

This form allows you to assemble a set of contiguous sequences (contigs) with the CAP3 program.
If you use CAP3 in any published work, please cite the following reference:

Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. Genome Res., 9, 868-875.

Documentation is available [here](#).

生物資訊分析服務平台(一)

-序列分析工具CAP3與SSR

生
物
資
訊

農試所作物組 呂椿棠 邱金水 黃彥華

農試所副所長 呂秀英

一、前言

分子生物學是近十幾年來發展最迅速的科學之一，大量的生物序列資料必須透過快速、精確且有系統的收集、整理、檢索與分析，才能於最短時間內從中找出表達基因、蛋白質等結構與功能的規律，進而發現生物學的新觀點，此必須整合生物、統計數學和資訊的跨領域科學，因此一個新的跨領域研究學門－生物資訊學(bioinformatics)於是衍生盛行。在生物資訊學的研究及應用過程中，依照資訊的性質與需求，會運用到各式各樣的分析軟體。這些軟體有些是免費，有些則需付費購買；有些軟體可以直接從網路上下載程式使用，但有些軟體只能透過網站的瀏覽介面線上操作；此外，不同軟體各有其適用範圍、安裝環境與特定的資料格式，因此一旦研究內容必須運用到不同來源的分析軟體時，往往必須耗費很多時間和人力於各分散網站上進行操作與資料輸出入格式之重整。本所作物組的生物統計與生物資訊研究室近十年來在生物資訊研究

上已有顯著成果，基於研究需要，已建構完成生物資訊分析服務平台，陸續將一些常見且公開免費使用的生物資訊分析軟體進行整合，以簡化資料處理流程、縮短分析時間及提高結果準確度，並設計一個具有親和性的操作介面，讓一般使用者皆能透過網頁瀏覽器直接且容易地使用平台內各種分析工具。本文將先針對平台內所提供之序列分析工具的序列接合及簡單重複序列分析兩個功能的內容做一摘要介紹，並以實例說明操作步驟，以供資源分享。

二、分析服務平台內容

平台是在Linux作業環境下架設，所提供之服務內容分為三大類別：

(一) 序列分析工具(Sequence Analysis Tools)，(二) 美國國家生物技術資訊中心(National Center for Biotechnology)的序列相似度搜尋程式(NCBI BLAST)，

(三) 其他生物資訊工具(Other Sequences Tools)，平台首頁如圖一所示。平台內的分析工具或軟體皆源自於網路上可公開下載的免費程式，而本研究室是利用perl、CGI與PHP等工具語言將這些工具軟體進行整合或直接連結，以達成自動化分析與分析結果加值化再利用的目標，並以網頁的顯示介面呈現給使用者。

通訊作者：呂椿棠副研究員
連絡電話：04-23317127

CAP3 Sequence Assembly Program

This form allows you to assemble a set of contiguous sequences (contigs) with the [CAP3](#) program. If you use CAP3 in any published work, please cite the following reference:

Li, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, 9, 868-877. Documentaion is available [here](#).

三、序列分析工具之功能介紹

進入本平台，在序列分析工具 (Sequence Analysis Tools) 選項功能下 (圖一)，目前提供兩種分析程式：序列接合與簡單重複序列分析，分別點選首頁左側視窗畫面之「Sequence Assembly Program (CAP3)」及「Sequence Single Repeat (SSR)」即可進入分析畫面，其功能說明如下：

(一) 序列接合程式 (Sequence Assembly Program (CAP3))

本平台所提供序列接合的功能，主要執行CAP3序列接合程式。CAP3尤其適用於較短序列的拼接組合，例如表現序列標幟 (expressed sequence tag, EST)。EST是由訊息RNA(messenger RNA, mRNA)經反轉錄酶(reverse transcriptase)

產生互補DNA (complementary DNA, cDNA)的5'端或3'端進行單一自動化定序所得到的短序列(長度通常200-500 bp)，為特定生長發育期或特定生理狀態下之組織的基因表現，以提供基因辨識、新基因發現及基因表現差異性等研究。由於EST序列是隨機取樣之各轉錄產物mRNA，表現量多的EST定序可能具有較高的重複性而導致產生多餘序列，故可能會有多个EST代表同一個轉錄產物，因此進行EST分析之前，為維持序列的品質，通常都會運用聚集與接合的方法，減少EST的重疊性，並將其連接成更長、更高質量的序列。利用本平台安裝的CAP3程式，可去除上傳序列的冗餘性，使聚集及接合後的接合序列(contig或consensus sequence)及單獨成群序列 (singleton or singlet)所組成的單一序列



圖一、本所的生物資訊分析服務平台(<http://128.1.3.206/bioinfo/saos.html>)目前提供三大類別的服務內容：序列分析工具 (Sequence Analysis Tools)、NCBI的序列相似度搜尋程式 (NCBI BLAST) 及其他生物資訊工具 (Other Sequences Tools)。

For a more advanced usage of CAP3, it is recommended to install the original software on your local computers. CAP3 is freely available from Xiaoqiu Huang's [web site](#).

