

壹、國家衛生研究院生物資訊服務簡介	
一、前言	
貳、SeqWeb使用說明	5
-  Login SeqWeb	
二、SeqWeb 主網頁簡介	6
三、Sequence Manager功能	7
四、Job and Result Manager	
五、Preterence Manager	
<b>参、系統、資料庫及程式簡介</b>	
一、系統及資料庫簡介	
二、GCG核酸及蛋白質資料庫	
三、SeqWeb程式簡介	
<b>肆、臣列牧士销</b> 众	20
年、戶列俗式間介	20
一、广列俗式裡類 二、SeaWeb所齡出的過安刑式	
二、pedineo/// 抽页相关主式	
伍、以文字搜尋資料庫	
一、StringSearch:以字串尋找所要的序列	
二、LookUp:以keyword尋找所要的序列	
陸、以序列搜尋比對資料庫	
一、序列比對分析基本概念	
二、BLAST程式操作	
法、名应列并列入长	17
ホ、タケク 业 列 カ 机 ·································	
- DesungOAI · 支厅列亚列刀柯	
三、Pretty:找出Consensus sequence	
· 潮、尋找ORF及圖譜	
ー、Frames: 尋找Open Reading Frame	
二、Map·尋找圖譜	
ニ、Iransiale+・ 轉译/予列	
練 習:透過實例學習SeqWeb	
附件	63
參考資料	65

NHRI SeqWeb3.1 講義 v1.0

# **壹、國家衛生研究院生物資訊服務简介**

### 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。

#### ■ 分析工具之提供

#### A. Blast 2 序列搜尋服務 (Telnet blast2.nhri.org.tw)

BLAST 序列搜尋軟體會根據使用者提供的序列與指定的資料庫做搜尋。使用者可藉此做 DNA 和蛋白質的比對分析。目前於 BLAST 2 主機所安裝的乃是 美國 National Center for Biotechnology Information (NCBI)於 1997 年提供的 BLAST 2.0 版序列分析工具。

### B. 提供「生物資訊資料分送站」NHRI Bioinformatics FTP Server

為增進國內生物資訊發展,解決下載國外資料頻寬不足之問題,本院特別設置本生物資訊資料分送站,本資料分送站是以收集生物資訊相關軟體及資料庫為 目標,初期先以收集常見之重要生物資訊資料庫為主,並定期將最新資料放置於 本站,提供國內研究人員申請下載使用。

#### ■ 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,免於取道國際網路臃塞的限制。

- A. The Genome Database (GDB)基因體資料庫鏡相站
- B. Protein Data Bank (PDB) 蛋白質結構資料庫鏡相站 (與清大合作)
- C. ExPASy 蛋白質資料庫鏡相站
- D. FlyBase 果蠅生物資訊資料庫鏡相站
- E. Mouse Genome Informatics (MGI)老鼠生物資訊資料庫鏡相站
- 國內資料庫的建置與維護
  - A. Chinese Gene Mutation Database「中國族群基因突變資料庫」(與陽明大學 合作建置)
  - **B. NHRI Liver EST Database**

# 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,除了解決舊版操作尚的一些問題之外,也更新部份功能及程式。

SeqWeb v3 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! SeqWeb 中包含了多數使用者最常用的程式,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。

### - Login SeqWeb

SeqWeb v3 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 (<u>http://v8803.nhri.org.tw:8003/mgr.shtml</u>) 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」(<u>http://www.tbi.org.tw</u>) 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。

<u>Login 畫面</u>	連線到 v8803.nhri.org.	tw	?×
	Seq Web 使用者名稱(U): 密碼(P):	gusername password	<u> </u>
		□ 記憶我的密碼(R) 確定	取消

# 二、SeqWeb 主網頁簡介

Login 之後即可看到如下的 SeqWeb 主網頁。

🚰 Seq Web - Web based sequence	analysis tool - Microsoft Intern	net Explorer		_ 🗆 ×			
檔案(F) 編輯(E) 檢視(V)	我的最愛(A) 工具(T)	說明(H)					
🗲 上一頁 🔹 📀 🖌 💌	🔁 🏠 🔎 搜尋 🦻	≿ 我的最愛 🥝 🔗	- 🕹				
網址(D) 🛃 http://v8803.nhri.or	g.tw:8003/mgr.shtml			🔹 🔁 移至 🛛 連結 👋 📆 🗸			
SeqWeb v <sub>3.1</sub>	程式選項	Manager 功能選項					
	Programs	Managers		Help Topics   Support			
Managers Proiect		程式工作區	<u>1</u> <u>1</u>				
Sequence Job	Project Manager A project is where sequence files and their associated result files are stored. Using the Project Manager you can create, modify or delete a project. To create a project, you must be "project enabled". All users have a 'Default' project.						
Preference 選單區	Sequence Manager Sequence files are stored in a project. Using the Sequence Manager you can add sequence file(s) to a project or delete sequence file(s) from a project. The Sequence Manager also allows you to copy or move sequence file(s) between projects.						
	<b>Job Manager</b> When an analy manages these 'saved'.	vsis program is run, this e jobs. The Job Manage	; creates a job. The er has two views - '	Job Manager submitted' and			
🙆 Accelrys Inc Software for Pl	narmaceutical, Chemical, and M	laterials Research		🥂 🥝 網際網路			

### ■ 程式工作區

在主網頁右邊是各程式之主要工作區,在開始時選擇了那一種分析程式,就會在 程式工作區呈現該所有分析程式的內容與選項。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<u>核酸</u>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。

<u>Locally align two nucleic acid sequences.</u>
 (核酸序列分析)
 <u>Locally align two peptide sequences.</u>
 (蛋白質序列分析)

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 使用手册及 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

SeqWeb v3 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 SeqWeb v3 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。

SeqWeb v3 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 GCG 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 SeqWeb v3 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。

SeqWeb v <sub>3.1</sub>						2	(Sjac	celrys®
	Pr	ograms	Mana	gers			lelp Topi	cs   Support
Managers	Sequ	ence Manage	er					?
Project Sequence							Projec	t: Default 💌
Job	Rec 59	ords: Disp	laying: 1- 1	O Page: :	1 of 6 Pages: <b>1</b> <u>2</u> <u>3</u>	456	Sh	now: 10 💌
Preference		▲ <u>Sequ</u>	ience	1	Description	Тур	<u>e Length</u>	Modified On
		<u>af123456</u>		Influenza (A/Chicke (H5N1))	A virus n/Hong Kong/y388,	/97 N	1726	May 8 10:10:35 2006
		<u>af144305</u>		Influenza (A/Goose/ (H5N1)) h	A virus 'Guangdong/1/96 emagglutinin (HA)	N	1760	May 8 10:10:35 2006
		<u>ay618086</u>		ay618086	5	N	432	May 8 10:10:35 2006
		<u>capb_bovin.ur</u>	iiprot sprot	F-actin c subunit (C	apping protein bet CapZ beta).	ар	301	May 8 10:10:35 2006
		capb_chick.un	iprot sprot	F-actin c subunit is 36/32)	apping protein bet oforms 1 and 2 (Ca	a apZ P	277	May 8 10:10:35 2006

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:

	<u>capb h</u> i	uman.uniprot	<u>sprot</u>	F-actin cappiı beta subunit (	ng proteii CapZ bet	n a).	Ρ	276	May 8 10:10:35 2006
	<u>capb m</u>	ouse.uniprot	<u>sprot</u>	F-actin cappi beta subunit (	ng protei CapZ bet	n :a).	Р	276	May 8 10:10:35 2006
	<u>capb y</u>	east.uniprot s	prot	F-actin cappiı beta subunit.	ng protei	n	Ρ	287	May 8 10:10:35 2006
	<u>capzb t</u>	oovin.uniprot	<u>sprot</u>	F-actin cappiı beta subunit (	ng protei CapZ bet	n :a).	Р	301	May 8 10:10:35 2006
Ad	d From:	Select 💌	Sele	ct a project 💌	Сору	Mov	e	Edit	Delete
		Select Clipboard Database Local File					© 1	997-2006	6 Accelrys Inc.

# 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
- 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。
- 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!

	Programs	Managers			Help T	opics   Support
Managers	Sequence Manage	er				?
Project Sequence	檔名不同時 1	,建議以 Modified On 怎	\$排序較容易找出檔	案 へ	Proje	ect : Default 🔻
Job	Records: Dis	playing: 1-61 F	Page: 1 of 1 Page	s: 1		Show: 200 🔽
Preference	Seque	nce l	<u>Description</u>	Туре	Length	Modified On
	gi 202718.ssf	gi_2027 bp linea	18 M96160 4131 r 01–JAN–1970	N	4131	May 29 17:03:32 2006
	EMBOSS 3393	© EMBOSS B.ssf linear 01	6_33933 154 bp -JAN-1970	N	154	May 29 16:13:28 2006
	D HD HUMAN.ssf	HD_HUM JAN-197	1AN 3144 aa 01- 0	Ρ	3144	May 23 09:49:45 2006
	🗖 <u>101115.gb ro</u>	101115		N	6036	May 11 18:19:12 2006

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。
- 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:ABCDEFGHIKLMNPQRSTUVWXYZ(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。
- > 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。

Search Database	Results			3
Selec	t a Database and en	ter the Entry name or a	ccession number	to search
Project: Default				
Database: nuclei	c: genbank DNA Data	abases (GenBank w/o EST,	GSS, HTC)	
Entry Name OR #	Accession Number:	ay069515 🛛 🛶 輸入關	<b>鍵字(Accession</b> )	aumber或 Entry name)
Note: Use '*' to repre name. ( e.g.: AA0036	esent zero or more chara * or AA00368? )	acters in the name. Use '?' to	represent a single	character in the
Search 重設 Ca	ancel /1	<b>搜尋結果於下方呈現</b>		
Records: 1	Displaying: 1- 1	Page: 1 of 1	Pages: 1	Show: 10 🔽
□ ▲ <u>Name</u>		Descrip	otion	
☑ gb in:ay06951	15 LOCUS AY069515 melanogaster LD23	1946 bp mRNA linear IN' 3533 full length cDNA. Ad	V 17-DEC-2001 D CCES	EFINITION Drosophila
Default Add	← 加入勾選序列	ի		Done

