台灣農業研究 (J. Taiwan Agric. Res.) 57(2):103–118 (2008)

水稻表現序列標幟 (EST) 功能分析自動化系統[1]

呂秀英[2] 陳政道[2] 呂椿棠[2,4] 邱怡嘉[3] 魏夢麗[2]

摘　要

呂秀英、陳政道、呂椿棠、邱怡嘉、魏夢麗。2008。水稻表現序列標幟 (EST) 功能分析自動化系統。台灣農業研究 57:103–118。

表現序列標幟 (expressed sequence tag, EST) 是 cDNA 序列片段，直接研究 EST 可獲取基因表現訊息。透過 EST 的功能註解、分類和表現量統計，能瞭解不同 cDNA 基因庫間之基因表現型式的差異性及表現基因的功能類別。現存於公共資料庫的水稻 EST 序列數量相當龐大，可作爲水稻基因組功能註解的重要資源。但針對這些巨量 EST，僅靠人工是不可能在短時間內完成所有分析與計算。因此，本研究在 Linux 伺服器以 MySQL 資料庫管理系統建立本地資料庫，並綜合 Perl 和 PHP 程式語言之運用，成功研發了一套完整的水稻 EST 功能分析自動化系統，能快速且準確完成 EST 功能註解、分類及表現量統計的全程分析工作，大大提高了工作效率。使用者無須登錄到 Linux 分析平台，只要在 Microsoft Window 操作環境下透過網頁瀏覽器，即可直接使用該系統。系統以 Web 界面爲基礎所設計的輸入畫面在操作上非常簡便，且分析完成後主動將輸出結果透過電子郵件寄給使用者。本系統亦可容易地擴展運用到其他物種之 EST 功能分析工作，對分子遺傳實驗室發展 EST 分析工作具有重要意義。

關鍵詞： 水稻、表現序列標幟、功能註解及分類、基因表現量統計、自動化系統、Perl 語言。

前　言

1991 年 Adams 等人從三種人腦組織的 cDNA 基因庫 (library) 中隨機挑選出 609 個選殖體 (clone) 進行定序，從而得到一組人腦組織的表現序列標幟 (expressed sequence tag, EST)，並將其與資料庫進行序列同源性對比，結果顯示該組中有 36 個代表已知基因，337 個代表未知基因，這是關於 EST 技術應用的首次報導，並首次提出了 EST 的概念 (Adams *et al.* 1991)。EST 雖然只是 cDNA 的片段序列，但具快速且低廉的優點，故成爲連接結構基因組學和功能基因組學的一道橋

1. 行政院農業委員會農業試驗所研究報告第 2318 號。接受日期：97 年 6 月 17 日。
2. 本所作物組研究員、計畫助理、助理研究員與助理研究員。台灣 台中縣 霧峰鄉。
3. 私立朝陽科技大學生物技術研究所碩士。台灣 台中縣 霧峰鄉。
4. 通訊作者，電子郵件：tang@wufeng.tari.gov.tw；傳真機：(04)23390528。

樑，進行 EST 分析可提供基因辨識、新基因發現及基因表現差異性等研究，是最快速獲得序列資訊及提供功能性基因組研究之來源 (Adams *et al*. 1991, 1992; Polymeropoulos *et al.* 1993; Bonaldo *et al*. 1996; Rounsley *et al*. 1996; Ewing *et al*. 1999; Jongeneel 2000; Kantety *et al.* 2002; Rudd 2003; Wei *et al*. 2005a, b)。隨著 EST 定序的快速發展，至 2008 年 3 月 28 日止，美國國家生物技術資訊中心 (National Center for Biotechnology Information, NCBI, http://www.ncbi.nlm.nih.gov/) 的 dbEST 資料庫已登錄不同物種之不同組織的 EST 共 50,818,319 條，其中以人和鼠的最多 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html)。目前全世界研究單位都在針對禾穀類作物或其他重要作物進行大規模的 EST 計畫，EST 資料仍在快速增加中 (Kantety *et al*. 2002)。至 2008 年 3 月 28 日止，登錄在 NCBI 之 dbEST 資料庫的水稻 EST 序列數量已累積近 1,220,261 條 (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html)。從公共資料庫下載取得 EST 以進行分析之前，爲維持序列的品質，通常都會利用 CAP3 或其他接合軟體的聚集 (clustering) 與接合 (assembly) 之常規化處理，減少 EST 之重疊性，並將其連接成更長、更高質量的序列，此稱爲單一序列 (unique sequences) (Adams *et al*. 1995)；然後再透過該 EST 之單一序列的功能註解 (functional annotation)，以瞭解這個序列在生物體中所扮演的功能。透過已註解的 EST 序列，可以進行其上下游資訊的全面分析，包括新基因發現、重複序列測定、調控基因確定、完整基因分析等 (Jongeneel 2000; Wolfsberg & Landsman 2001; Sreenivasuls *et al*. 2002; Wei *et al*. 2005a)。

