部份(也就是儘量沒有 gap 的產生)，而 Gap 則是幾乎是兩條序列從頭到尾完全並列的分析，會在序列中間插入許多的 gap。

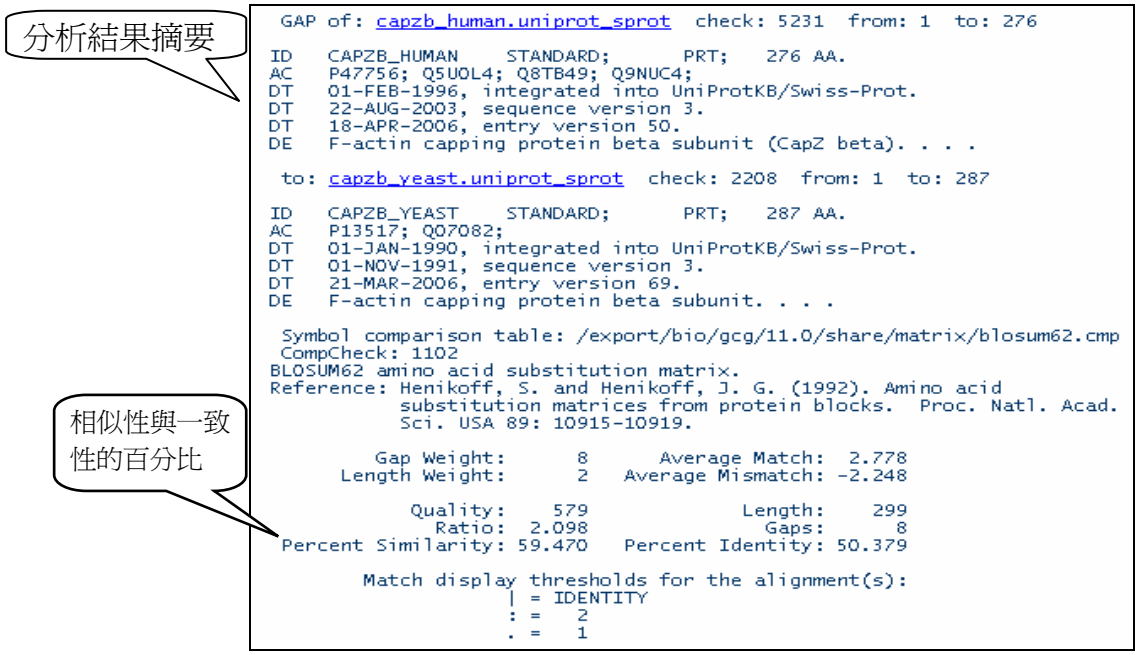
一般來說，當兩條序列相似度甚高時，執行這兩個程式並無法看出明顯的不同，使用者需了解自己要比對的目的為何，再選擇適合的程式。



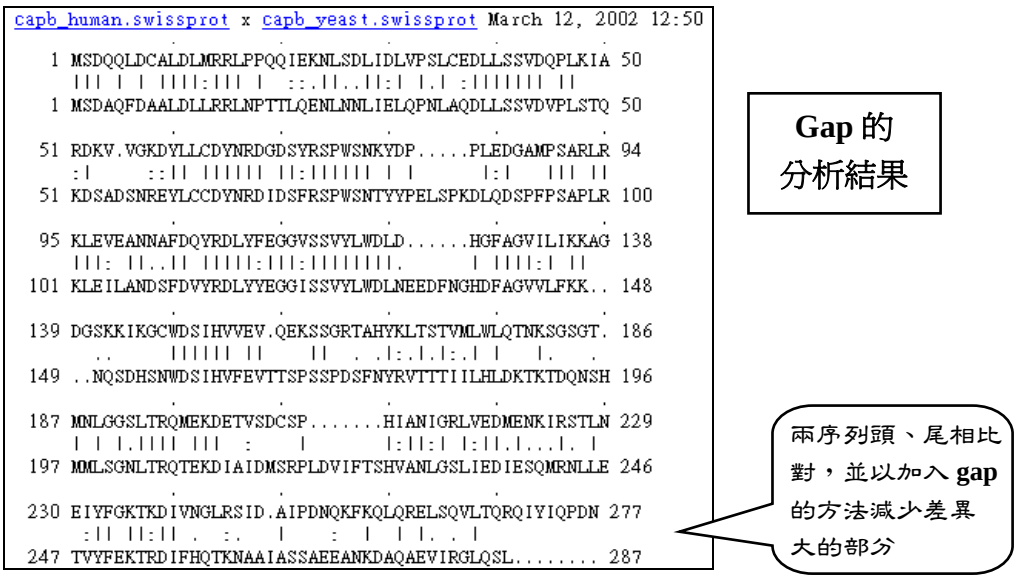




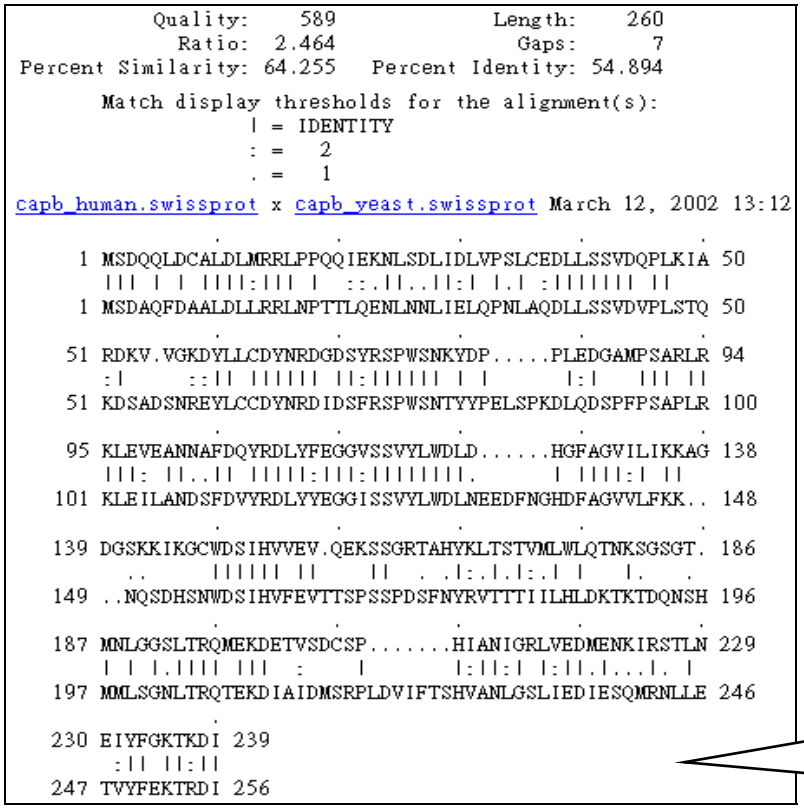
Gap 與 BestFit 均是對兩條序列比對的程式，進入這兩者的程式畫面後，會發現參數設定及選項完全相同。操作時，從 Sequence Manager 中勾選兩條序列加入分析，按執行後很快就能得到結果。



執行完 Gap 程式的結果後，首先會先告知結果的一些摘要，包括比對後序列全長、加入空隙數、Percent Similarity 和 Percent Identity 等統計數值。其中當比對的是核酸序列時，那麼 Similarity 和 Identity 會完全一樣，而在 display 時也只有 Identity 和 Mismatch(空白)兩種，若是蛋白質序列，則會依兩個胺基酸的相同、相似或不同而在 Similarity 與 Identity 兩者的數值上出現差異。



Gap 的並列分析結果(上圖)和 BestFit 的並列分析結果(下圖)，可以明顯看到 Gap 為了使兩條序列儘量從頭到尾並列在一起，加入了很多的 Gap，並能看出其實整體的相似度普通。而 BestFit 就僅顯示出兩段序列最相似的區域，其他差異大的區域將不列出。



```
          Quality:    589              Length:    260
            Ratio:  2.464               Gaps:      7
Percent Similarity: 64.255   Percent Identity: 54.894

      Match display thresholds for the alignment(s):
                  | = IDENTITY
                  : =   2
                  . =   1

capb_human.swissprot x capb_yeast.swissprot March 12, 2002 13:12

     1 MSDQQLDCALDLMRRLPPQQIEKNLSDLIDLVPSLCEDLLSSVDQPLKIA 50
     1 MSDAQFDAALDLLRRLNPTTLQENLNNLIELQPNLAQDLLSSVDVPLSTQ 50
    51 RDKV.VGKDYLLCDYNRDGDSYRSPWSNKYDP.....PLEDGAMPSARLR 94
    51 KDSADSNREYLCCDYNRDIDSFRSPWSNTYYPELSPKDLQDSPFPSAPLR 100
    95 KLEVEANNAFDQYRDLYFEGGVSSVYLWDLD......HGFAGVILIKKAG 138
   101 KLEILANDSFDVYRDLYYEGGISSVYLWDLNEEDFNGHDFAGVVLFKK.. 148
   139 DGSKKIKGCWDSIHVVEV.QEKSSGRTAHYKLTSTVMLWLQTNKSGSGT. 186
   149 ..NQSDHSNWDSIHVFEVTTSPSSPDSFNYRVTTTIILHLDKTKTDQNSH 196
   187 MNLGGSLTRQMEKDETVSDCSP.......HIANIGRLVEDMENKIRSTLN 229
   197 MMLSGNLTRQTEKDIAIDMSRPLDVIFTSHVANLGSLIEDIESQMRNLLE 246
   230 EIYFGKTKDI 239
   247 TVYFEKTRDI 256
```