Search Database Results	4
Select a Databa	se and enter the Entry name or accession number to search
Project: Default	
Database: protein: uniprot U	IniProt (SWISS-PROT plus Translated EMBL)
Entry Name OR Accession	Number: [capzb_* 🛛 🖛 輸入關鍵字(Accession number或 Entry name
Note: Use '*' to represent zero or name. (e.g.: AA0036* or AA00368	more characters in the name. Use '?' to represent a single character in the ? )
Search 重設 Cancel	ノ捜尋結果於下方呈現
Records: 10 Displayir	g: 1- 10 Page: 1 of 1 Pages: 1 Show: <mark>10 _</mark>
■ <u>Name</u>	Description
🗖 uniprot sprot:capzb arat	DE Probable F-actin capping protein beta subunit (CapZ-beta). GN OrderedLocusNames=At1g71790; ORFNames=F14
🔽 uniprot sprot:capzb_chic	DE F-actin capping protein beta subunit isoforms 1 and 2 (CapZ 36/32) DE (CapZ B1 and B2) (Beta-actinin su
🗖 uniprot sprot:capzb_dicc	DE F-actin capping protein beta subunit (CAP32). GN Name=acpA; Synonyms=abpE; OS Dictyostelium discoideum
🔽 uniprot sprot:capzb_droi	DE F-actin capping protein beta subunit. GN Name=cpb; Synonyms=ANCP-BETA; ORFNames=CG17158; OS Drosophila
🗹 uniprot sprot:capzb hum	DE F-actin capping protein beta subunit (CapZ beta). GN an Name=CAPZB; OS Homo sapiens (Human). OC Eukaryota;
uniprot sprot:capzb mou	DE F-actin capping protein beta subunit (CapZ beta). GN Name=Capzb; Synonyms=Cappb1; OS Mus musculus (Mous
Default 🔥 🛶 加入	勾選序列 Done Done

### ■ 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。

capb_drome.uniprot_sprot_subunit.	Ρ	276	10:10:35 2006
Capb human.uniprot sprot subunit (CapZ beta).	Ρ	276	May 8 10:10:35 2006
□ <u>capb mouse.uniprot sprot</u> □ <u>capb mouse.uniprot sprot</u> □ <u>capb mouse.uniprot sprot</u>	Р	276	May 8 10:10:35 2006





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

<u>步驟</u>:

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
- 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
- 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



!!AA_SEQUENCE 1.0       Text View         WPDEF       F-actin capping protein beta subunit (CapZ beta).         ID       CAPB_HUMAN       STANDARD;       PRT; 276 AA.         AC       P47756; 08TB49; 09NUC4;       SText View"
WPDEF F-actin capping protein beta subunit (CapZ beta). ID CAPB_HUMAN STANDARD; PRT; 276 AA. AC P47756: 08TB49: 09NUC4: <b>點選 "Text View"</b>
ID       CAPB_HUMAN       STANDARD;       PRT;       276 AA.         AC       P47756;       08TB49;       09NIIC4;
AC P47756: 08TB49: 09NIIC4: 點選 "Text View"
DT 01-FEB-1996 (Rel. 33, Created)
DT 10-OCT-2003 (Rel. 42, Last sequence update)
DT 05-JUL-2004 (Rel. 44, Last annotation update)
DE F-actin capping protein beta subunit (CapZ beta).
GN Name=CAPZB;
OS Homo sapiens (Human).

儲存網頁				? ×
儲存於(I):	🕝 点面		- 🕑 🔊 📂 🖽-	
<ul> <li>表最近的文件</li> <li>表面</li> <li>美面</li> <li>我的電腦</li> <li>一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、一、</li></ul>	→ 我的文件	請自行決 (如 "桌面	定儲存位置 " 或 "D:/" 等)	
	檔名(N): 🧲	pb_human.txt>	記得更改檔案名稱	儲存(S)
	存檔類型(T):	(字檔 (*.bxt) 💙 🗲	選擇 "文字檔" 儲存	取消
	編碼(E): 西	電語系 (ISO)	•	

### ■ 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

 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。

- 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。
- 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

<u>步驟</u>: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

<u>步驟</u>:

- 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<u>核酸序列</u>
- 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 互换得到互補序列,
- 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 序列特徵的美工编輯

步驟:

- 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. 删除序列特徵: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。

Job Manage	r.					
		Proj	ect: All 🗾 Jol	bs: 🖸 Submitte	ed <mark>O</mark> Saved	Refresh
Records: 18	Displayir	ng: 1- 10	Page: 1 of 2	Pages: <b>1</b>		Show: 10 💽
<b></b> <u>Job #</u>	<u>Task</u>	•	Start Time	<u>Run Time</u>	Project	<u>Status</u>
<b>23957</b>	bestfit	Jul 2 19:28	:44 2006	00:00:00	Default	× Failed
<b>1</b> 6375	lookup	Jul 2 19:18	:06 2006	00:00:00	Default	× Failed
16020	clustalw+	Jul 2 19:12	:04 2006	00:07:39	Default	Completed
<b>1</b> 5450	prime	Jun 30-10:5	54:10 2006	00:00:00	Default	× Failed
<u> </u>	lookup	May 22 10:	46:10 2006	00:00:25	Default	Completed
<u>25779</u>	lookup	May 22 09:	46:49 2006	00:00:00	Default	× Failed
Refine Vie	🔪 🗲 🛛 Refin	e: 重作・View	: 看結果	St	op:中止 job	-> Stop

#### 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— 可自行编輯紀錄的 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。

#### 第 17 頁

Job	Manager			
		Project : Default 💌	Jobs: OSubmitted	• Saved Refresh
Rec	ords: 50 Displaying: 1-	10 Page: 1 of 5	Pages: <b>1</b> <u>2</u> <u>3</u> <u>4</u> <u>5</u>	Show: 10 🖃
	<u>File</u>	Descri	<u>ption</u>	▼ <u>Modified On</u>
	<u>k02938 stemlo 22212</u>	StemLoop Results 07/A	ug/2006:10:51:14	Aug 7 10:51:14 2006
	<u>capzb human bestfi 22009</u>	BestFit Results 07/Aug	/2006:10:49:03	Aug 7 10:49:03 2006
	Hong Kong 15 prime 19402	Prime Results 24/Jul/20	06:13:41:44	Jul 24 13:41:44 2006
	Hong Kong 15 map 24847	Map Results 03/Jul/200	6:12:02:14	Jul 3 12:02:14 2006
	<u>capzb human bestfi 24317</u>	BestFit Results 02/Jul/2	2006:19:58:06	Jul 2 19:58:06 2006
	lookup 1927	LookUp Search Results 08/Jun/2006:19:10:34		Jun 8 19:10:34 2006
	lookup 3772	LookUp Search Results 29/May/2006: 14: 12: 50		May 29 14:12:50 2006
	<u>lookup 29850</u>	LookUp Search Results 23/May/2006:18:05:55		May 23 18:05:55 2006
E	dit Refine Se	ect a project 🝷 🚺 Copy	Move	Delete

# 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<u>存檔的格</u> 式。較需注意的是,若是<u>麥金塔電腦(mac)、或是 Uuix 系統</u> (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<u>新密碼登入</u>才 行!



# **參、系統、資料庫及程式简介**

# 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95年5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

Description	GenEMBL (GenBank + EMBL)	GenBank
Entire sequence Database	re sequence GenEMBLPlus:* Database GEP:*	
All database divisions except EST, and GSS sequences	GenEMBL:* GE:*	GenBank:* GB:*
Only EST, and GSS Sequences	Tags:*	GB_Tags:*
Bacterial sequences	Bacterial:* Ba:*	GB_Ba:*
Expressed sequence tag (EST) sequences	EST:*	GB_EST:*
Genome survey sequences (GSS)	GSS:*	GB_GSS:*
High throughput genome	HTG:*	GB_HTG:*
Invertebrate sequences	Invertebrate:* In:*	GB_In:*
Organelle sequences	Organelle:* Or:*	
Non-rodent, non-primate mammalian sequences	Other_Mammalian:* Om:*	GB_Om:*
Non-mammalian, vertebrate sequences	Other_Vertebrate:* Ov:*	GB_Ov:*
Sequences from patents and patent applications	Patent:*	GB_Pat:*
Phage sequences	Phage:* Ph:*	GB_Ph:*
Plant and Fungal Sequences	Plant:* Pl:*	GB_Pl:*
Primate sequences	Primate:* Pr:*	GB_Pr:*

Rodent sequences	Rodent:* Ro:*	GB_Ro:*
Structural RNA sequences (such as rRNAs)	Gb_St:* St:*	GB_St:* St:*
Sequence-tagged site (STS) sequences	STS:*	GB_STS:*
Synthetic sequences (plasmids, vectors)	Synthetic:* Sy:*	GB_Sy:*
Unannotated sequences	Unannotated:* Un:*	GB_Un:*
Viral sequences	Viral:* Vi:*	GB_Vi:*

GenBank 的分類可以分為兩大類,一種是以功能分類的像 EST、GSS、STS、HTG、 Pattern 等,另外則是依 Organism 來分類像 Plant、Invertebrate、Primate 等等。以下僅 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。

- 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。
- 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

Description	PIR-Protein UniProt		GenPept
Entire sequence database	PIR:* P:*	UniProt:*	
Annotated sequences	Protein:* PIR1:*	UniProt_Sprot:*	
Preliminary sequences	PIR2:*	UniProt_trembl:*	
Unverified sequences	PIR3:*		
Unencoded or untranslated	PIR4:*		
GenBank Translated Proteins			GenPept:* Gp:*

蛋白質資料庫是由數個單位自行建立收集的,例如 GenPept,乃是 NCBI 以 GenBank 核酸序列為基礎而建立的;Swiss-Prot 則是歐洲 EBI 組織下的 SIB 所建立的, 資料庫裡包含了最完整的蛋白質序列資訊;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<u>最重要</u>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 uses a profile as a query to search the datable 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 Searchin	lg			
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	)n			
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 compares one or more sequences to a database of p hidden Markov models, such as the Pfam library, in order to id- known domains within the sequences.					
HTHScan scans protein sequences for the presence of helix-turn- motifs, indicative of sequence-specific DNA-binding structures 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 am sequence.					
PeptideStructure PeptideStructure makes secondary structure predictions for sequence.					
SPScan Scans protein sequences for the presence of secretor peptides (SPs).					
* TransMem	Scans for likely transmembrane helices in a peptide sequence.				
Nucleic Acid Secon	dary 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.					
<b>** Utilities</b>	** 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。 NHRI SeqWeb3.1 講義 v1.0