一般針對 EST 功能註解的方法，通常是將 EST 單一序列直接利用 BLAST (basic local alignment search tool) 去比對 NCBI 資料庫 (http://www.ncbi.nih.gov/BLAST/) 或 MIPS (Munich Information Center for Protein Sequences) 之資料庫 (MAtDB, http://mips.gsf.de/proj/plant/jsf/athal/searchjsp/searchSequence.jsp)，然後再將所比對到的相似性序列進一步利用 MIPS 功能分類系統 (MIPS Functional Catalogue, MIPS_FunCat, http://mips.gsf.de/proj/funcatDB/search_main_frame.HTML) 來進行功能分類 (Lai *et al*. 2004; Shim *et al*. 2004; Brenner *et al*. 2005)。此程序必須經由一再核對及篩選的反覆過程，否則無法迅速且準確地獲知其真正的功能。基此，Chiu *et al*. (2007) 成功地建立了一套完整且快速的新流程：將 EST 單一序列比對 TIGR (The Institute for Gene Research) 之 Gene Index (TGI) 資料庫 (http://compbio.dfci.harvard.edu/tgi) 的假設性一致性序列 (tentative consensus sequence, TC) 進行功能註解，再藉由 TC 搜尋結果連結到 GO (Gene Ontology, http://www.geneontology.org/) 與 MIPS 之功能分類系統進行功能分類，最後進行基因功能表現量之統計。對於少量資料，可將其提交到上述各相關網站，即可完成分析；但對於大量的 EST 資料，所需花費的時間和人力仍相當可觀，因爲利用網際網路在多個網站間交互操作，不僅存在速度慢、步驟繁瑣、缺乏針對性等缺點，而且保密性差，難以滿足研究工作需要。另外由於不同網站資料庫的操作方式與資料儲存格式不盡相同，故常常得花很多時間於各輸入輸出檔案間的格式轉換與資料整理，此過程中也容易出錯。隨著 EST 量越來越多，處理巨量生物資料的高度複雜性，僅靠人工完成所有分析與計算是難以想像的。而唯一解決之道，是將整個分析流程予以全面自動化。因此，本研究以 Chiu *et al*. (2007) 所建立的 EST 功能註解分析流程爲基礎，先從網路上將 TGI 中 TC 資料庫的相關序列及其訊息敘述、GO 分類資料庫的代碼及敘述、MIPS 分類目錄以及 BLAST 序列比對軟體，下載到自有 Linux 主機所構築的資料庫內，以建置一個可以在本地直接進行序列比對、功能註解和分類的操作平台；然後編寫程式，將整個分析流程予以自動化並設計簡單易用的使用介

面，使能快速準確地完成 EST 功能註解、分類及表現量統計的全程分析工作。此本地序列資料分析平台及自動化系統之建構完成，對於分子遺傳實驗室發展 EST 分析工作具有重要意義。

材料與方法

EST 功能註解分析之流程

Chiu *et al*. (2007) 擷取 NCBI 之 dbEST 資料庫中受褐飛蝨誘導之 188 條水稻 EST 序列，將其經 CAP3 軟體聚集和接合後，共獲得 155 條較高品質和去除多餘性的單一序列，以此單一序列爲材料建立了一套 EST 功能的註解、分類及表現量統計的分析流程，如圖 1 所示：(1) 首先將 EST 經聚集及接合後所產生的單一序列，去比對 TIGR Gene Index (TGI) 資料庫的 TC，以進行功能註解；(2) 藉由 TC 搜尋結果連結至相對應之 GO 分類代碼及敘述，可進行功能註解之粗分類；(3) 再經由 GO 分類代碼連結至相對應之 MIPS 分類代碼目錄，以進行功能註解之細分類；(4) 最後統計各功能類別的基因表現量及其比例。