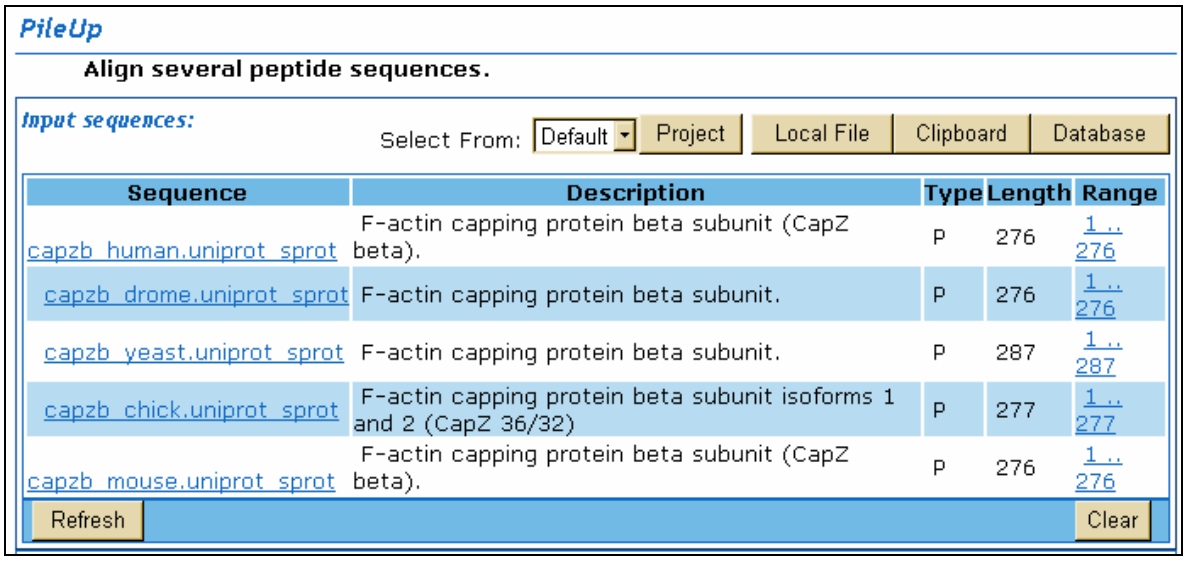
BestFit 的分析結果

以相似度較高的區域比對，相異性太大的部分將不列出

二、PileUp：多序列並列分析

PileUp 程式，是將一群序列進行比對，並得到一個好的比對結果。若要進一步呈現這群比對好的序列之“共有序列”(conserve sequence)時，還必須多執行一次 Pretty 程式。但在 SeqWeb 中，其實可以跳過 PileUp，直接執行 Pretty。

要注意的是：PileUp 是使用 **global alignment** 的比對方法，因此用來分析的序列間必須，最好有一定程度的相似性，否則會無法得到好的結果或是完全無法執行。



PileUp

Align several peptide sequences.

Input sequences:

Select From: Default | Project | Local File | Clipboard | Database

Sequence	Description	Type	Length	Range
capzb_human.uniprot_sprot	F-actin capping protein beta subunit (CapZ beta).	P	276	1 .. 276
capzb_drome.uniprot_sprot	F-actin capping protein beta subunit.	P	276	1 .. 276
capzb_yeast.uniprot_sprot	F-actin capping protein beta subunit.	P	287	1 .. 287
capzb_chick.uniprot_sprot	F-actin capping protein beta subunit isoforms 1 and 2 (CapZ 36/32)	P	277	1 .. 277
capzb_mouse.uniprot_sprot	F-actin capping protein beta subunit (CapZ beta).	P	276	1 .. 276

Refresh | Clear

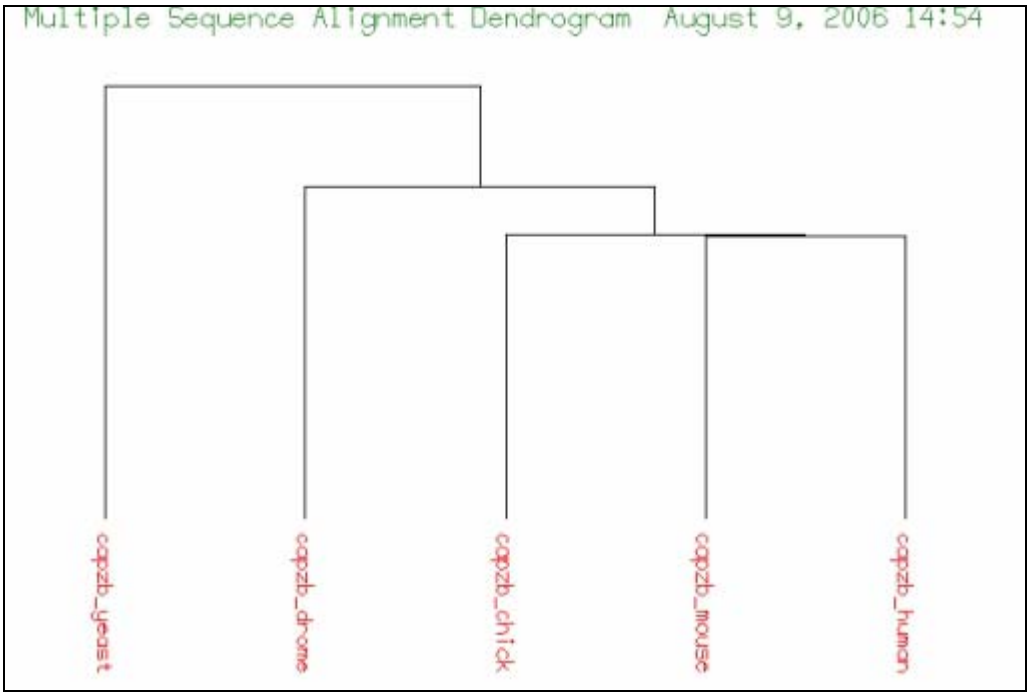
操作時，在 Sequence Manager 中勾選需要分析的序列(三條以上)，再按 run 就能進行分析工作。

```
                1                                                    50
capzb_human ~SDQQLDCAL DLMRRLPPQQ IEKNLSDLID LVPSLCEDLL SSVDQPLKIA
capzb_mouse ~SDQQLDCAL DLMRRLPPQQ IEKNLSDLID LVPSLCEDLL SSVDQPLKIA
capzb_chick MSDQQLDCAL DLMRRLPPQQ IEKNLSDLID LVPSLCEDLL SSVDQPLKIA
capzb_drome MSEMQMDCAL DLMRRLPPQQ IEKNLIDLID LAPDLCEDLL SSVDQPLKIA
capzb_yeast MSDAQFDAAL DLLRRLNPTT LQENLNNLIE LQPNLAQDLL SSVDVPLSTQ

               51                                                   100
capzb_human RDKV.VGKDY LLCDYNRDGD SYRSPWSNKY DP.....PLE DGAMPSARLR
capzb_mouse RDKV.VGKDY LLCDYNRDGD SYRSPWSNKY DP.....PLE DGAMPSARLR
capzb_chick RDKV.VGKDY LLCDYNRDGD SYRSPWSNKY DP.....PLE DGAMPSARLR
capzb_drome KDKE.HGKDY LLCDYNRDGD SYRSPWSNSY YP.....PLE DGQMPSERLR
capzb_yeast KDSADSNREY LCCDYNRDID SFRSPWSNTY YPELSPKDLQ DSPFPSAPLR

               101                                                  150
capzb_human KLEVEANNAF DQYRDLYFEG GVSSVYLWDL D......HGF AGVILIKKAG
capzb_mouse KLEVEANNAF DQYRDLYFEG GVSSVYLWDL D......HGF AGVILIKKAG
capzb_chick KLEVEANNAF DQYRDLYFEG GVSSVYLWDL D......HGF AGVILIKKAG
capzb_drome KLEIEANYAF DQYREMYYEG GVSSVYLWDL D......HGF AAVILIKKAG
capzb_yeast KLEILANDSF DVYRDLYYEG GISSVYLWDL NEEDFNGHDF AGVVLFKK..
```