# 建、序列格式简介

# 一、序列格式種類

雖然在 GCG command mode 僅接受 <u>GCG 格式</u>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 <u>ascii code 的純文字檔類型。換言之,若</u>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,<u>純粹只有連續的序列</u>,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 <u>GenInfo Identifier</u>,這是序列檔在變動時會給的一個版本編號。

gb|U49845--這是這個序列的 GenBank 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 除了 A12345 這種格式之外,還有 AB123456 這一種格式。在進行序 列分析時可以以指定 Accseeion number 的方式指定要分析的序列, 例如: Genbank:U49845。要注意的是,在 NCBI 的另一個資料庫 RefSeq 的序列格式和 GenBank 很像但 Assession 格式不同,如 NM\_123456,如果以這個 accession number 到 GenBank 指定就會找 不到序列。

- SCU49845--是這個序列的 Locus name,這原本是設計來 group 同一 locus 的 不同的序列,但因為不敷使用,現在比較不具功能了。
- Saccharomyces ce.....-接下來的這個部份是這個序列的 Definition,是對這條序列的精簡描述。

### ■ 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 SCU	J49845 5028 bp	DNA	PLN	21-JUN-1999
DEFINITION Saco	charomyces cerevisiae T	CP1-beta ger	e, partial cds, and A	xl2p
(AXL	2) and Rev7p (REV7) g	enes, comple	te cds.	-
ACCESSION U4	9845			
VERSION U49	0845.1 GI:1293613			
KEYWORDS .				
SOURCE bak	er's yeast.			
ORGANISM Sa	ccharomyces cerevisiae			
Eukar	yota; Fungi; Ascomycot	a; Hemiascor	nycetes; Saccharomy	cetales;
Sacch	aromycetaceae; Sacchar	omyces.		
REFERENCE 1	(bases 1 to 5028)			
AUTHORS To	rpey,L.E., Gibbs,P.E., No	elson,J. and L	awrence,C.W.	
TITLE Clon	ing and sequence of REV	V7, a gene wł	nose function is requi	ired for
DNA	damage-induced mutage	enesis in Sacc	haromyces cerevisia	e
JOURNAL Yea	st 10 (11), 1503-1509 (1	1994)		
MEDLINE 951	76709			
•••••		••••		
FEATURES	Location/Qualify	iers		
source	15028			
	/organism="Saccha	aromyces cer	evisiae"	
	/db_xref="taxon:4	932"		
	/chromosome="IX			
<b>~ ~ ~</b>	/map="9"			
CDS	<1206			
	/codon_start=3			
	/product="TCP1-b	eta"		
	/protein_id="AAA	98665.1"		
	/db_xref="GI:1293	3614"		
//				
/translation= 551 r in		A D D D T A N D A	ES I KLKKAV VSSA DUM''	SEA
	AEVLLKVDNIK	ARPRIANK	2HM	
gene	08/3138 /aama   AVI 2			
CDG	/gene=AAL2			
CDS	08/3138 /como_"AVI 2"			
	/gelle AAL2	mbrana aluas	protain"	
	/note- plasma men	morane gryco	protein	
	/couoii_start=1 /function="require	d for avial by	dding nattorn of S	
	/iunction= require	u ioi axiai du	using patient of S.	
	cerevisiae			



- 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	Saccharomyce	s cerevisiae T	CP1-beta gen	e, partial cds; and Ax	l2p
(AXL2) and Rev7p (REV7) genes, complete cds.					
ACCESSION	U49845		_		
VERSION	U49845.1 G	I: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 <1206 /codon start=3
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         source       15028         /organism="Saccharomyces cerevisiae"         /db_xref="taxon:4932"         /chromosome="IX"         /map="9"         CDS       <1206
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         source       15028         /organism="Saccharomyces cerevisiae"         /db_xref="taxon:4932"         /chromosome="IX"         /map="9"         CDS       <1206
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         source       15028         /organism="Saccharomyces cerevisiae"         /db_xref="taxon:4932"         /chromosome="IX"         /map="9"         CDS       <1206
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         source       15028         /organism="Saccharomyces cerevisiae"         /db_xref="taxon:4932"         /chromosome="IX"         /map="9"         CDS       <1206
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         source       15028         /organism="Saccharomyces cerevisiae"         /db_xref="taxon:4932"         /chromosome="IX"         /map="9"         CDS       <1206
DNA damage-induced mutagenesis in Saccharomyces cerevisiae         JOURNAL       Yeast 10 (11), 1503-1509 (1994)         MEDLINE       95176709         FEATURES       Location/Qualifiers         source       15028         /organism="Saccharomyces cerevisiae"         /db_xref="taxon:4932"         /chromosome="IX"         /map="9"         CDS       <1206
JOURNAL Yeast 10 (11), 1503-1509 (1994) MEDLINE 95176709 FEATURES Location/Qualifiers source 15028 /organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
MEDLINE 95176709 FEATURES Location/Qualifiers source 15028 /organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
FEATURES Location/Qualifiers source 15028 /organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
FEATURES Location/Qualifiers source 15028 /organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
source 15028 /organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
source 15028 /organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
/organism="Saccharomyces cerevisiae" /db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
/db_xref="taxon:4932" /chromosome="IX" /map="9" CDS <1206 /codon start=3
/chromosome="IX" /map="9" CDS <1206 /codon start=3
/map="9" CDS <1206 /codon start=3
CDS <1206 /codon start=3
/codon start=3
/product="TCP1-beta"
/protein_id="AAA986651"
$\frac{1}{2} \frac{1}{2} \frac{1}$
/translation="SSIVNCISTSCI DI NNICTIADMDOI CIVESVELED AVVSSASEA
A LEVEL A DEPTA NO ADDITA NO OLAN
AEVLLKVDNIKARPKIANKQHM
gene 0873138
/gene="AXL2"
CDS 6873158
/gene="AXL2"
/note="plasma membrane glycoprotein"
/codon_start=1
/product="Axl2p"
/protein_id="AAA98666.1"
/db_xref="GI:1293615"
/translation-"MTOLOISLLLTATISLLHLVVATOVEAVDIGKOVDDVADVNESE
DASE OUTINE 1510 = 1074 = 0.025 = 16004
BASE COUNT 1510 a 10/4 c 855 g 1009 t
ORIGIN
$\sim$
U49845 Length: 5028 November 30, 2000 10:22 Type: N Check: 3941
$\bigcup_{k}$
1 GATCCTCCAT ATACAACGGT ATCTCCACCT CAGGTTTAGA TCTCA
5001 TTTTAAGCTA TTCAATTTCT CTTTGATC
( 註解和序

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戴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。

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

SeqWeb 所輸出的檔案類型如下表:

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。

NHRI SeqWeb3.1 講義 v1.0

# 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。

StringSearc Search	<del>h</del> for character patterns.		
Input Parame	ters:	輸入關鍵字	搜尋
<u>String to sea</u>	rch for	capping protein beta	
Search Set	protein: uniprot UniProt (SWISS-PROT	plus Translated EMBL)	
	search definition line only search entire annotation section	○ ← 以定義做搜尋 ○ ← 以註解做搜尋	選擇序列 資料庫
<u>Find entries:</u> Include docu	with ANY of the specified patterns with ALL of the specified patterns mentation in output file	ন হ	
Width of doci	umentation in the output file	100 (range 0 thru 220)	

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。

	SeqWeb v <sub>3.1</sub>							
	StringSearch Results							
		Page 1 of 10						
	! STRINGSEARCH from: uni	prot:* August 9, 2006 01:26						
勾選所需的序列,	! searching for: "cappin	19"						
並可加入至 sequence								
TT JULY (T sequence			Add selected to Project					
manager 中	Sequence	Description						
	Uni_sprot:Capg_Human	P40121 homo sapiens (human). macrophage capping p cap-g). 4/2006 348	protein (actin-regulatory protein					
	Uni_sprot:Capq_Mouse	P24452 mus musculus (mouse). macrophage capping p 1) (actin-capping p	protein (myc basic motif homolog					
	Uni_sprot:Capza_Arath	082631 arabidopsis thaliana (mouse-ear cress). f- !subunit (capz-alpha). 4	-actin capping protein alpha					
	Uni_sprot:Capza_Ashqo	Q75ds4 ashbya gossypii (yeast) (eremothecium goss !alpha subunit. 2/200	sypii). f-actin capping protein					
	Uni_sprot:Capza_Caeel	P34685 caenorhabditis elegans. f-actin capping pr	rotein alpha subunit. 4/2006 282aa					
	Uni_sprot:Capza_Canga	Q6fn48 candida glabrata (yeast) (torulopsis glabr alpha subunit. 2/2006	rata). f-actin capping protein					
	Uni_sprot:Capza_Dicdi	P13022 dictyostelium discoideum (slime mold). f-a subunit (cap34). 2/2006	actin capping protein alpha					
	Uni_sprot:Capza_Drome	Q9w2n0 drosophila melanogaster (fruit fly). f-act 4/2006 286aa	in capping protein alpha subunit.					
	Uni_sprot:Capza_Klula	074232 kluyveromyces lactis (yeast). f-actin cap; 262aa	oing protein alpha subunit. 2/2006					
	Uni_sprot:Capza_Neucr	Q9p5k9 neurospora crassa. probable f-actin cappir 269aa	1g protein alpha subunit. 2/2006					
	Uni_sprot:Capza_Schpo	Q10434 schizosaccharomyces pombe (fission yeast). !alpha subunit. 3/	. probable f-actin capping protein					

### 第 35 頁

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。



# 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>

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 第 37 頁

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 (Basic Local Alignment Search Tool)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。

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

Global Alignment

 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:

	А	R	Ν	D	С	Q	Е	G	Η	Ι	L	K	М	F	Р	S	Т	W	Y	V
А	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
Ν	-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 0 0 -3 7 2 -2 1 -3 -2 2 0 -4 -1 0 -1 -1 -1 -3 0 -1 1 Ε -1 0 0 2 -3 2 **6** -3 0 -4 -3 1 -2 -3 -1 -1 -1 -3 -2 -3 -3 0 -1 -3 -2 -3 8 -2 -4 -4 -2 -3 -4 -2 0 -2 -3 -3 -4 G 0 H -2 0 -2 10 -4 -3 0 -1 -1 -2 -1 -2 -3 1 -1 -3 1 0 2 -4 5 2 -3 Ι -1 -4 -3 -4 -2 -3 -4 -4 -4 2 0 -3 -3 -1 -3 -1 4 -2 -3 -4 -4 -2 -2 -3 -4 -3 2 **5** -3 3 -4 -3 -1 -2 -1 L 1 1 3 1 -2 0 -3 -3 6 -2 -4 -1 0 K -1 0 -1 -3 2 -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 -3 -3 -4 -5 -2 -4 -3 -4 -1 F 0 1 -4 0 8 -4 -3 -2 1 4 -1 -1 -3 -2 -1 -4 -1 -1 -2 -2 -3 -4 -1 -3 -4 10 -1 -1 -4 -3 -3 Р S 1 -1 1 0 -1 0 -1 0 -1 -3 -3 0 -2 -3 -1 5 2 -4 -2 -2 Т -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 2 5 -3 -2 0 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 Ε Ε --Р W Η E A Ε sequence 2: ----Α ----**Score=1** -8 -8 -1 -8 5 15 -8 10 6 6 -8

#### Local Alignment

- 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。
- 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)
- 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

Searches for sequences similar to a query sequence. The query and the database searched can be either peptide or nucleic acid in any combination.