Enter your sequence in FASTA format:

(unique sequences)的長度增加。經由本平台執行CAP3序列接合，可獲得這些接合序列、單獨成群序列及其各種基本訊息，CAP3程式內設各以6個檔案分別存放，這些分析結果不但可供下載到使用者自己個人電腦進行其他後續分析，也可用來繼續執行下述之簡單重複序列分析。

(二) 簡單重複序列分析程式 (Sequence Single Repeat (SSR))

簡單重複序列簡稱SSR，又稱微衛星DNA (microsatellite DNA)，是基因組中以2-6個鹼基為基本重複單位組成的長達幾十個鹼基的重複序列。植物基因組中存在豐富的SSR標記，且均勻分布於整個植物基因組中，其兩端序列一般為相對保守的單複製序列，故據此可以設計特異引子(primer)進行SSR-PCR，以擴增串聯重複序列，根據串聯重複數的不同，可以探討SSR的多態性，是一種非常有用、使用成本甚低的標記。傳統上SSR標記是建立在PCR基礎上的DNA分子標記，近來也發現EST也可提供發掘SSR標記的一個很有用的來源，目前EST-SSR已用於建構遺傳圖譜、分離與鑒定新基因、基因表現差異、比較基因組及製備DNA晶片等各方

面研究。藉由本平台執行該程式，能判定出所上傳序列中是否含有SSR。而該上傳序列的來源可來自於上述執行序列接合程式(CAP3)的分析結果之直接連結使用，也允許使用者自行上傳序列檔案。

四、序列分析工具之實例操作

現以一個水稻抗稻熱病之EST序列資料為例子，說明從序列接合到完成所有單一序列的簡單重複序列分析的完整過程。

(一) 序列接合

連線至本平台，點選左側視窗「Sequence Assembly Program (CAP3)」項目（圖一），即可進入序列接合操作畫面（圖二）。

The screenshot shows the CAP3 Sequence Assembly Program web interface. It includes a title bar, a description of the program, a reference citation, and instructions for usage. The main form area has a text input field for FASTA format sequences, a file selection button, and a 'Show CAP3 Default values' button. Below the form are fields for 'E-mail' and 'password', and a '執行' (Execute) button. At the bottom, there are links for '檔案合併系統' (File merging system) and '取得分析後的檔案' (Get files after analysis). Four red boxes with arrows point to specific elements: 1. The FASTA format text input field. 2. The file selection button. 3. The 'E-mail' and 'password' fields. 4. The '執行' (Execute) button.

1.可直接貼上FASTA格式的序列

2.或指定FASTA格式序列的檔案路徑

3.設定E-mail帳號與密碼後，按「執行」

4.執行後，按「取得分析後的檔案」

圖二、CAP3序列接合的操作畫面。

CAP3 Sequence Assembly Program

This form allows you to assemble a set of contiguous sequences (contigs) with the [CAP3](#) program.
If you use CAP3 in any published work, please cite the following reference:

X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, 9, 868-877.
Documentation is available [here](#).