TGI 資料庫蒐集的基因資料不僅簡單對 NCBI 之 GenBank 資料庫的基因和 EST 數據按物種分類，也提取 GenBank 資料庫中不同物種的 EST 和註解的基因序列，經由聚集、接合等分析過程以產生單一的、具有高可信度的 TC。因此可藉此網站輸入查詢序列進行比對後，找出相似的 TC，來推測基因之功能註解。GO 主要提供結構及動態模式的控制字彙和分類系統，真核生物的基因和蛋白質也可透過 GO 的整合性分類系統得到分子功能 (molecular function)、生物程序 (biological process)、細胞成分 (cellular component) 三個層次的註解，可對基因組資訊進行簡單分類及描述 (GO 2006)。MIPS 分類系統的功能分類目錄 (MIPS_FunCat, http://mips.gsf.de/proj/funcatDB/search_main_frame.HTML) 是一個階層系統，由二十多個主要功能種類所組成，主功能底下又有其細分類別，迄今最新 2.1 版已增加到 28 個主功能共包含 1362 項之功能類別，可用來對不同生物之基因組進行註解之分類，是研究者對基因及其產物進行功能分類時常參考的資料庫 (Andreas *et al*. 2004; Mewes *et al*. 2006, 2008)。

序列資料分析平台之建置

本研究所研發的自動化分析系統安裝在 Linux 伺服器內，用於建構該 Linux 序列資料分析平台的硬體配備爲 Pentium III、CPU 650 MHz、1GB RAM，及內含 80G 的 SCSI 硬碟。作業系統爲 RedHat Linux 9.0，可從 http://www.redhat.com 或其他鏡像站點取得。該類小型伺服器的花費較少，大部分實驗室都有能力配置。

由於生物資訊分析通常涉及巨量資料，資料的儲放必須有一個高效率的資料庫，才能利於管理、存取及後續分析。因此，本研究在 Linux 作業平台上，利用 PHP 及 MySQL 建立一個資料庫，專門用來儲存待分析之 EST 單一序列和所有分析結果。而爲了建置一個可以在本地直接進行序列比對、功能註解和分類的操作平台，也從網路上下載 TGI 中 Rice (*Oryza sative*) 之 TC 資料庫 (含序列及相關訊息敘述)、GO 分類資料庫 (含代碼及敘述) 和 MIPS 分類目錄，以及 BLAST 序列比對軟體，並將它們儲存在本地資料庫中。根據 Chiu *et al*. (2007) 所建立之水稻 EST 功能註解分析流程 (圖 1)，再利用 Perl 語言來撰寫水稻 EST 功能分析自動化系統之 Perl 程式模組。

相關軟體之獲取與安裝

建構分析平台和自動化系統所使用的軟體皆爲免費軟體，可以從網際網路下載或直接向作者索取，而在 RedHat Linux 9.0 作業系統安裝後即已內建 Perl 5.6、PHP 2.6.4 及 MySQL 3.23.5。由於這些軟體皆安裝在 Linux 操作系統中，不能移植到 Microsoft Windows 系統，但 Windows 使用者仍可以透過一般網頁瀏覽器 (如 IE, Netscape, Sleipnir 等) 輸入 Linux 伺服器的所在 IP 網址，就能連結使用本系統，而未必要藉由終端模擬軟體或遠端連線軟體登錄到 Linux 伺服器上使用。本研究相關軟體之功能，簡單說明如下：

序列相似性比對軟體 BLAST：BLAST 爲目前生物資訊學 (Bioinformatics) 研究上極爲重要的工具，特別是在核酸或胺基酸序列資料庫的搜尋，堪稱目前最廣泛使用的序列相似性比對方法。此法的演算式和統計學基礎由 Altschul 等人於 1990 年發表 (Altschul *et al*. 1990) 後，經陸續改進，至 1996 年成爲 NCBI 最重要的網路伺服工具。BLAST 使用了局部 (local) 而非全域性 (global) 的序列比對，而再採用字串比對 (word hashing) 時，加入了相似度門檻值 (similarity threshold) 的概念使得長字串不需全部一致而大幅提高搜尋速度，並利用極端分布統計 (extreme value distribution statistics) 的概念進行信賴度分析而設計了 E-value，在速度與搜尋靈敏度上獲得極佳的平衡 (Altschul & Gish 1996)。在給定條件的搜尋下，所找到的相似序列之清單中的 Score 是這一相似序列中最相似的區域之得分，代表統計學上的可能性，得分越高，越不可能是逢機發生，即越有可信度。但該值爲絕對值，與所搜尋的資料庫大小無關。而 E-value 是一種期望值，與搜尋空間大小有關，其意義是在這樣的資料庫搜尋空間中可逢機獲得這樣高分之序列的可能性，因此 E-value 越高，