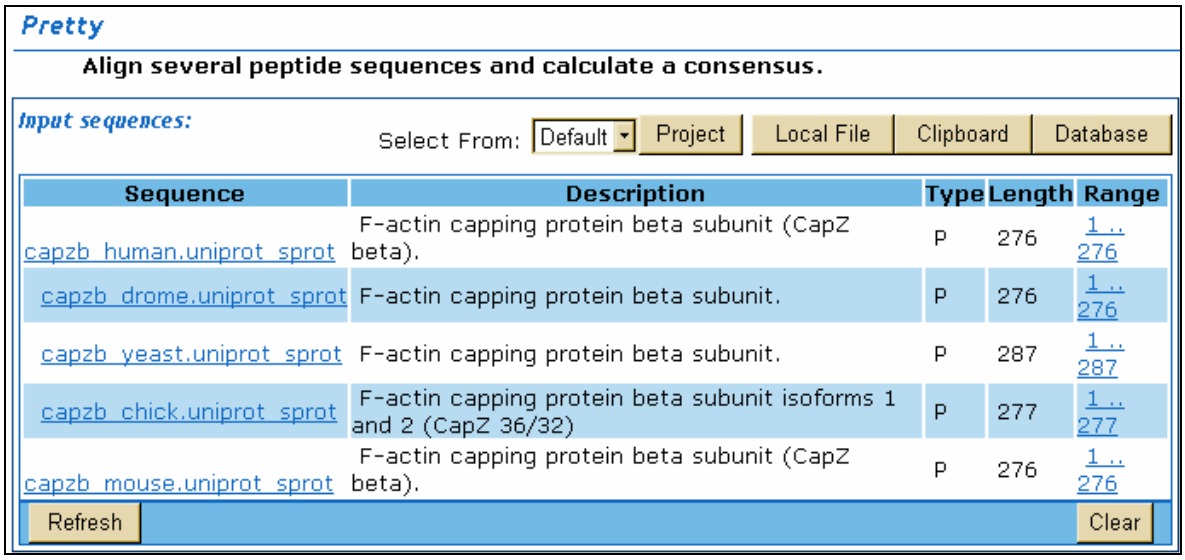
在 SeqWeb 中 PileUp 的結果中會以彩色字形顯示排序結果，另一方面這也代表著相同與相似的胺基酸 residue 可以較清楚的分辨出來。但如果使用者不想顯示彩色字形，可以至 Preference Manager 中再做調整。



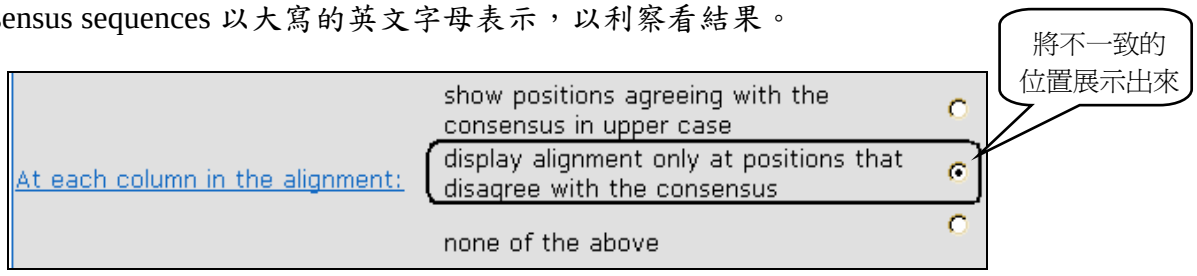
PileUp 輸出結果還包括了如上方的樹狀圖，但那實際上是表示程式在執行時，比對各別序列的順序及方式，雖然很像進行演化分析的 phylogenetic tree，但**實際上並不是一個正確的演化樹圖**，敬請注意及避免混淆。

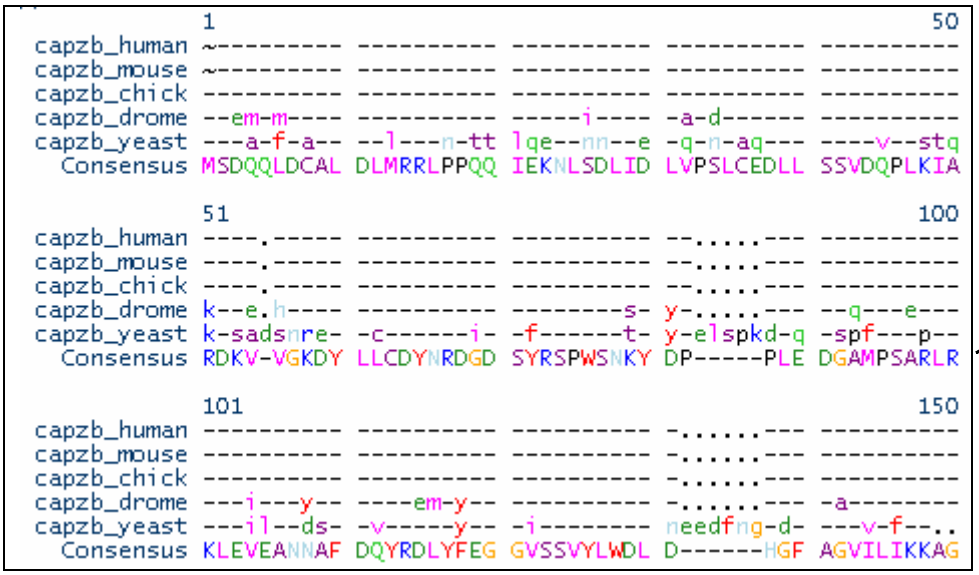
三、Pretty：找出 Consensus sequence

Pretty 這個程式可以用來找出某一群具相似性的序列間的 conserved region。在執行時每個選項都要留意，若選擇不同，出來的結果也會有不同的差異。此處選用預設值進行分析，並觀察結果：

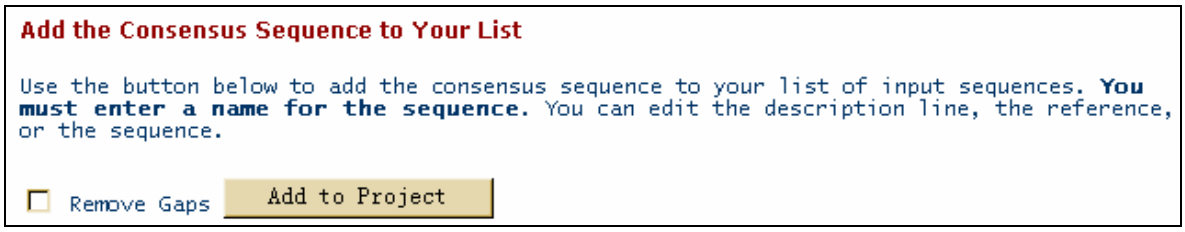


如果使用者希望較容易分辨出 conserved region 的部分，建議選擇「display alignment only at positions that disagree with the consensus」的選項，此功能可將相同的 residue 以橫線(-)符號作為標記；相異的 residue 以小寫英文字母做區分，最後再將 consensus sequences 以大寫的英文字母表示，以利察看結果。





在得到結果之後，使用者可以依照自己的所需，將 Pretty 做出的 consensus sequences 之結果加入至 Sequence Manager 之中，以備將來察看。