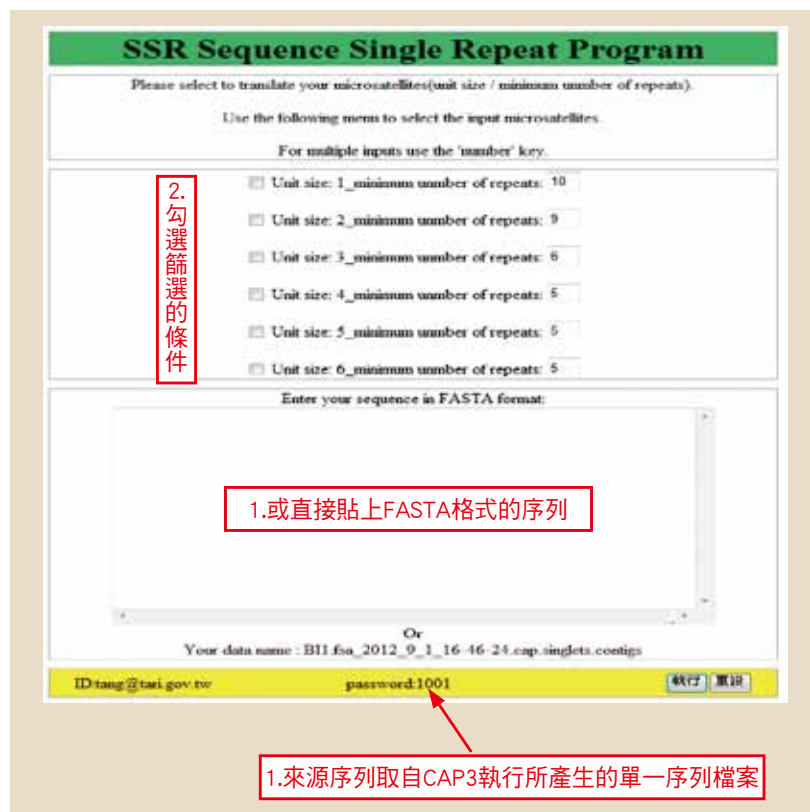
生
物
資
訊

在序列接合操作畫面的空白視窗貼上序列資料或利用下方「瀏覽」視窗指定序列檔案的路徑，這些序列資料須以FAST格式建立，並於E-mail及password欄位內輸入E-mail帳號與密碼，作為個人分析資料查詢與辨識之用，最後點選「執行」鍵，即進行序列接合工作，如圖二所示。

程式完成分析後，可進而在操作畫面下方點選「取得分析後的檔案」（圖二），當上述所輸入的E-mail帳號與密碼被系統辨識成功，本平台會依個人帳號表列出執行CAP3所有分析結果的執行日期（upload time）及檔案(file name)之清單，如圖三-A所示，本例點選畫面中的BI1.fsa檔案，即可顯示出如圖三-B所示左框視窗內的6個分析結果檔案清單，此時可直接點選各分析結果檔案以瀏覽與下載，或繼續執行後



圖三、序列接合的結果。(A)先依個人Email帳號表列出所有分析結果清單，可點選左框視窗內各清單檔案；(B)點選清單檔案後顯示CAP3執行產生的6個檔案，可點選右框視窗內的CAP3 Result analysis；(C)平台自動整併產生2個檔案，可供下載利用或點選右框視窗內的SSR analysis進行簡單重複序列分析。



圖四、簡單重複序列分析的操作畫面。

For a more advanced usage of CAP3, it is recommended to install the original software on your local computers. CAP3 is freely available from Xiaoqiu Huang's [web site](#).

Enter your sequence in FASTA format:

續的SSR分析工作。

(二) 簡單重複序列分析

欲利用前述已獲得的單一序列資料繼續進行簡單重複序列分析，只要點選圖三-B右框視窗內的「CAP3 Result analysis」項目，本平台會將6個序列接合檔案內所分別存放的接合序列、單獨成群序列及其各種基本訊息，再整併成可供執行後續SSR分析的兩個檔案，如圖三-C左框視窗所示，第1個檔案涵蓋所有接合序列與單獨成群序列所組成的單一序列資料，第2個檔案是這些單一序列的相關資訊，此時點選右框視窗內的「SSR analysis」項目，可進入SSR分析執行畫面（圖四）。該分析執行的來源序列，未必要自動取自於前述CAP3分析結果，也可直接貼上使用自己輸入且符合FASTA格式之序列資料。

在SSR分析執行畫面中，如圖四所示，先針對基本重複單位大小(Unit size)勾選

id	SequenceID	Item	UnitSize	Repeats	Start	End
1	gi BI807492	1	cag	7	223	243
2	gi BI807560	1	gaa	6	32	49
3	gi BI807665	1	tgt	7	346	366
4	gi BI807886	1	tggtg	6	482	505
5	gi BI807980	1	tgc	6	383	400
6	gi BI808063	1	atc	6	145	162
7	gi BI808188	1	gaa	8	178	201
8	gi BI808726	1	ccg	6	494	511
9	gi BI808965	1	cca	6	127	144
10	gi BI808965	2	cgcg	5	212	231
11	gi BI809012	1	tat	7	388	408
12	gi BI809083	1	cagca	5	114	138
13	gi BI809237	1	ccg	7	526	546
14	gi BI809251	1	atat	9	130	165
15	gi BI809291	1	atc	6	261	278

You can download: [The SSR RESULTS](#)

將分析結果下載另存新檔

圖五、簡單重複序列分析的結果。

篩選條件，同時視實際情況調整各勾選項中基本重複單位之最小重複次數(minimum number of repeats)，以本例水稻基因組以重複3、4、5個鹼基為最常見，故可勾選Unit size 3、4、5三項，而這三項的最小重複次數分別為6、5、5（此也是本平台的內設值），最後按視窗最下面的「執行」鍵，即開始執行分析。

分析結果內容顯示如圖五，顯示出符合篩選條件下各序列的SSR之樣態、重複次數及位置等相關訊息，如以本實例分析結果的編號BI807492序列來看，發現在該序列223-243bp位置上有連續7次重複的cag鹼基存在，最後在該結果視窗下方提供有下載功能，可將分析結果另存新檔，以利未來進一步的引子設計。