- 🖑 <u>Nucleotide query against a nucleotide database (BLASTN).</u>
- 🍇 <u>Peptide query against a peptide database (BLASTP).</u> 🕚
- 🔏 <u>Nucleotide query against a peptide database (BLASTX).</u>

<u>Position Specific Iterated BLAST of a peptide query against a peptide database (PSI-BLAST).</u>

🍇 Peptide query against a nucleotide database (TBLASTN).

🖀 <u>Nucleotide query against a database with translation of both to protein (TBLASTX).</u>

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。

BLAST				
Nucleotic	le query against a nucleotide database (BLASTN).			
Input sequence:	Select From: Default <mark>Project Local File</mark>	Clipb	oard	Database
Sequence	Description	Туре	e Lengt	h Range
<u>k02938.gb_ov</u>	X.laevis 5S RNA gene transcription factor (TFIIIA) mRNA, complete	Ν	1518	<u>1</u> <u>1518</u>
Refresh				Clear

#### C. 選取資料庫。

Input Parameters:	選擇資料庫
Search Set	est_human Human Expressed Sequence Tags (GenBank and EMBL) 🗾
Ignore hits that might occur more than how many times by chance alone	est_human Human Expressed Sequence Tags (GenBank and EMBL) est_mouse Mouse Expressed Sequence Tags (GenBank and EMBL) est_other All Other Expressed Sequence Tags (GenBank and EMBL) genbank GenBank
Number of processors to use for the search	gss Genome Survey Sequences (GSS from GenBank and EMBL) htc HTC htg High Throughput Genomes (HTG from GenBank and EMBL) rs ma Refseg RNA

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:

- 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 sequence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<sup>2597</sup>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)

```
Alignment of k02938.gb_ov to <u>GB_OV:XELTFIIIA</u>
       K02938 X.laevis 55 RNA gene transcription factor (TFIIIA) mRNA,
         complete cds. 4/1993
        Length = 1518
Score = 2159 bits (1089), Expect = 0.0 

Identities = 1152/1152 (100%) ← 相同 base 百分比

Strand = Plus / Plus
           gaattccggaagccgagggctgttcagttgctgaaggagagatggggagagaggggctgc 60
Query: 1
Sbjct: 1
           gaattccggaagccgagggctgttcagttgctgaaggagagatgggagagaaggcgctgc 60
Query: 61
            cggtggtgtataagcggtacatctgctctttcgccgactgcggcgctgcttataacaaga 120
            Sbjct: 61
           cggtggtgtataagcggtacatctgctctttcgccgactgcggcgctgcttataacaaga 120
           actggaaactgcaggcgcatctgtgcaaacaccaggagagaaaccatttccatgtaagg 180
Query: 121
Sbjct: 121
           actggaaactgcagcgcatctgtgcaaacacacaggagagaaaccatttccatgtaagg 180
           Query: 181
Sbjct: 181
```

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<u>直接送到美國 NCBI</u> 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 Searches for sequences similar to a query sequence. The query and the database searched can be either peptide or nucleic acid in any combination.
💑 <u>Nucleotide query against a nucleotide database (BLASTN).</u>
🍇 <u>Peptide query against a peptide database (BLASTP).</u>
🥳 Peptide query against a peptide database (TBLASTN).
💑 <u>Nucleotide query against a nucleotide database (BLASTX).</u>
Nucleotide query against a database with translation of both to protein
(TBLASTX).

n_alu Select Alu Repeats from REPBASE	•
n epd Eukaryotic Promotor Database	٠
n_est Non-redundant Database of GenBank+EMBL+DDBJ EST Division	
n_gss Genome Survey Sequence, includes single_pass genomic data, exon-trapped sequences, and Alu PCR sequences.	
n_htgs High Throughput Genomic Sequences	
n_kabat Kabat Sequences of Nucleic Acid of Immunological Interest	
n_mito Database of mitochondrial sequences, Rel. 1.0, July 1995	
n_month All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days	
n nr Non-redundant GenBank+EMBL+DDBJ+PDB sequences (but no EST's or STS's)	
n_pat Nucleotide sequences from the Patent division of GenBank 😽	
n_pdb PDB nucleotide sequences	
n_sts Non-redundant Database of GenBank+EMBL+DDBJ STS Division	-

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。

$\vartheta$ <sub>NCBI</sub>			GATCCCCGGC TACACACAC CATACGTCTC TAACCAATTCG	Nucleoti		
PubMed Nuc	leotide Protein	Genome	Structure	PopSet	Taxonomy	OMIM
Search Nucleotide 💌	for mblZ18854.11CEC	PBSMR		Go Clear		
	Limits	Preview/Index	History	Clipboard		
About Entrez	Display GenBank	Save Text	Details	Add to Clipboard		
Search for Genes LocusLink provides curated information for	□1: Z18854. C.el	egans mRNA for[g	i:6691] Pu	bMed, Protein, Related	d Sequences, Ta	«onomy, LinkOut
human, fruit fly,	LOCUS	CECPBSMR 1201	bp mRNA	INV	10-mar-1	994
mouse, rat, and	DEFINITION	C.elegans mRNA for	capping p	rotein beta subun	it.	
zebralish	ACCESSION VERSION	Z18854 Z18854.1 GI:6691				
AND THE REAL OF THE REAL	SOLIDGE SOLIDGE	Capping protein be	ta subunit	1		
Entrez Nucleotide	ORGANISM	Caenorhabditis ele	gans,			
Help   FAQ	OKOHNISM	Eukaryota; Metazoa Rhabditoidea: Rhak	.; Nematoda .ditidae: P	; Chromadorea; Rh eloderinae: Caeno	abditida; rhabditis	
Retrieve large data	REFERENCE	1 (bases 1 to 120	11)	croacrinac, cacho	indodi (ib)	
sets	AUTHORS	Waddle, J.A., Coope	r,J.A. and	Waterston, R.H.		
	TITLE	Analysis of the ge	nes encodi	ng actin-capping	protein in C.	elegans:
Check sequence		Complementation of	yeast cap	ping protein null	mutants with	. the
revision history		nematode genes				

# ■ 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.

 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						
		Globally align two p	eptide sequenc	æs.			
選擇兩條對的序列	要比	sequences:	Select From: D	efault <mark>🔹 Project</mark>	Local File	Clipboard	Database
		Sequence	)	Description	Туре	Length	Range
		Sequence ozb human.uniprot sp	<mark>:</mark> rot	Description capzb_human	<b>Type</b> P	Length 276	Range <u>1 276</u>
		<b>Sequence</b> ozb human.uniprot sp ozb yeast.uniprot spr	e rot ot	Description capzb_human capzb_yeast	P P P	<b>Length</b> 276 287	Range <u>1 276</u> <u>1 287</u>

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。

第 48 頁

NHRI SeqWeb3.1 講義 v1.0



# 二、PileUp:多序列並列分析

PileUp 程式,是將<u>一群序列進行比對</u>,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 <mark>- Project Local File</mark>	Clipboa	rd D	atabase
Sequence	Description	Туре	Length	Range
capzb_human.uniprot_sprot	F-actin capping protein beta subunit (CapZ beta).	Ρ	276	<u>1</u> 276
capzb_drome.uniprot_sprot	F-actin capping protein beta subunit.	Ρ	276	<u>1</u> 276
capzb_yeast.uniprot_sprot	F-actin capping protein beta subunit.	Ρ	287	<u>1</u> <u>287</u>
capzb_chick.uniprot_sprot	F-actin capping protein beta subunit isoforms 1 and 2 (CapZ 36/32)	P	277	<u>1</u> 277
capzb_mouse.uniprot_sprot	F-actin capping protein beta subunit (CapZ beta).	Ρ	276	<u>1</u> <u>276</u>
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
capzbchick	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	DPPLE	DGAMPSARER
capzb_mouse	RDKV.VGKDY	LLCDYNRDGD	SYRSPWSNKY	DPPLE	DGAMPSARER
capzb_cnick	KDKV.VGKDY	LLCDYNKDGD	SYRSPWSNKY	DPPLE	DGAMPSAKER
capzb_drome	KDKE.HGKDY	LECDYNKDGD	SYKSPWSNSY	YPPLE	DGQMPSERLR
capzb_yeast	KUSAUSINKEY	LCCDYNKDID	SEKSEWSNIT	TPELSPRULQ	DEPERSAPLE
	101				150
canzh human	VIEVEANNAE	DOVPDI VEEG	CVSSVVI WDI	D HGE	AGVTL TKKAG
capzb_numan	KLEVEANNAE	DOVPDI VEEG	GVSSVVI WDI	D HGE	AGVILIKKAG
capzb_nouse	KLEVEANNAE	DOVRDL VEEG	GVSSVVLWDL	D HGE	AGVILIKKAG
canzh_drome	KLETEANYAE	DOVREMYYEG	GVSSVYLWDL	DHGE	AAVTI TKKAG
canzb veast	KLEILANDSE	DVYRDLYYEG	GISSVYLWDL	NEEDENGHDE	AGVVLEKK.
capro_J cap c		D THE LITES	0100110400		A COLUMN TO A C

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:

Pretty					
Align several peptide	sequences and calculate a consen	sus.			
Input sequences:	Select From: Default - Project	Local File	Clipboa	ard	Database
Sequence	Description		Туре	Lengt	th Range
capzb human.uniprot sprot	F-actin capping protein beta subunit beta).	: (CapZ	Ρ	276	<u>1</u> 276
capzb_drome.uniprot_sprot	F-actin capping protein beta subunit	i.	Ρ	276	<u>1</u> 276
capzb_yeast.uniprot_sprot	F-actin capping protein beta subunit	i.	Ρ	287	<u>1</u> 287
capzb_chick.uniprot_sprot	F-actin capping protein beta subunit and 2 (CapZ 36/32)	t isoforms 1	Ρ	277	<u>1</u> 277
capzb mouse.uniprot sprot	F-actin capping protein beta subunit beta).	t (CapZ	Ρ	276	<u>1</u> <u>276</u>
Refresh					Clear