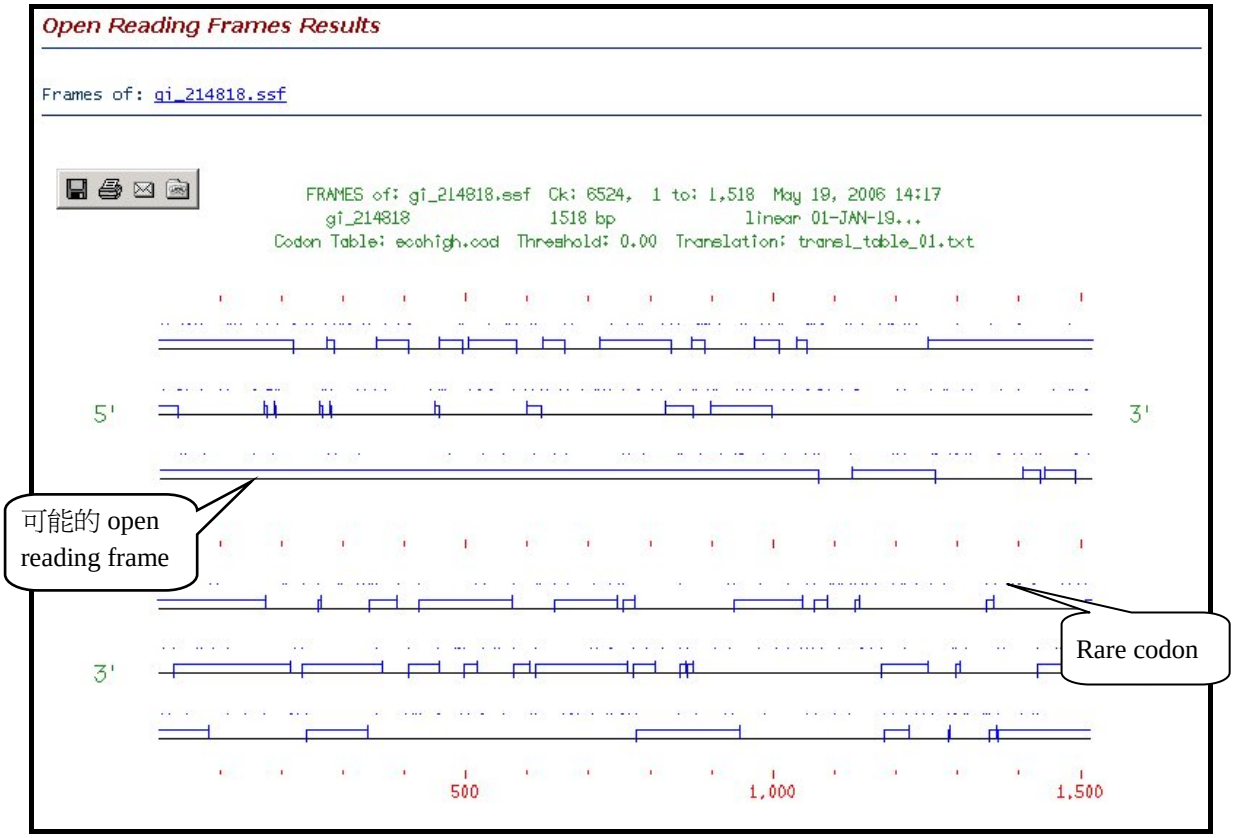
因為在 SeqWeb 中 Pretty 可以不先經過 PileUp 分析而直接跑，但在 GCG Command Mode 卻不行！因此在 GCG 要進行 Pretty 程式之前，一定要先用 PileUp 的程式跑過，其 output 的結果為一個 msf 的檔案格式，它才能作為 Pretty 程式的 input file！另外 GCG Command Mode 中也可以使用另一個的程式-- PrettyBox，此程式可將 conserved sequence 標定起來，並以 “*.ps”的格式輸出圖形，在結果的呈現上較為美觀。

捌、尋找 ORF 及圖譜

若想找出一段 DNA 序列中可能 Coding 蛋白質的區域，並將這段序列轉譯成蛋白質序列，在 SeqWeb 中必須連著執行三個程式。

一、Frames：尋找 Open Reading Frame

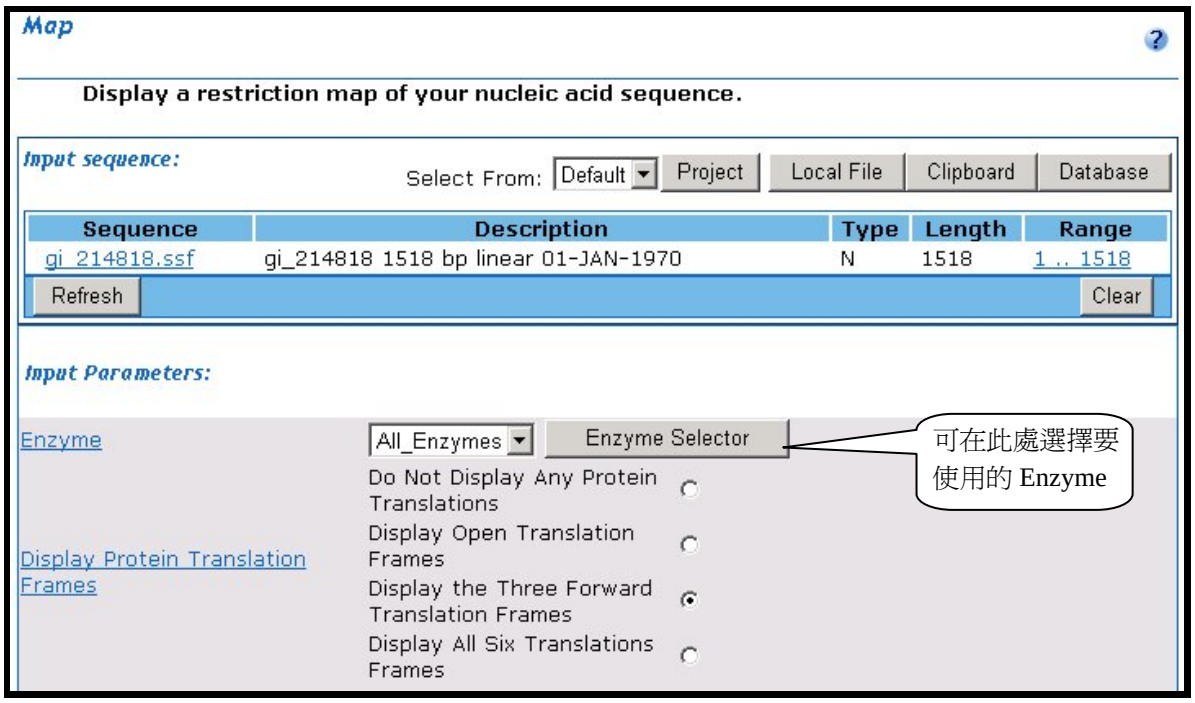
在核酸定序之後，可以利用 Frames 來尋找序列中可能的 open reading frame，所得的結果除了可以顯示六個 Frames 中可能的 ORF，同時也可列出 codon usage 的情形，在 ORF 上面所出現的小點，表示有 rare codon 的出現，當 rare codon 出現的多時，這個 ORF 的可信度相對的較低。此外，Frames 只能顯示出 ORF 大約的位置，要真正找出 ORF 開始和結束的正確位置，還是要參考 map 的結果。