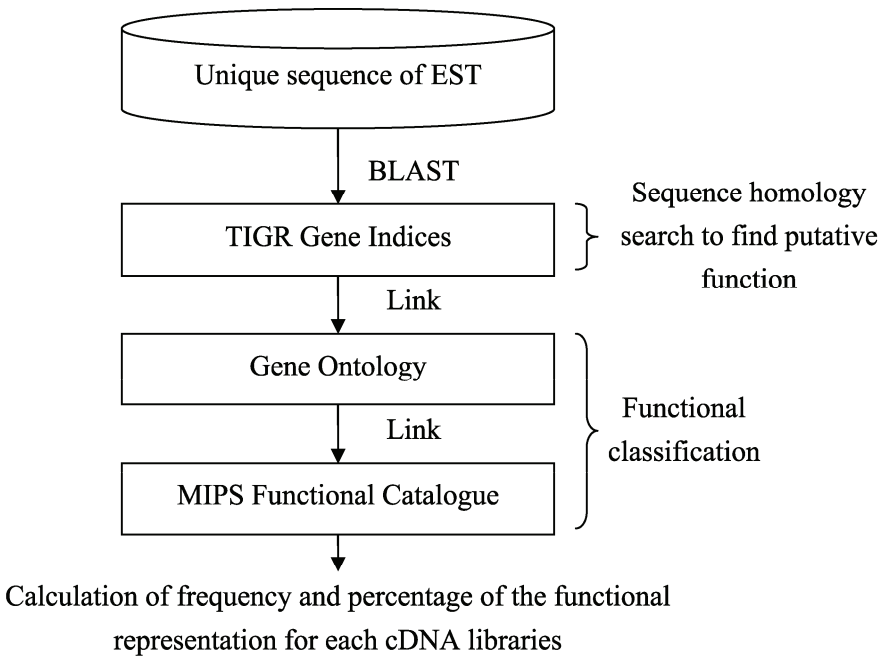


圖 1. Chiu *et al*. (2007) 所建立之 EST 功能註解分析的流程圖。

Fig. 1. Flowchart showing the analysis procedures of EST functional annotation established by Chiu *et al*. (2007)

則代表越有可能是逢機獲得的，也就越不可信，或越不具有生物意義。BLAST 可由 NCBI 網站下載 (http://www.ncbi.nih.gov/BLAST/)，用來進行本地序列的相似性分析。

程式語言 Perl：Perl 是 practical extraction and report language (實用擷取與報告語言) 的縮寫，爲 Larry 於 1980 年代中期所創，最初是 UNIX 作業系統所開發的編製程式語言，現在具有強大的跨平台特性，幾乎不受限制，如 Linux、Microsoft Windows、Macintosh、MVS、VMS、OS/2、Plan9、Amiga 等作業系統也都有相應的 Perl 語言解釋器；Perl 是一種腳本語言 (script language)，對程序、檔案和文字有很強的處理與變換能力，因此凡是有關系統工具、軟體工具、系統管理、資料庫連結、圖像程序設計、網路連結和 WWW 程序設計等任務，都特別適合用 Perl 來執行 (Larry *et al*. 2000; Schwartz *et al*. 2005)。Perl 十分容易學習和掌握，大多數的任務只需要 Perl 語言的一小部分即可完成。Perl 也是一種解釋式的語言，即只要寫了程式後，不需經由一道中間的編碼過程即可測試，可快速、方便地測試及除錯。這些優點使得 Perl 成爲普及於全球資訊網中的編製程式語言，如同系統和資料庫與使用者之間的通道，並已成立專屬官方網站 (http://www.perl.org/)。現在 Perl 語言已廣泛應用在生物資訊學研究 (Jagota 2004; Jan *et al*. 2004; Haynes *et al*. 2006; Zhou *et al*. 2007)。