如果使用者希望較容易分辨出 conserved region 的部分,建議選擇「<u>display</u> <u>alignment only at positions that disagree with the consensus</u>」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。

Add the Consensus Sequence to Your List	
use the button below to add the consensus sequence to your list of input sequences. <b>You</b> nust enter a name for the sequence. You can edit the description line, the reference r the sequence.	:e,
Remove Gaps Add to Project	

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 反圖譜

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。

Мар		3
Display a rest	riction map of your nucleic acid seque	nce.
Input sequence:	Select From: Default 💌 Pr	roject Local File Clipboard Database
Sequence	Description	Type Length Range
<u>gi 214818.ssf</u>	gi_214818 1518 bp linear 01-JAN-1970	N 1518 <u>11518</u>
Refresh		Clear
Input Parameters: Enzyme	All_Enzymes  Enzyme Se	elector 可在此處選擇要 使用的 Enzyme
<u>Display Protein Trans</u> Frames	lation Frames Display the Three Forward Translation Frames	
	Frames	, ,

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。

🚰 Enzyme Selector - Microsoft Internet Explorer		
Enzyme Selector		3
左方的 column 可以 把未用到的 Enzyme 加入至右方 es:	t enzymes from the Available List Selected Enzymes:	右方的 column 顯 示 Map 程式中將 用到的 Enzyme
Aasl7.GA       nn'nnGTC2         Aatl3.AGG'CCT0       Acc16I3.TGC'GCA0         Acc36I10.ACCTGCnnnn'nnnn4       Acc87I4         AccB7I7.CCAn_nnn'nTGG3       Acc87CTC0         AccB8I3.CCG'CTC0       Acc8II1.T'CCGG_A4         AccIII1.3.CAGCTCnnnnn'nnnn1       AcceIII1.T'CCGG_A4         AceIII1.3.CAGCTCnnnnn'nnnn1       AccIVI9.GGATCnnnn'n1         AcvI3.CAC'GTG0       AcvIVI9.GGATCnnnn'n1         AcsI4       AcIVI4         AclVI4       AcIVI4         Actor	▲         Aarl11.CACCTGCnnnn'nnnn4           Aatll5.G_ACGTC4         Acc65l1.G'GTAC_C4           Acc65l1.G'GTAC_C4         Acc65l1.G'GTAC_C4           Accl2.GTmk_AC2         Acil2.A           Acil2.CTGAA'G_TT2         Acol1.Y'CCGG_T4           Acul2.CTGAAGnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn	i'5 2 ▼
Enzyme Files All_Enzymes Sa	ve Save As Delete Show Isos	chizomers
	OK Cancel	
<ul><li>▲ Ênzyme 種類</li></ul>		🤮 網際網路 🛛 🎢



# 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。

Translate+	2
Translate	e nucleic acid sequences into peptide sequences.
Input sequence:	Select From: Default 💌 Project 🛛 Local File 🛛 Clipboard 🔹 Database
Sequence	Description Type Length Range
XELTFIIIA.ssf	X.laevis 5S RNA gene transcription factor (TFIIIA) mRNA, N 1518 <u>11518</u> complete
Refresh	Clear
Codon Frequence	rs: cv Table enteric bacterial (highly expressed) genes 7/19/83 •
	Use all Frames
<u>Reading Frame</u>	1 Vse selected frame
Translation Tab	le Standard [1]
Window length : overlapping wind Overlap to use	for splitting each input sequence to dows     0       when splitting sequences     0
Run 重設	

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。

Translate Results	5
-------------------	---

IIRICH_SEQUENCE 1.0            {             imame XELTFIIIA_p1             descrip Translation of XELTFIIIA in frame 1             type PROTEIN             descrip Translation of XELTFIIIA in frame 1             type PROTEIN             checksum 7861             creation-date 05/19/2006 14:50:18             strand 1             sequence             efrkpravqllkerwerrrcrwcisgtsalsptaallitrtgncrricantgernhfhvr             kkdvrkalprfit*pathslilarktshvtmMdvt*dllgrqt*rstltdsitsrsasMc             ailrtvakhsrntin*rfissvthsschtnvlMkavtsgflclpv*nvMkksMqaipakr             Milahlwerlghyt*ntwgnairt*qyvMcviensgtlt*gin*pltkkselcisaleM             avtapiplhsileaiynhfMrnndllfvsMlaagnalq*kka*kdiqlWmigrgs*rrn             alaqreawplasldtypprakkMMpfreqkrlihl*kisplalkqMahwf*in*lynni             rkhlnlffyllklpsgwlth*cgfflflgl*fifrl*qkesvida*fvl*tavlaMpt             gktvlMatylfypMfaiksevqqplvclftihsfsktlysfskeslrecakllslyckh             kctatllvglflgrltdpvffltef         }         •          C          XELTFIIIA_transl_11275.pep Translation Spanning the entire Length
Select All Add selected to Project
Input Sequence: XELTFIIIA.ssf: !!NA_SEQUENCE 1.0 WPDEF X.laevis 5S RNA gene transcription factor (TFIIIA) mRNA, complete LOCUS XELFFIIIA 1518 bp mRNA linear VRT 27-APR-1993 DEFINITION X.laevis 5S RNA gene transcription factor (TFIIIA) mRNA, complete cds. ACCESSION K02938.1 GI:214818 VERSION K02938.1 GI:214818 KEYWORDS developmental regulation; transcription factor.

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。

 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	

 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。由於 ExPASy 鏡相站設於國衛院,您可以很 快的得到結果。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

Sequence Manag	er							3
							Project	Default 💌
Records: 21	Displaying	: 21- 21	Page: 3 of 3	Pages:	<u>1 2</u> 3	e	Sho	w: 10 💌
📕 🔺 <u>Sequence</u>		Des	<u>cription</u>		Туре	Length	Modif	ied On
XELTFIIIA.ssf	X.laevis 5S mRNA, comp	RNA gene tr lete	anscription factor	(TFIIIA)	N	1518	May 19 2006	14:48:46
Add From: Selec	t 💌	Select a p	roject 💌 Copy	Move			Edit	Delete
Select Clipbo Datab Local	ard ase File 😽							

Search Database I	Results			3
-	elect a Database and ente	er the Entry name or acc	cession number to sea	arch
Project: Default				
Database: nucleic	: genbank DNA Databases (i	GenBank w/o EST, GSS, H	TC) 🗾	
Entry Name OR A	ccession Number: m9616	60		
Note: Use '*' to repre: or AA00368?) Search 重設 Car	sent zero or more characters in	the name. Use '?' to represe	ent a single character in th	ne name. ( e.g.: AA0036*
Records: 1	Displaying: 1- 1	Page: 1 of 1	Pages: <b>1</b>	Show: 10 💌
☐ ▲ <u>Name</u>		Descripti	on	
☑ gb ro:m96160	LOCUS RATADCYB 4131 bp adenylyl cyclase type VI ml	mRNA linear ROD 27-AP RNA, complete	PR-1993 DEFINITION F	Rattus norvegicus
Default - Add				Done

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。

#### 第 58 頁

Мар						3
Display a restriction (	map of your nucleic acid seq	uence.				
Input sequence:	Select From: Default 💌	Project L	ocal File.	Clip	board	Database
Sequence	Description			Туре	e Length	Range
RATADCYB.ssf cds.	vegicus adenylyl cyclase type \	/I mRNA, con	nplete	N	4131	<u>1 413</u>
Refresh						Clear
Enzyme	All_Enzymes 💌 Enzyme	Selector				
	Do Not Display Any Protein Translations	0				
Display Protein Translation	Display Open Translation Frames	c				
Frames	Display the Three Forward Translation Frames	ſ				
	Display All Six Translations Frames	С				
Treat input sequence as eirc	ular					
Show enzymes that cut only	once					
Minimum number of cuts		1 (	(range 1	thru 1	00000)	
Mayimum pumbar of outo		100000	(rongo t	++	000001	

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 Ecll36II 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。