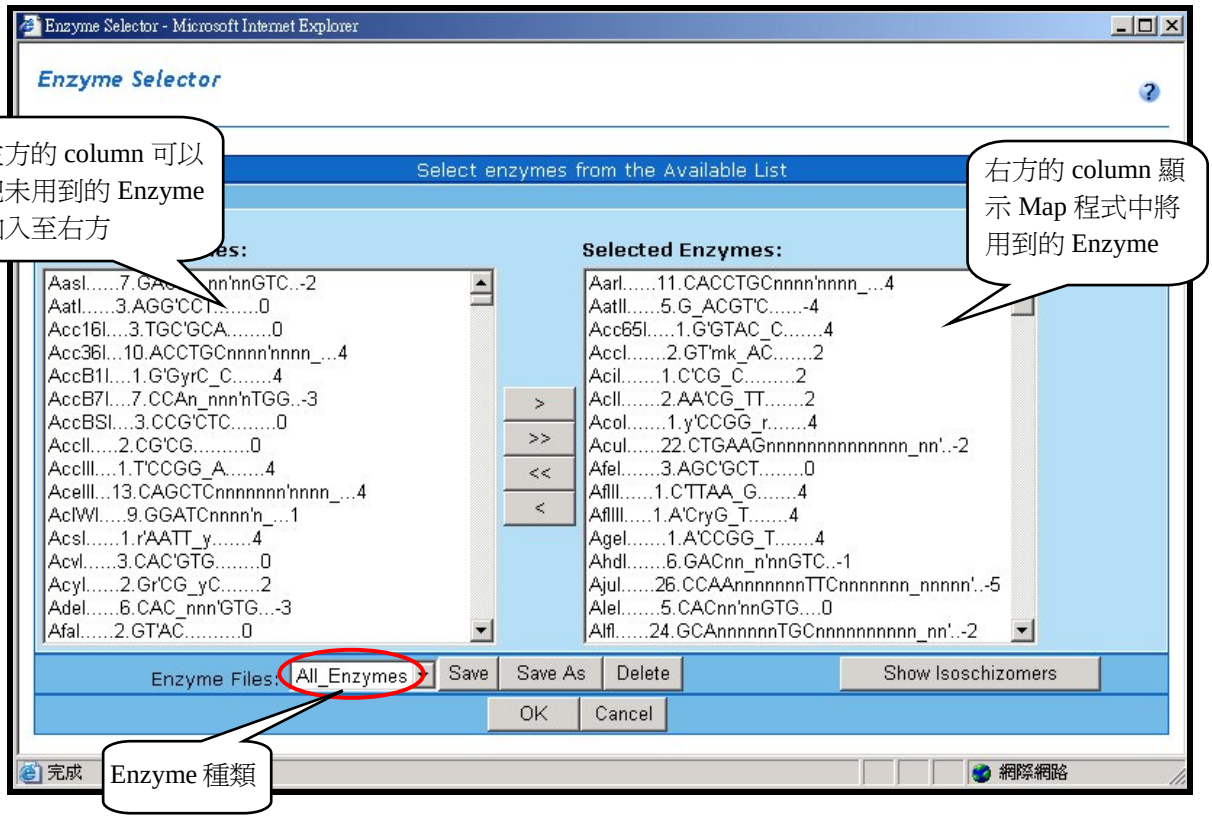
二、Map：尋找圖譜

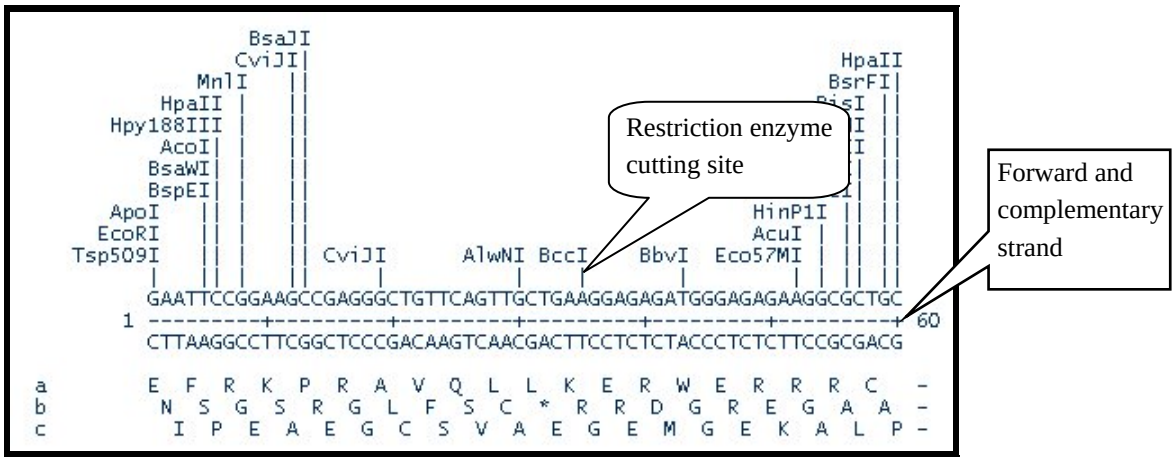
Map 除了可以協助找到序列中可能的限制圖譜(restriction map)之外，因為可以同時顯示六個 reading frame 轉譯出的結果，也可以和 frames 的結果互相參考而找到正確的 start 和 stop codon。

若是選擇做蛋白質序列的圖譜，做出來的會是 protease 的切位圖譜。



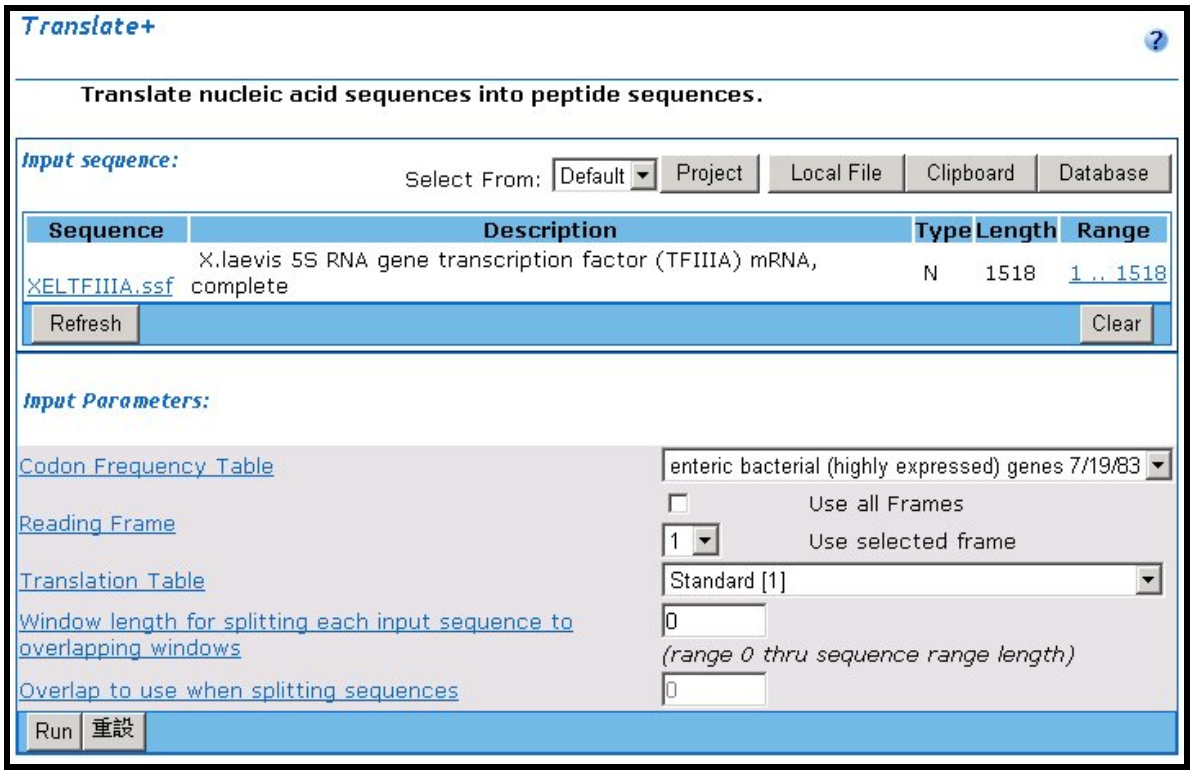
若是使用者按下“View Chosen Enzyme”的按鈕，將開啟新視窗，並於右方欄位顯示程式將使用的 Enzyme 種類；此外，也可以由左方的欄位加入其他的 Enzyme。但若是使用者有自己的特殊 Enzyme 時，就無法加入至選單裡，因此算是一種限制。





三、Translate+：轉譯序列

在找出可能的 ORF 後，可以將其 translate 為蛋白質序列，以進一步分析其可能的 motif 或預測其性質及二、三級結構。在 SeqWeb 之 Sequence Manager 中使用 function 的功能也可以得到 translated 的核酸序列。Translate 在 command mode 中可以直接以核酸序列來執行程式，只需輸入 start 及 stop 即可，還可以直接轉譯 reverse 序列。在 SeqWeb 中，由於是直接從 Sequence Manager 中取得序列，因此對於一段本身就是 coding sequence 的序列而言，將可以得到很好的 translated 蛋白質序列。另一方面，若是序列為使用者本身的 novel sequence，若要獲得正確的 translated 蛋白質序列，就必須先知道真正的 coding region (以 Map 的程式尋找)，再以手動方式修改序列內容，並另存新檔後分析。



若是要 translate 的序列屬於 GenBank 或 EMBL 格式，則以 Translate 程式分析完後，將出現兩個結果：其一為整個序列直接 translate；其二則是從序列原來的基本資料中已知的 coding region 進行 translate，因此後者得到的才是正確的蛋白質序列。

此外還有一些要注意的事項：如果使用者要分析的是 **complementary** 的序列時，必須以 Reverse 程式將其反轉成正向的序列，才能得到正確的蛋白質序列。

練　習：透過實例學習 SeqWeb

實例 1、請逐步操作,填寫下列問題, 找到 Xenopus borealis (Kenyan clawed frog), Xenopus laevis (African clawed frog), yeast, Bufo americanus (American toad),及 Rana pipiens (Northern leopard frog)這幾個 species 的 TFIIIa protein sequence，並將 sequence files 存入自己的 project 中。

1. 進入 LookUp 程式，在"Database"欄位選擇 Uniport，然後在 Alltext 欄位鍵入 "TFIIIa", 接受其他欄位的 default 值,然後按 Run in background,稍候 Job Manager 畫面出現,確認剛才執行的 StringSearch 已完成，點選 view 看結果.

   ＊＊請問您是否查到 7 筆資料？請填寫下列 protein 的 accession number.