HTML 輸出程式 PHP：PHP 是 hypertext preprocessor (超文件前置處理器) 的縮寫，用來產生 HTML 網頁原始檔的中介程式及語言，屬於一種伺服端內嵌式 HTML 的應用程式；不是用一大堆指令來輸出 HTML 程式，而是直接可以在 PHP 和 HTML 間切換，尤其適合用來 Web 開發。它容易學習，執行速度快，可同時運行於 Windows、UNIX 和 Linux 不同作業平台的 Web 程序，內置了對文件上傳、密碼認證和郵件收發等功能 (Williams & Lane 2004; Trachtenberg & Sklar 2006)。PHP 也有其專屬的官方網站 (http://www.php.net/)。透過 PHP 程式之編寫設計，即便本系統是建立在 Linux 分析平台上，也能讓 Windows 使用者直接透過網頁瀏覽器來輸入資料並讀取輸出結果，而未必要登錄到 Linux 伺服器才能使用該系統。

資料庫管理系統 MySQL：MySQL 是一個快速、容易使用且功能強大的關聯式資料庫管理系統 (relational database management system, RDBMS)，可容易地與 C、C++、Java、Perl、PHP 等語言連結，且可運作在許多作業平台上，例如 Linux、HP-UX、Windows 等平台。自 1996 年以來，MySQL 得到很多大公司的支持，其功能性和穩定性都有很大的發展，目前共超過 40 個資料庫，包含 10,000 個表，其中 500 多個表超過 7 百萬行，大約有 100GB 容量的主要應用資料，能快速且靈活地儲存並記錄文件和影像，故已有許多配置和管理 MySQL 伺服器的圖形工具被開發，爲系統管理員提供極大的幫助 (Williams & Lane 2004; DuBois 2006)。MySQL 現也已成立專屬官方網站 (http://www.mysql.com)。透過 MySQL，能提供各研究室的 Web 網站或個人網站一個簡單、方便及高效能的資料庫系統。

系統之效率評估

爲確認系統的效益，有必要以實際序列資料進行評估。Chiu *et al.* (2007) 擷取 NCBI 之 dbEST 資料庫中受褐飛蝨誘導之 188 條水稻 EST 序列，將其經 CAP3 軟體聚集和接合後所得到的 155 條單一序列爲材料，採用人工方式進行功能註解，自 TC 資料庫相似性序列的蒐集比對至最終完成功能的註解、分類及統計，前後共花了大約 3 個月的時間。因此，本研究將首先以此 155 條單一序列來測試以自動化系統完成全程分析工作所需的時間。另外，Lu *et al*. (2007) 取自 NCBI 之 dbEST 中受稻熱病菌誘導之共 84,705 條水稻 EST，利用 CAP3 將其聚集和接合後，共得到 32,165 條 EST 單

一序列，此巨量的單一序列是絕不可能利用人爲方法來完成功能註解之分析工作，因此也是利用此自動化系統來完成所有分析。

結　果

使用介面操作

本研究在 Linux 主機以 MySQL 資料庫管理系統建立本地資料庫，並綜合 Perl 和 PHP 程式之運用，成功研發了一套完整的水稻 EST 功能分析自動化系統。透過 PHP 程式設計，使用者無須登錄到 Linux 主機平台，只要在 Microsoft Window 操作環境下透過網頁瀏覽器，即可使用該系統。使用者執行介面之畫面，設計如圖 2。它在操作上十分簡單易明，只要先將 FASTA 格式的序列資料，