				Edit S		ATABCVR	1 ggf				
				Eult St	equence r	MINDCID	_1.551				
		Feature	e Status Fil	ter: 🔽 Inv	alid 🔽 Su	spect 🔽 P	ending 🔽	Validated			
se-over t	o show featu	re descript	ion. Click t	o select	50000		Active Indexed				
										CDS 14 35	56
	1K	21	(	3K	4K						
										Y I	
Enable Mu	ltinle Seler	tions 💽	Feature C	ORF							
				10721308 2010		anti anti		~			
L1101	L1111	L1121	<u>-</u> 1131	-1141	41151	<b>L</b> 1161	L1171	L1181	L1191	L1201	
		L1121	- - 1131 GCCCAGGAAC	- - 1141 IGGTCATGAO		L1161	L GTTOGACAA		L1191 GAGAATCACTI	L1201 STCTGAGGATCAA	
-1101 TTCACCAC -1211	-1111 SOCTIGEOCTOC -1221	-1121 CAGTGCACTO -1231	LII31 GCCCAGGAAC LI241	L1141 IGGTCATGAO L1251	LI 151 CTTGAATGAGC L1261		L1171 GGTTCGACAAC		L1191 GAGAATCACTI	LI201 STCTGAGGATCAA	
LIDI TTCACCAC LIZII	LIIII GCCIGGCCTCC LI221	L1121 CAGTGCACTO L1231	L1131 GCCCAGGAAC L1241	L1141 IGGTCATGAO L1251	LI151 CTTGAATGAGC L1261	-1161 TCTTIGCCC -1271	LITT GGTTCGACAAC LI281	LII81 CTGGCTGCG LI291	L1191 GAGAATCACTU L1301 CATIGATCGAGG	LIZOI STCTGAGGATCAA LIJII SCCATCTCCSCTUR	~
-1101 TTCACCAC 1211 GATCTTAC	LIIII GCCTGGCCTCC LI221 GGAGACTGTTA LI331	L1121 CAGTGCACTU L1231 CTACTGTGTGTU L1341	-1131 GCCCAGGAAC -1241 GTCGGGGCTGG -1351	-1141 IGGTCATGAC -1251 COGGAGGCCCC -1361	LII51 CTTGAATGAGC LI261 GGGCAGACCAT LI371	<sup>1</sup> 1161 71CTTTBCCCC <sup>1</sup> 1271 BCCCACTBCT <sup>1</sup> 1381	GTTCGACAAC GTTCGACAAC <sup>L</sup> 1281 IGTGTGGAGAT L1391	LI181 CTGGCTGCG L1291 TGGGGGTAGA	LI91 GAGAATCACTU LI301 CATGATOGAGG	L1201 STCTGAGGATCAA L1311 SCCATCTOSCTGG	
-1101 TTCACCAC -1211 GATCTTAC -1321	LIIII GOOTGGOOCTOO LI221 GGAGACTGTTA LI331	L1121 CAGTGCACTU L1231 CTACTIGTIGTIC L1341	L1131 GCCCAGGAAC L1241 GTCGGGGCTGG L1351	L1141 IGGTCATGAC L1251 CCGGAGGCCC L1361	LI151 CTTGAATGAGC L1261 GGGCAGACCAT L1371	1161 TCTTBCCCC 1271 GCCCACTGCT 1381	LITI GGTTCGACAAC LI281 GGTGTGGAGAT LI391	LI181 CTGGCTGCG L1291 IGGGGGTAGA L1401	LII91 GAGAATCACTU LI301 CATGATOGAGU LI411 GGCAGTTUGA1	<sup>4</sup> 1201 STCTGAGGATCAA <sup>4</sup> 1311 SOCATCTOSCTGG <sup>4</sup> 1421 USTCT05CTGG	
1101 11CACCAC 1211 GATCTTAC 1321 16C51GAC	LIIII GCCTGGCCTCC LI221 GGAGACTGTTA LI331 GGTAACGGGTT	L1121 CAGTGCACTU L1231 CTACTGTGTU L1341 GTAAATGTGAA	LII31 GCCCAGGAAC LI241 GTCGGGGCTGG LI351 ACATGCGCGTT	-1141 IGGTCATGAC -1251 CCGGAGGCCC -1361 GGGCATCCAC	LII51 CTTGAATGAGC LI261 GGGCAGACCAT LI371 AGCGGGGTGTGT LI491	41161 TCTTIGCCCC 41271 GCCCACTGCI 41381 ACACTGCGGG 40401	LITI GGTTCGACAAC LI281 GGTGTGGAGA3 LI391 IGTCCTTGGTC LI501	L1181 CTGGCTGCG L1291 IGGGGGTAGA L1401 TIGCGGAAAT	L1191 GAGAATCACTU L1301 CATGATCGAGG L1411 GGCAGTTTGAT	4201 STCTGAGGATCAA 4311 SOCATCTOSCTGG 4421 IGTCTGGTOCAAC	
1101 TTCACCAC 1211 GATCTTAC 1321 IGCDIGAC 1431	LIII GCCTGGCCTCC LI221 GGAGACTGTTA LI331 GGTAACGGGTT LI441	LI21 XAGTGCACTU L231 XCTACTGTGTU L341 STAAATGTGAA L451 XAGATGCAC	-1131 5000A6GAAC -1241 5100666CT60 -1351 ACATGO30370 -1461	-1141 IGGTCATGAO -1251 COGGAGGOOO -1361 GGGCATOCAO -1471	LI151 CTTGAATGAGC L1261 GGGCAGACCAT L1371 AGCGGGC3TGT L1481	<sup>1</sup> 1161 TCTTIGOCO <sup>1</sup> 1271 GOCCACTECT <sup>1</sup> 1381 ACACTECGGT <sup>1</sup> 1491 CTCTECCC	LITTI GGTTCGACAAI LIZBI IGTGTGGAGAT LIZBI IGTCCTTGGTC LISDI	LII81 CTGGCTGCG L1291 IGGGGGTAGA L14D1 CTGCGGAAAT L1511	LI91 GAGAATCACTU L301 CATGATCGAGG L411 GGCAGTTTGAT L521 GACTATGACC	<sup>1</sup> 1201 STCTGAGGATCAA <sup>1</sup> 1311 SOCATCTCG5CTGG <sup>1</sup> 1421 IGTCTG6TCCAAC <sup>1</sup> 1531	
1101 TTCACCAC 1211 GATCTTAC 1321 TGC5TGAC 1431 GATGTGAC	LIII GCCTGGCCTCC LI221 GGAGACTGTT/ LI331 GGTAACOGGTC LI441 CCCTGGCCAAC	<sup>5</sup> 1121 CCAGTIGCACTIC <sup>5</sup> 1231 CTACTIGTIGTIC <sup>5</sup> 1341 STAAATIGTIGAA <sup>5</sup> 1451 CCAGTIGGAGG	1131 50004664A0 1241 5105666000 1351 40416030370 1461 5036666000	1141 IGGTCATGAC 1251 CCGGAGGOCC 1361 GGGCATCCAC 1471 GGGCGGGCCAC 1471 GGGCGGGCCAC	41151 CTTGAATGAG 41261 GGGCAGACCAT 41371 AGCGGGGGGTGT 41481 CATCCACATCA	41161 TCTTIGCCCC 41271 GCCCACTECT 41381 ACACTECTECTE 41491 41491 CTCCEECCAC	41171 GGTTOGACAAC 41281 IGTOGGAGAA 41391 IGTOCTTGGTO 41501 CACTGCAGTAC	-1181 SCTGGCTGO3 -1291 TGGGGGTAGA -1401 TGCGGAAAT -1511 SCTGAAC3GG	41191 GAGAATCACTU 41301 CATGATOGAGU 41411 GGCAGTTTGAT 41521 GACTATGAGG	<sup>1</sup> 201 GTCTGADGATCAA <sup>1</sup> 311 SOCATCTOSCTGG <sup>1</sup> 421 IGTCTGGTOCAAC <sup>1</sup> 531 IGGAGCCAGGCCG	
1101 11CACCAC 1211 GATCITAC 1321 IGOGIGAC 1431 GATGIGAC 1541	LIII SCCTGGCCTCC LI22I GGAGACTGTTA LI33I SGTAACOSGTC LI441 CCCTGGCCAAC LI55I	<sup>5</sup> 1121 CAGTIGCACTIC <sup>5</sup> 1231 CTACTIGTIGTIC <sup>5</sup> 1341 STAAATIGTGAA <sup>5</sup> 1451 CACATIGGAGC <sup>5</sup> 1561	-1131 500CA6GAAC -1241 5T05666CT64 -1351 ACAT6050370 -1461 5036666600 -1571	1141 IGGTCATGAC 1251 COGGAGGCOO 1361 SGGCATOCAC 1471 GGGCGGGCOGG 1581	41151 2111GAATGAGG 41261 3666CAGAQCAT 41371 460366G03101 4481 2470CACATCA 21591	1161 TCTTTGCCC 1271 GCCCACTGCT 1381 ACACTGCGGT 1491 CTCGGGCCAC 1601	<sup>1171</sup> GGTTOGACAAO <sup>1281</sup> IGTGTGGAGAT <sup>1391</sup> IGTCCTTGGTO <sup>1501</sup> CACTGCAGTAO <sup>1611</sup>	1181 5CT66CT603 1291 106666TAGA 1401 7T6056AAAT 1511 2CT6AA0366 1621	L191 GAGAATCACTU L301 CATGATCGAGG L1411 GGCAGTTTGAT L521 GACTATGAGG L631	<sup>1201</sup> STCTGAGGATCAA <sup>1311</sup> SCCATCTOSCTGG <sup>1421</sup> IGTCTGGTOCAAC <sup>1531</sup> IGGAGOCAGGCOS <sup>1641</sup>	
41101 TTCACCAO 41211 GATCTTAO 1321 TGCGTGAO 41431 GATGTGAO 41541 TGCGGGTO	LIII SCCTGGCCTCX LI221 GGAGACTGTT# LI331 GGTAACOGGTU LI441 SCCTGGCCAAC LI551 GAGCGCAACGC	LI121 CAGTIGCACTU L1231 CTACTIGTGTU L1341 STAAATIGTIGAA L1451 XCACATIGGAGU L1561 DITACCTICAAU	-1131 500046GAAC -1241 51005666C164 -1351 ACATGC030371 -1461 50366666000 -1571 50366666000	1141 IGGTCATGAC 1251 COGGAGGCCC 1361 GGGCATCCAC 1471 GGGCGGGCCGG 1581 ATTGAGACCT	LII51 CITIGAATGAGG LI261 GGGCAGACCAT LI371 AGCGGGCGGTGT LI481 CATCCACATCA LI591 ICCTCATACTA	1161 1271 5271 500040560 1381 4491 51605 51601 5646000460	51171 GGTTOGACAAG 5281 IGTGTOGAGAT 5391 IGTCCTTIGGTG 501 CACTGCAGTAG 5611 CAGAAACTGGA	-1181 GCTGGCTGCG -1291 IGGGGGTAGA -1401 TGCGGAAAAT -1511 CCTGAACGGG -1621 IAGAGGAGAA	LI91 GAGAATCACTU L301 CATGATOGAGU L411 GGCAGTTIGAT L521 GACTATGAGGT L631 GGCCATGCTGU	<sup>1</sup> 1201 STCTGAGGATCAA <sup>1</sup> 311 GCCATCTOGCTGG <sup>1</sup> 421 IGTCTGGTCCAAC <sup>1</sup> 531 IGGAGCCAGGCCG <sup>1</sup> 1641 GTCAAGCTGCAGC	
1101 TTCACCAU 1211 GATCTTAU 1321 TGCGTGAU 1431 GATGTGAU 1541 TGGCGGTU 1651	-1111 GOCTGGOCTOC -1221 GGAGACTGTTA -1331 GGTAAOGGGTU -1441 COCTGGOCAAC -1551 GAGGCCAACGC -1661	LI121 CAGTIGCACTIC L231 CTACTIGTIGTIC L341 STAAATIGTGAA L451 CACATIGGAGG L561 DTACCTICAAC L671		1141 IGGTCATGACI 1251 COGGAGGCCCI 1361 GGGCATCCACI 1471 GGGGGCCGG 1581 ATTGAGACCT 1691	LISI CTTGAATGAGC L261 GGGCAGACCAT L371 AGCGGGCGTGT L481 CATCCACATCA L591 TCCTCATACTA L701	51161 TCTTT6CCC 51271 GCCCACT6CT 51381 SACACT6CG6 54491 SCCT6C66CCAG 51601 GGAGCCAGCC 51711	LITI GTTCGACAAU L281 IGTCTGGAGA3 L391 IGTCCTTGGTC L501 CACTGCAGTAC L611 CAGAAAO3GAJ L721	LISI CTGGCIGCG L291 CGGGGTAGA L401 TGCGGAAAT L511 CTGAAC566 L621 VAGAGGAGAA L731	LI91 GAGAATCACTC L1301 CATGATOGAGC L1411 GGCAGTTIGAT L1521 GACTATGAGGT L631 GGCCATGCTGC L1741	<sup>1</sup> 12D1 STCTGAGGATCAA <sup>1</sup> 311 SOCATCTOSCTGG <sup>1</sup> 421 IGTCTGGTOCAAC <sup>1</sup> 531 IGGAGOCAGGCOG <sup>1</sup> 1641 STCAAGCTGCAGC <sup>1</sup> 751	
-1101 TTCACCAC 1211 GATCTTAC 1321 TGC3TGAC 1431 GCATGTGAC 1541 TGCC3GTC 1651 GGACGCGC	41111 SOCIGOCOCO 4221 GGAGACIGITA 41331 GGTAAOGGGIU 41441 SOCIGOCAAG 41551 GAGOGCAAGGO 41661 GGOCAACICO	LI121 CAGTIGCACTII L231 CTACTIGTIGTIC L1341 STAAATGTGA/ L451 CACATIGGAGG L561 DITACCTICAA0 L671 CTGGAAGGACTI	-1131 5000A6GAAC -1241 5103666CT04 -1351 ACATGO30370 -1461 5036666000 -1571 5036666000 -1571 5036746700 -1681 104750000064	-1141 IGGTCATGAO -1251 2036AGGCOO -1361 2066CATOCAC -1471 3060GGCOO -1581 ATTGAGACCT -1691 2169GTTCCTI	41151 2711GAATGAGC 41261 3660CAGACCAT 41371 460066003161 41481 2410CACACTAC 1591 1001CACACATACTA 1701 540035160011	1161 TCT TECCC 4271 GCCCACTECT 4381 ACACTECE 4491 CTCGEGCCAC 4601 GGAGCCAGC 4711 CTCCCCGCAGC	41171 GGTTOGACAAG 4281 GTGTGGAGA3 4391 IGTCCTTGGTG 4501 CACTGCAGTAG 4611 CAGGAACGGAA 4722 CAGGAACGGACTCT/	LISI CTGGCTGCG L291 CGGGGTAGA L401 CTGCGGAAAT L511 CTGAACGGG L621 LAGAGGAGAA L731 LAGGCATTCC	1191 GAGAATCACTU 1301 CATGATOGAGO 1411 GGCAGTTTGAT 1521 GACTATGAGGT 1631 GGCCATGCTGC 1741 GACAGATGGGC	<sup>L</sup> 12D1 STCTGAGGATCAA <sup>L</sup> 1311 SOCATCTOSCTGG <sup>L</sup> 1421 IGTCTGGTCCAAC <sup>L</sup> 1531 IGGAGOCAGGCOG <sup>L</sup> 1641 STCAAGCTGCAGC <sup>L</sup> 1751 CATQGATGACTCT	