   Uniprot:Tf3a_Bufam_____ Uniprot:Tf3a_Ranpi_____

   Uniprot:Tf3a_Xenbo_____ Uniprot:Tf3a_Xenla_____

   Uniprot:Tf3a_Yeast_____

2. 請在結果網頁勾選上一題的五個 TFIIIa protein，點選"Add selected"，按 Close 關閉視窗，回到 SeqWeb 首頁，進入 Sequence manager，按 Edit，選擇 refresh，確認這些檔案已經加入，即完成。

   ＊＊ 您也可使用 StringSearch 程式尋找 protein sequences，由 SeqWeb 首頁進入 StringSearch 程式，在"String to search"欄位鍵入 "TFIIIa", Search set 選擇 protein:swissprot, 點選"search definition line only", 再點選"Find Entries with all of the specific patterns", "include documentation in the output file,並設定 100 為"width of documentation in the output file",然後按 Run in background,稍候 Job Manager 畫面出現,確認剛才執行的 StringSearch 已完成，點選 view 看結果.

   ＊＊ 您也可以去 ExPASy （http://tw.expasy.org/）尋找 protein sequences，先在 database 欄位選擇"**SWISS-PROT and TrEMBL**"，然後在檢索欄位 key in "TFIIIa"，再按"quick search"即可。
   請問:在 **SWISS-PROT** 是否也查到 7 筆資料？請填寫在 **TrEMBL** 您查到幾筆？請觀察比較您在 **TrEMBL** 和在 **SWISS-PROT** 查到的資料,最大的不同在哪裡？

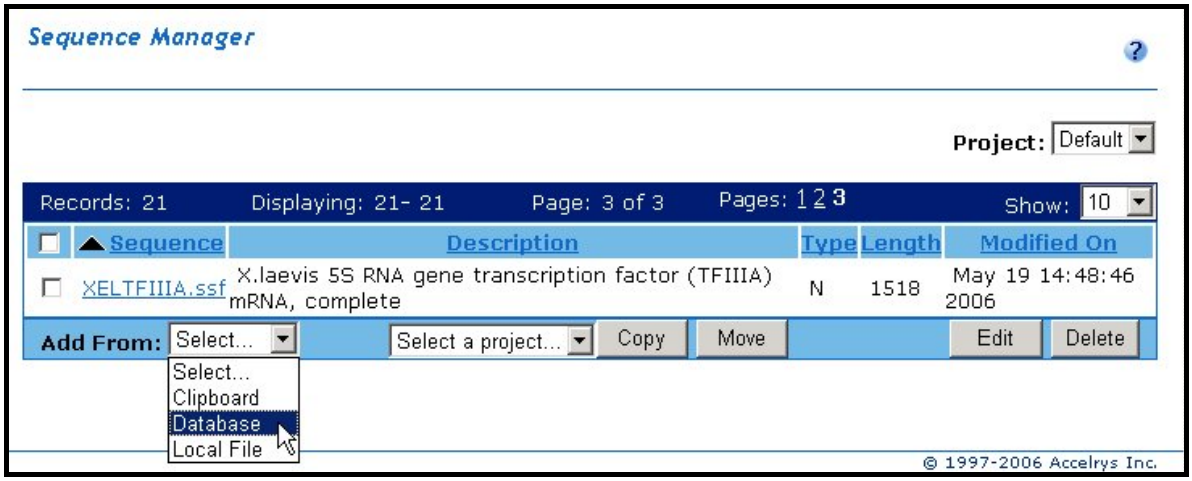
   ＊＊在 ExPASy 查到序列後，您可以點選序列名稱的 hyperlink，在 Sequence Information 欄位點選 FastA format 的 hyperlink，將序列以 copy-paste 的方式用記事本（notepad）存成純文字檔，然後再進入 SeqWeb 的 sequence manager，將序列檔案以 Add from local file/refresh 的方式加入您在 SeqWeb 的 project 中。
   請注意由於 SeqWeb 會自動將 FASTA format 的檔案轉檔為 GCG format,檔名會因此改變,您可再將檔案 rename 回自己所需的檔名。麻煩多了，不是嗎？不過，這些步驟您如果熟悉了，將來如有需要去一些特別的資料庫擷取 sequence file，會用得著。

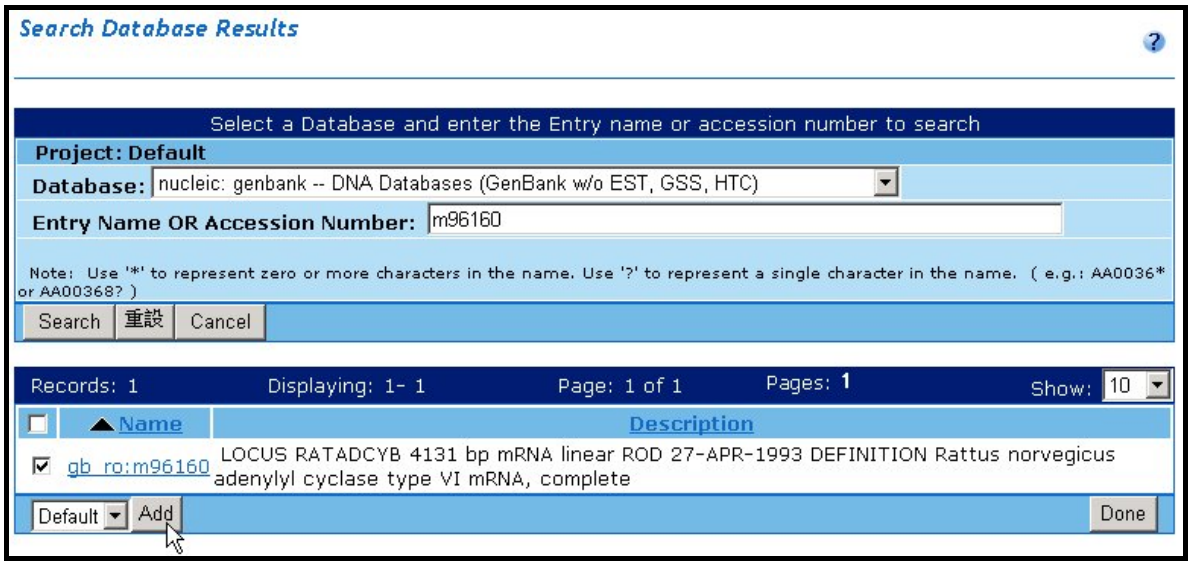
實例 2、請利用 SeqWeb 找出兩條序列，一條的 accession number 是 M96160（Mouse Adenylyl Cyclase type VI）當作是 insert，另一條是 pGEM vector（Accession number 是 X65313）。請將 insert 中一段大約 740 個鹼基對（1271-2010）接到 pGEM vector 的 multiple cloning site，以便爾後實驗大量複製此段序列。做完剪接請用 mapplot 印出結果。

建議解答：

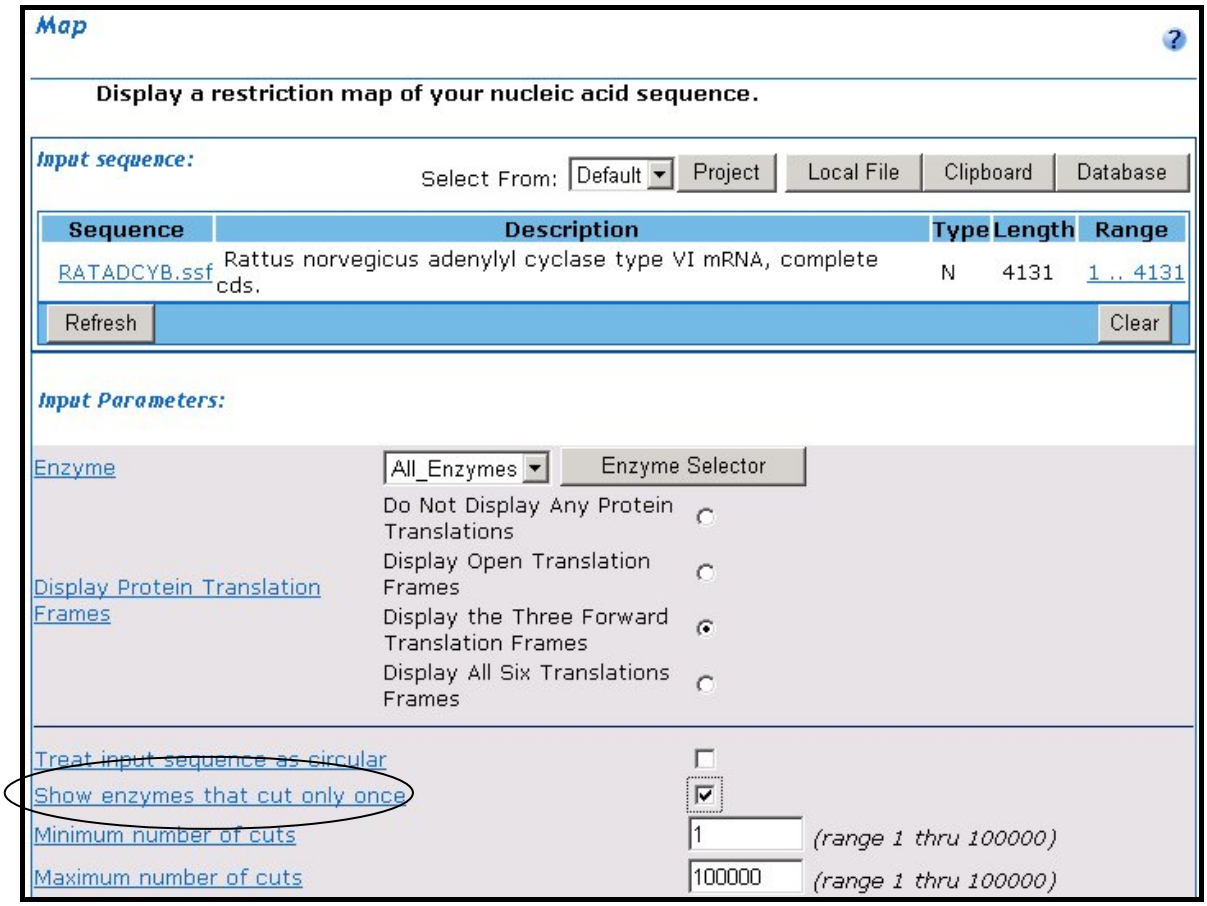
步驟一、先在自己的帳號中的 sequence manager 抓到兩條已知 accession number 的核酸序列。如下圖：