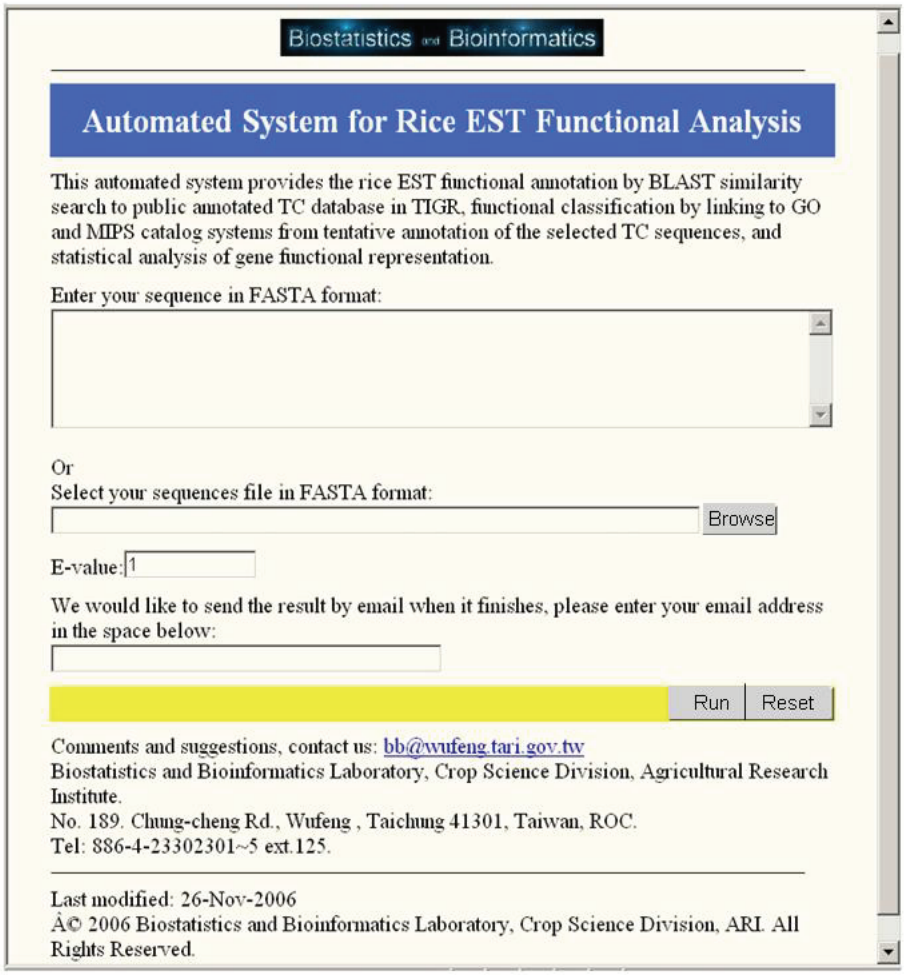


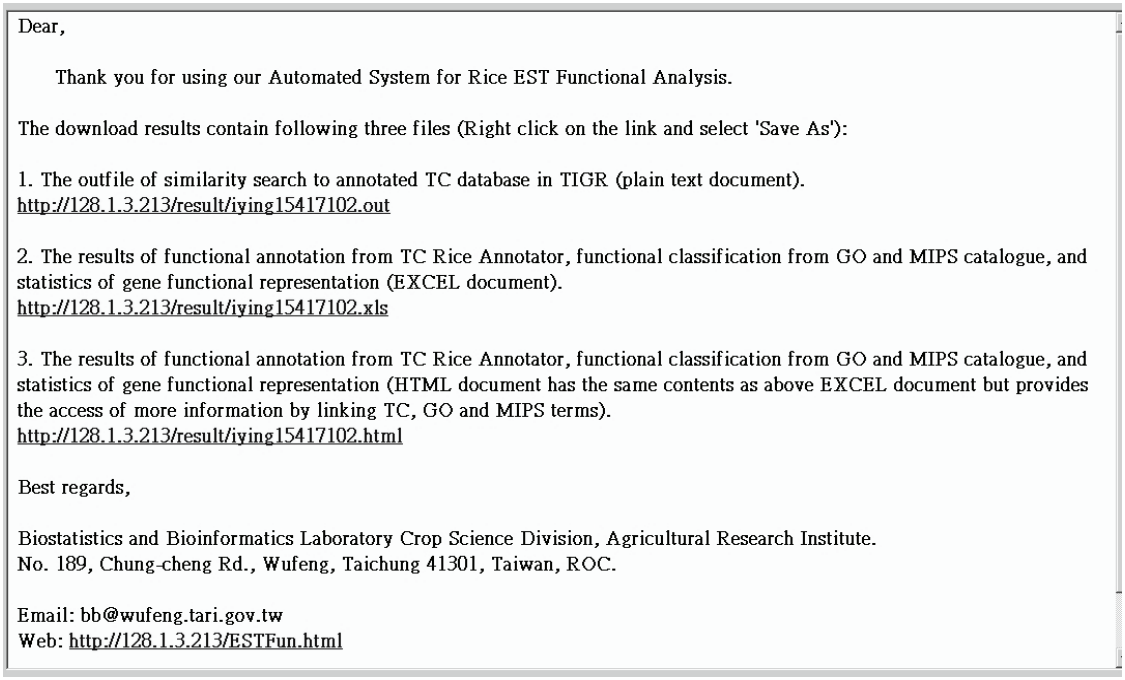
圖 2. 水稻 EST 功能分析自動化系統的輸入介面。

Fig. 2. Input interface of automated system for rice EST functional analysis.