41101 TTCACCAC 4211 GATCTTAC 41321 TGC3TGAC 41321 GATGTGAC 1541 TGCCGGTC 4651 GGACGCGC 41761	4111 SOCTIGGOCTOC 4221 GGAGACTIGTI, 41331 GGTAAOSGGTO 41441 COCTIGGOCAAC 41551 SAGGGCAACGOC 41661 GGOCAACTOCA 4771	41121 CAGTGCACTU 41231 CTACTUFTGTU 41341 FTAAATGTGAA 4451 CCACATTGCAGU 41561 GTACCTCCAAU 41671 LTGGAAGGACT 41781	-1131 GCCCA6GAAC -1241 STCGGGGCTG4 -1351 ACATGCGGGTG4 -1461 GCGGGGGGCCC -1681 GGAGCAGTGC4 -1681 GGATGCCCCCG4 -1791	1141 IGGTCATGAC 1251 CCGGAGGCCC 1361 GGGCATCCAC 1471 GGGCGGGCCCG 1581 ATTGAGACCT 1691 CTGGGTTCCTI 1801	41151 20164416460 201641 20174 20174 20174481 2010404104 201041041 201041401 201041401 201041401 201041401 201041401 201041401 201041401 201041401 201041401 201041401 201041401 20104140 20104110 201041400 201041400 201041400 201041400 201041400 201041400 201041400 201041400 201041400 2010410000000000	-1161 тстттессо: -1271 сосодствоет -1381 -1381 -1491 стобебссаю -1601 ведаессаюс -1711 стособасо -1711 стособасо -1821	41171 GGTTOGACAAG 1281 GTGTGGAGAG 1391 IGTCCTTGGTC 1501 CACTGCAGTAC 1611 CAGAAACGGAJ 1721 CAGGACTCTA 1831	1181 GCTGGCTGCG 1291 ICGGGGTAGA 1401 ICGGGAAAT 1511 ICTGAACGGG 1621 AAGGAGGAGAA 1731 AAGGATTCC 1841	41191 GAGAATCACTC 41301 CATGATGSAG 41411 GGCAGTTIGAT 4521 GACTATGAGG 4631 GGCCATGCTG 4741 GACAGATGGC 4851	L1201 STCTGAGGATCAA L1311 SCCATCTOSCTGG L421 IGTCTGGTCCAAC L531 IGGAGOCAGGCCG L1641 STCAAGCTGCAGC L751 CATOSATGACTCT L1861	
41101 TTCACCAC 1211 GATCTTAC 1421 TGC3TGAC 1431 GATGTGAC 1541 TGC3GTC 1651 GGACGCGC 4761 AGCAAAG6	-1111 GCCTGGCCTCC -1221 GGAGACTGTTA -1331 GGTAACGGGTC -1441 CCCTGGCCCAAC -1551 GAGCGCAACGCC -1661 GGCCAACTCC2 -1771 AGAACGGGGG	L121 CAGTECACTU L231 CTACTETETU L341 TAAATETGAA L451 CCACTEGAAG L561 DFTACCTCAAU L671 LT61 L781 GCCCAAGATU	-1131 GCCCAGGAAC -1241 GTCGGGGCTG4 -1351 ACATGGGGGTG4 -1461 GCGGGGGGCC3 -1571 GGAGCAGGGGCC3 -1571 GGAGCAGGCAGGGCC3 -1781 JT91 GCTCTGAACCC	1141 IGGTCATIGACI 1251 2006A660000 1361 3660CATOCACI 1471 3660C660000 1471 3660C660000 1471 3660C60000 1581 ATTGA660C0T 1691 2006GFTCCTT 1801 2006ATGATGATGATGAT	Ч151 СТГСААТСААС Ч261 ССССАТСАТС Ч371 АСССОСССТС Ч481 САТСССАТСАТСА Ч701 ССССАТАСТА Ч701 БАСССТСССТТ Ч811 ЭСТССАСАССА	-1161 тстттососо -1271 Босодотвост -1381 -1491 стозбеоссас -1601 вбабосодосо -1711 стособеосо -1821 ттотобеосо	41171 GGTTCGACAAG 4281 GGTGTGGAGAT 4391 GGTCCTTGGTG 4501 CACTGCAGTAG 4611 CAGAAACDGA/ 4721 CAAGGACTCT/ 4831 GAGCCATCGAT	41181 GCTGGCTGGG 4291 GGGGGTAGA 4401 TGCGGAAAT 4511 XTGAACGGG 4621 AGGGGAGAA 4731 AGGCATTCC 4841 GCCCGAAGC	LI91 GAGAATCACTU L301 CATGATOJAGO L411 GGCAGTTTGAT L521 GACTATGAGGT L631 GGCCATGCTGO L741 GACAGATGGGG L851 ATOGACCAGC	<sup>1</sup> 1201 STCTGAGGATCAA <sup>1</sup> 1311 GCCATCTOGCTGG <sup>1</sup> 1421 IGTCTGGTOCAAC <sup>1</sup> 531 IGGAGCCAGGCCG <sup>1</sup> 1641 STCAAGCTGCAGC <sup>1</sup> 1751 CATCGATGACTCT <sup>1</sup> 861 IGGGTAAGGACCA	
41101 TTCACCAC 1211 GATCTTAC 1321 IGCOTGAC 4131 GATGTGAC 1541 TGGCGGTC 1551 GGACGCGTC 1761 AGCAAAG6/ 41871	-1111 GOCTGGOCTOX -1221 GGAGACTGTTA -1331 GGTAAOSGGTU -1441 COCTGGOCAAC -1551 GGCCAACTOX -1661 GGCCAACTOX -1771 AGAACOSGGG -1881	LI121 CAGTIGCACTIL L231 CTACTGTGTGT L341 TTAAATGTGAA L451 CACATGGAGG L561 CTACCTCAAG L671 L1671 L1671 L1671 L1781 GCCCAAGATIL L991	-1131 GCCCAGGAAC -1241 GTCGGGGCTG4 -1351 ACATGCGCGFT -1461 GCGGGGGGCC -1571 GGAGCAGGGGGCC -1581 IGATGCCCCGG -1791 GCTCTGGAACC -1901	1141 IGGTCATIGACI 1251 2036A660000 1361 2036A660000 1361 2036A0000 1471 2036A0000 1471 2036A0000 1471 2036A0000 1471 2036A0000 1471 2036A0000 2037 20	41151 CTTGAATGAGC 4261 666CAGACCAT 4371 AGCG60GGTGT 4481 CATOCACATCA 41591 FOCTCATACTA 41591 FOCTCATACTA 41701 SACCGTGCACGAGT 41921	-1161 лст посос -1271 боссаствост -1381 -1491 стозебосас -1601 вбабосабос -1711 стосозбасс -1821 -1225 -1821 -1231	41171 GGTTOGACAAU 4281 IGTOGFAGAGA 4391 IGTOCTTGGTO 4501 CACTGCAGTAC 4611 CAGAAACGGACTCT 4721 CAAGGACTCT 4831 GAGOCATCGAT 4941	<u>-1181</u> 3CTG6CTGC3 -1291 1GG6GGTAGA -1401 1TGG6GGAAGA -1511 2CTGAAC3GG -1621 4AG6GAGAGA -1731 4AG6CATTCC -1841 1GCCCGAAGC -1951	L191 GAGAATCACTC L1301 CATGATOGAGC L1411 GGCAGTTTGAG L1521 GACTATGAGGT L1631 GGCCATGCTGC L1741 GACAGATGGCC L1851 ATOGACCAGCC L1961	<sup>1</sup> 1201 STCTGAGGATCAA <sup>1</sup> 311 SOCATCTOSCTGG <sup>1</sup> 1421 IGTCTGGTOCAAC <sup>1</sup> 531 IGGAGCCAGGCOS <sup>1</sup> 1641 STCAAGCTGCAGC <sup>1</sup> 751 CATOBATGACTCT <sup>1</sup> 861 IGOSTAAGGACCA <sup>1</sup> 971	
41101 Treacear 4211 GATCTTAC 4321 TGOSTGAC 431 GATGTGAC 1541 TGOGGGTC 4761 AGCAAG6 1761 AGCAAG6 1871 JUDICTOC	4111 5000000000 4121 5000000000000000 5000000000000000000	LI121 CAGTIGCACTII L231 CTACTIGTIG L341 STAAATIGTGA4 L451 SCACATIGGAGG L561 DFTACCTICAAC L1671 LTGGAAGGACT L781 GCOCAAGATII L891 CACCTICCAI	-1131 5000A6GAAC -1241 5103666CT64 -1351 ACATGO30370 -1461 50366666000 -1571 566666000 -1571 5670 -1681 1681 1681 1681 1681 1681 1681 168	-1141 IBGTCATGAC -1251 COGGAGGCCCC -1361 GGGCATCCAC -1471 GGGCGGGCCCAC -1581 ATTGAGACCT -1691 CTGGGTTCCTC -1801 CTGGGGTTCCTC -1911 -TCTAGGAGAGA	1151 2717GAATGAGC 1261 366CAGACCAT 1371 460C660C71GT 41481 247CCACATCA 1591 102TCATACTA 1701 54CCG71GCCTT 1811 36TG6ACGAG4 1921 46TATTCACG9	1161 TCTTTGCCC 4271 GCCCACTGCT 4381 ACACTGCGGG 4491 CTCGGGCCAC 4601 GGAGCCAGCC 4711 CTCCCGGACC 4281 TTCTGGGCCC 4931 AAAGTAGACC	41171 GGTTOGACAAU 4281 IGTGTOGAGA3 4391 IGTCCTIGGTU 4501 CACTGCAGTAU 41611 CAGAAAOGGAU 4721 CAAGGACTCTA 4831 SAGOCATCGAT 4941 2003111030	1181 GCTGGCTGCG 1291 TGGGGGAGA 1401 TGCGGAAAT 1511 CTGAAC5GG 1521 AGGGGAGAA 1731 AGGGCATTCC 1841 TGCCGGAAGC 1951	LI91 GAGAATCACTU L301 CATGATGAGG L411 GGCAGTTIGAT L521 GACTATGAGGT L631 GGCCATGCTGU L741 GACAGATGGGC L1741 GACAGATGGGC L1851 ATGGACCAGCT L961 CTGCTGTGTGCC	<sup>1</sup> 12D1 STCTGAGGATCAA <sup>1</sup> 311 SOCATCTOSCTGG <sup>1</sup> 1421 IGTCTGGTOCAAC <sup>1</sup> 531 IGGAGOCAGGCOG <sup>1</sup> 1641 STCAAGCTGCAGC <sup>1</sup> 751 CATOGATGACTCT <sup>1</sup> 1861 IGGTTAAGGACCA <sup>1</sup> 971 CICCTGGTTTTCT	
41101 TTCACCAC 41211 GATCITAC 41321 GATGIGAC 41541 TGGOGGGT 41551 GGACGCGC 41761 AGCAAAG/ 41871 TGTGCGCC	41111 SOCTGGOCTOC 4221 GGAGACTIGTTA 4331 GGTAAO3GGTT 4441 COCTGGOCAAO 4551 GAGOGCAAO3C 4661 GGCCAACTOC2 41771 AGAACO3GGG 41881 COCTGCTCCTGCC 4000	41121 CAGTIGCACTII 41231 CTACTIGTIGTIC 41341 51AAATIGTGA/ 41451 51AOCTICAAI 41671 51AOCTICAAI 41671 51781 50OCAAGACIC 41891 5CAOCTICCAI 52001	-1131 GCCCAGGAAC -1241 GTCGGGGCTGG -1351 -1461 GCGGGGGGCCC -1571 GGAGGAGGATGC -1791 GCTCTGAACC -1791 GCTCTGAACC -1901 GAGGGAGGATC -2011	-1141 IGGTCATGAO -1251 CO36AGGCCO -1361 GGGCATCCAC -1471 -1691 CTGGGTCCTT -1691 CTGGGTTCCTT -1801 CTGAGAGAGAA -1911 CTGAGAGAGAA	-1151 CTTGAATGAGG -1261 GGGCAGACCAT -1371 AGCGGGGGGTGT -1481 CATCCACATCA -1591 TCCTCACATCA -1591 TCCTCACATCA -1701 GACGTGCCTT -1811 GGTGGACGAGT -1921 AGTATTCACGG -2031	-1161 тстттессо: -1271 GCCCACTEGE -1381 -1491 -1491 -1601 -1601 -1601 -1711 (СТССССБССАС -1711 (СТССССБССАС -1711 (СТССССБСССС -1821 ТТСТБЕБЕССС -1931 	41171 GGTTOGACAAG 1281 IGTGTOGAGAG 1391 IGTCTTIGGT( 1501 CACTGCAGTAG 1611 CACGAACGGA/ 1721 CAAGGACTCTA 1831 GAGCACTCTA 1941 CTOSTTTICGA CDS1	1181 GCTGGCTGGG 1291 ICGGGGTAGA 1401 ICGGGAAAT 1511 CTGAAGGG 1621 AGGAGGAGAA 1731 AGGCATTCC 1841 IGCCGAAGC 1951 GAGCCTACGT 2061	1191 GAGAATCACTU 1301 CATGATG3AG 1411 GGCAGTTIGAT 1521 GACTATGAGG 1631 GGCATGCTG 1741 GACAGATGGC 1851 ATGGACCAGCT 1961 CGCTGTGCCC 2021	<sup>1</sup> 1201 STCTGAGGATCAA <sup>1</sup> 1311 SOCATICTOSCTGG <sup>1</sup> 1421 IGTCTGGTOCAAC <sup>1</sup> 1531 IGGAGOCAGGCOS <sup>1</sup> 1641 STCAAGCTGCAGCC <sup>1</sup> 751 CATOSATGACTCT <sup>1</sup> 1861 IGOSTAAGGACCA <sup>1</sup> 1971 CTCCTGGTTTTC <sup>1</sup> 2081	