進入 Sequence manager (請參考本講義 p.8)，在 Sequence Manager 中從資料庫中加入序列

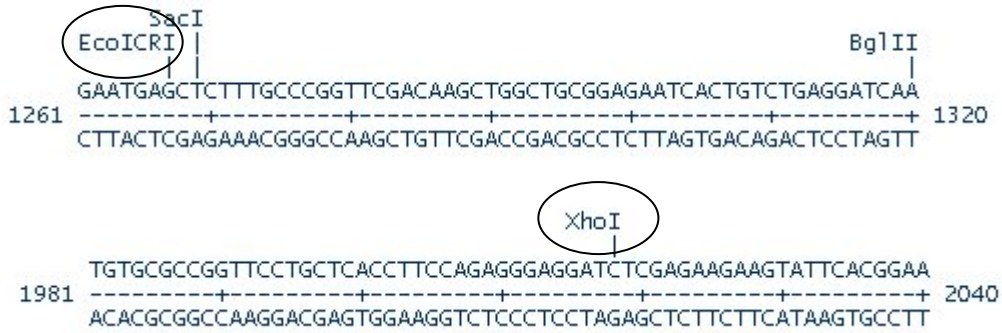




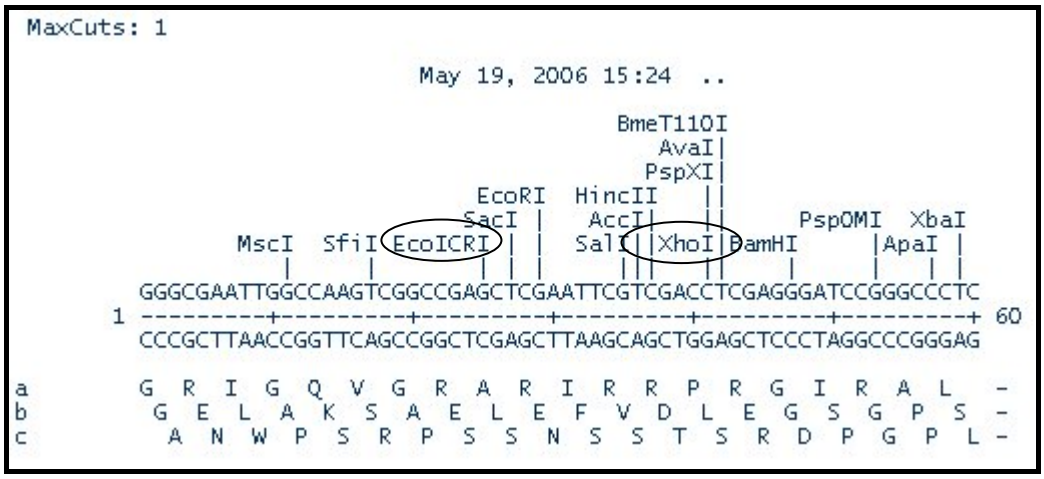
步驟二、使用 map 指令，加上"只切一刀"的參數（"只切一刀"可以使 insert 固定方向性），找出 insert（M96160）在 1271-2010 中可用的核酸限制酵素截切部位。



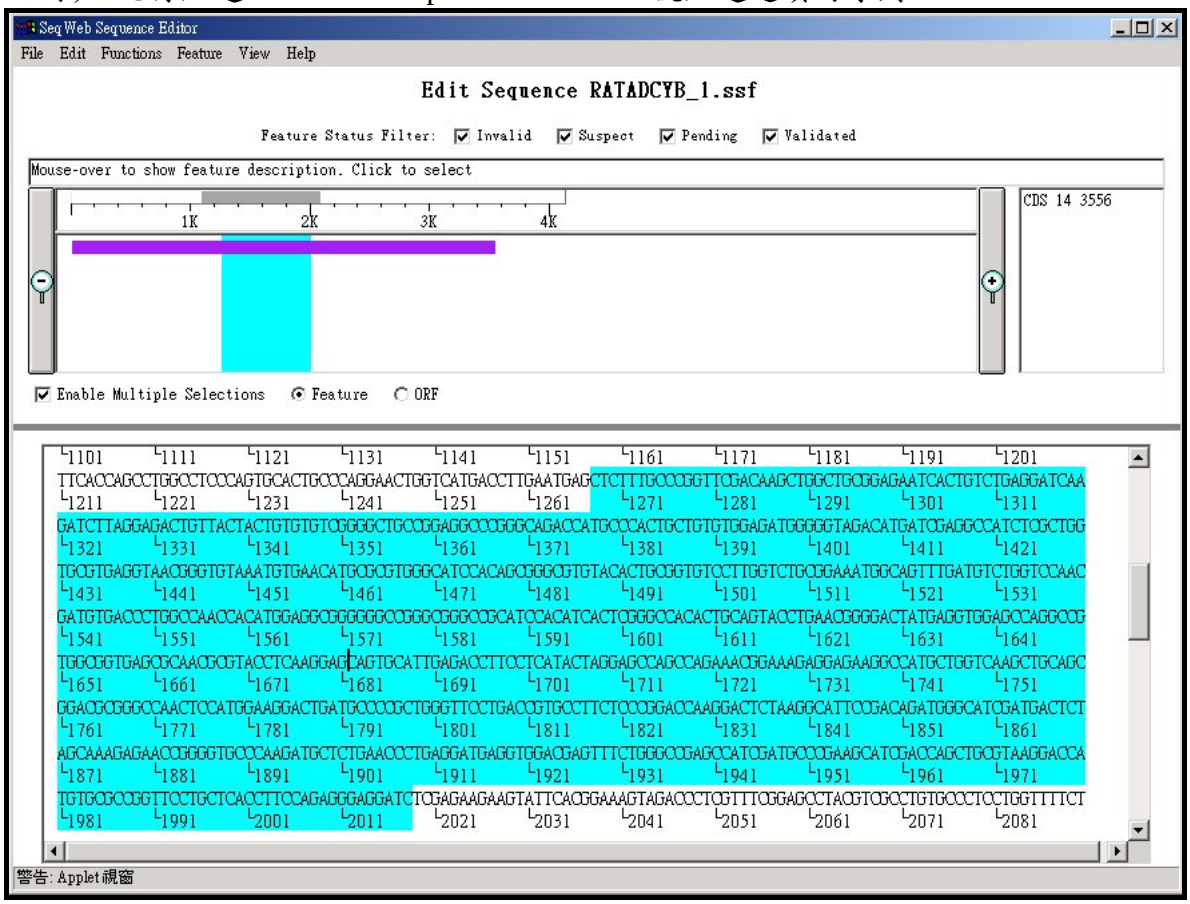
insert 的第一個核酸限制酵素是 EcoICRI，第二個核酸限制酵素是 XhoI