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。

🛚 Seq Web Sequence Editor	×
File Edit Functions Feature View Help	
Edit Sequence CVGEM11ZP.ssf	
Feature Status Filter: 🔽 Invalid 🔽 Suspect 🔽 Pending 🔽 Validated	
Mouse-over to show feature description. Click to select	
500       1000       1500       2000       2500       3000       3500         misc_feature       864       880       misc_feature       864       880         misc_feature       822       3777       misc_feature       322       3778         misc_feature       3898       3914       3898       3914	-
promoter 3941 3957 start 25 25 end 778 778	•
▶ Enable Multiple Selections   Feature   UKF	
	1
491 4101 4111 4121 4131 4141 4151 4161 4171	
CTOSCTEGTEQETEAEGETAAQEEGTETAAATETEAACATEQEOETEEGCATCCACAEGEGQETETACACTEGEETETCCTTEGTCTEGE	
-181 -191 -201 -211 -221 -231 -241 -251 -261	
GAAATGGCAGTTTGATGTCTGGTCCAACGATGTGACCCTGGCCAACCACATGGAGGGGGGGG	
<sup>2</sup> 271 <sup>2</sup> 281 <sup>2</sup> 291 <sup>3</sup> 301 <sup>3</sup> 311 <sup>3</sup> 321 <sup>3</sup> 331 <sup>3</sup> 341 <sup>3</sup> 351	
ĢGCCACACTGCAGTACCTGAACG6GGACTATGAGGTGGAGCCAGGCCGTGGCGGTGAGCGCAACGCGTACCTCAAGGAGCCAGTGCATTC	
<sup>1</sup> 361 <sup>1</sup> 371 <sup>1</sup> 381 <sup>1</sup> 391 <sup>1</sup> 401 <sup>1</sup> 411 <sup>1</sup> 421 <sup>1</sup> 431 <sup>1</sup> 441	
AGACCTTCCTCATACTAGGAGCCAGCAGAAACGGAAAAGAGGAGAAGGCCATGCTGGTCAAGCTGCAGCGGACGCGAGCCCAACTCCATGC	
451 461 471 481 491 501 511 521 531	
<u>ААББАСТБАТĢССССБСТББĢТТССТБАССЭТБССТТСТСССББАССААБĢАСТСТААББСАТТССБАСАĢАТБББСАТСЭАТБАСТСТА</u>	1
警告: Applet 視窗	8

步驟五、使用 mapplot 作一次驗證,看 Ecll36II 和 XhoI 是否還存在。



NHRI SeqWeb3.1 講義 v1.0

# 附件

FAQ	何時使用 GenWeb、SeqWeb 或 GCG Command Mode?		
GenWeb			
	<建議使用>	靈敏度較高,費時較久的序列搜尋比對程式,如 BLAST、	
		FASTA、S-W、FrameSearch、ProfileSearch 等。	
	<不建議使用>	BLASTN、TFASTX。	
SeqWeb			
	<建議使用>	含圖形輸出的 GCG 程式,和 GCG 常用程式,如 Map、	
		PileUp、Frames 等。	
	<不建議使用>	BLAST 、 FASTA 、 StringSearch	
GCG Command Mode			
	<建議使用>	需一次進行多條序列分析或 SeqWeb 沒有的分析程式以	
		及 BLASTN、NetBLAST 等,或是要發表用的.ps 圖形。	
	<不建議使用>	對初學者而言,含圖形輸出的 GCG 程式。	

Project 將 GCG 的序列檔案放入 SeqWeb 中

- Step 1 將 DNA 序列檔以 FTP 下載至個人電腦中,存成純文字檔
- Step 2 在 SeqWeb 的 Sequence Manager 中的 Add 功能,選 Add from local file 即可

### Project 將 SeqWeb 的序列檔案放入 GCG 中

- Step 1 在 SeqWeb 的 Sequence Manager 中將檔案叫出,直接以 Save As 存到個人電 腦中。
- Step 2 以 FTP 上傳至 GCG, 再 reformat 即可。

NHRI SeqWeb3.1 講義 v1.0

# 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.
- 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. 國家衛生研究院生物資訊首頁:http://bioinfo.nhri.org.tw
- 2. GCG Manual : http://bioinfo.nhri.org.tw/gcghelp/gcgmanual.html
- 3. POST 網頁: http://binfo.ym.edu.tw/post/
- 4. NCBI : http://www.ncbi.nlm.nih.gov
- 5. EMBL : http://www.embl-heidelberg.de/
- 6. DDBJ: http://www.ddbj.nig.ac.jp/
- 7. ExPASy : http://tw.expasy.org/
- 8. Uniprot : http://www.expasy.uniprot.org/

L SeqWeb v3.2 講義編輯

李桂玉	主任
汪詩海	先生
王旭川	女士