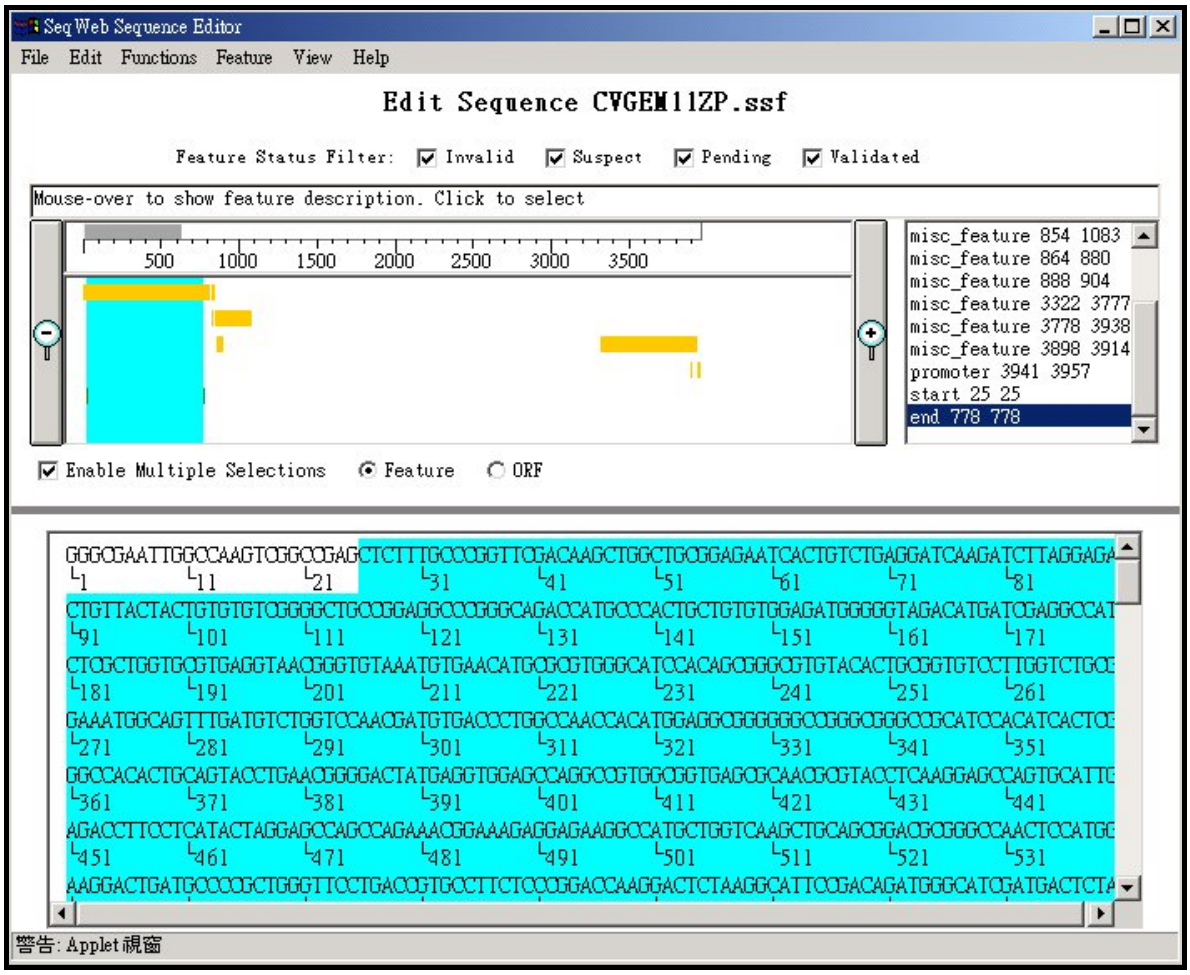
步驟三、使用 map 指令，加上”只切一刀”的參數，找出 pGEM vector（X65313）在 multiple cloning site 中可用的核酸限制酵素截切部位，結果發現也有 EclI36II 和 XhoI，這樣 cloning 就沒問題了。



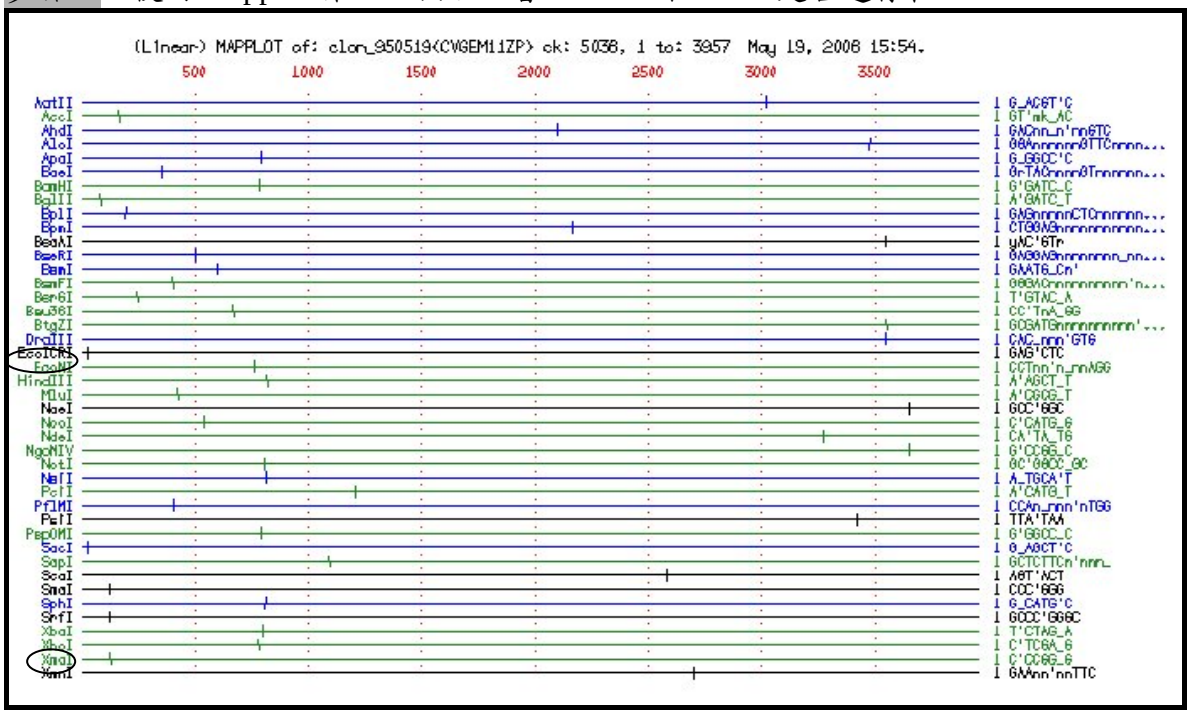
步驟四、在 sequence manager 編輯 M96160，複製適當部位的序列（選 Ec1136II 到 XhoI 之間），記得點選”Enable Multiple Selections”以便點選連續的序列。



在 sequence manager 編輯 X65313，貼到 pGEM vector 的適當位置，貼上後去除不必要的鹼基，另存新檔。



步驟五、使用 mapplot 作一次驗證，看 EcII36II 和 XhoI 是否還存在。



參考資料

1. 國家衛生研究院研究資源週 巨分子序列分析研習會講義 楊永正老師主編
2. 國家高速電算中心 生物資訊學初、中、高級課程講義 楊永正老師主編
3. 一天學好 GCG 入門 楊德勳編著
4. Baxevanis A. D., and B. F. F. Ouellette (1998). Bioinformatics, A practical guide to the analysis of genes and proteins. Wiley-Interscience Publication. New York. 370pp.
5. Bishop M. J. and C. J. Rawlings (1997). DNA and Protein Sequence Analysis. IRL Press. New York. 352pp.
6. Griffin A. M. and H. G. Griffin (1994). Computer Analysis of Sequence Data part I. Humana Press. New Jersy. 372pp.
7. Shpaer E G. Robinson M. Yee D. Candlin J D. Mines R. and T. Hunkapiller (1996) Sensitivity and selectivity in protein similarity searches: A comparison of Smith-Waterman in hardware to BLAST and FASTA. Genomics 38(2). p179-191.
8. Setubal J. and J. Meidanis (1997) Introduction to Computational Molecular Biology. PWS Publishing Company 296pp.

Useful Links:

1. 基因體醫學生技研發生物資訊核心(GMBD Bioinformatics Core)網站：http://www.tbi.org.tw
2. 國家衛生研究院生物資訊首頁：http://bioinfo.nhri.org.tw
3. GCG User Manual：http://bioinfo.nhri.org.tw/gcg/doc/11.0/gcghelp.html
4. NCBI：http://www.ncbi.nlm.nih.gov
5. EBI：http://www.ebi.ac.uk/
6. DDBJ：http://www.ddbj.nig.ac.jp/
7. ExPASy：http://tw.expasy.org/
8. Uniprot：http://www.expasy.uniprot.org/

📖 SeqWeb 3.1.2 講義編輯

李桂玉　主任

汪詩海　先生

王旭川　女士