用複製貼上之方式貼於畫面上的序列輸入框內，或者利用「Browse (瀏覽)」功能指定到個人電腦硬碟內儲存序列檔案的所在位置，再於最下面輸入框內填入使用者的電子信箱地址，最後按「Run (執行)」功能鍵即可。若在送出前按「Reset (重設)」功能鍵，則可清空已輸入的所有資料。系統在確認收到查詢序列後，會顯示出感謝使用本系統等字樣，然後將接下來的分析工作交給 Linux 主機執行。系統在完成分析後，會主動將結果以電子郵件寄給使用者，因此操作畫面中的電子信箱地址是必要輸入的。畫面中序列輸入框之下方的 E-value 欄位，系統內設值爲 1，但允許使用者自行修改將其降低或提高；降低 E-value 可將相似性比對的門檻變高，於是對 TC 所能搜尋的相似性序列會變得較少，反之若提高 E-value，則能比對到的序列可較多。

輸出結果內容

系統完成分析後，使用者會收到電子郵件通知，郵件內容如圖 3 所示。使用者可由收到的電子郵件中，直接用滑鼠點選相關檔案名稱，即可將它們直接開啓或下載。爲有利於使用者的各種後續運用，本系統將所分析的結果，提供了文字、Microsoft Excel 和 HTML 之三種不同儲存格式之輸出檔案，其附加檔名分別爲 out、xls 和 html。這三種檔案的命名原則，是以使用者所提供之電子信箱@之前的名稱，再加上一個流水號。第一個附加檔名爲 out 的輸出檔案，是可由網頁瀏覽器或任何文書製作軟體開啓的文字檔，內容爲查詢序列經 BLAST 比對 TIGR 的相似性搜尋結果，如圖 4 所示，輸出報表中會將所有小於指定 E 值 (系統內設值 = 1) 的相似性搜尋結果全部列出。至於


Dear,

Thank you for using our Automated System for Rice EST Functional Analysis.

The download results contain following three files (Right click on the link and select 'Save As'):

1. The outfile of similarity search to annotated TC database in TIGR (plain text document).
http://128.1.3.213/result/iying15417102.out

2. The results of functional annotation from TC Rice Annotator, functional classification from GO and MIPS catalogue, and statistics of gene functional representation (EXCEL document).
http://128.1.3.213/result/iying15417102.xls

3. The results of functional annotation from TC Rice Annotator, functional classification from GO and MIPS catalogue, and statistics of gene functional representation (HTML document has the same contents as above EXCEL document but provides the access of more information by linking TC, GO and MIPS terms).
http://128.1.3.213/result/iying15417102.html

Best regards,

Biostatistics and Bioinformatics Laboratory Crop Science Division, Agricultural Research Institute.
No. 189, Chung-cheng Rd., Wufeng, Taichung 41301, Taiwan, ROC.

Email: bb@wufeng.tari.gov.tw
Web: http://128.1.3.213/ESTFun.html

圖 3. 水稻 EST 功能分析自動化系統完成分析後主動寄給使用者的電子郵件訊息，隨函共附 3 個輸出檔。

Fig. 3. An e-mail message attaching with three output files from rice EST automated functional analysis system was sent to the user automatically after accomplishing the analysis.

```
BLASTN 2.2.11 [Jun-05-2005]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= Contig1
         (237 letters)

Database: tc
           36,381 sequences; 54,320,236 total letters

Searching..................................................done

                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

TC275121 GB|AAP47473.1|32331891|AY164899 RTNLB20 {Oryza sativa (...   389   e-108
TC261271 similar to GB|BAC22440.1|24414198|AP005452 alcohol dehy...    46   2e-04
TC256026 weakly similar to GB|AAP42759.1|30984592|BT008746 At1g2...    46   2e-04
TC258321                                                               42   0.002
TC248959 similar to UP|H136_ARATH (O82660) Photosystem II stabil...    42   0.002
TC261368 UP|O80422 (O80422) Carbonic anhydrase, partial (96%)          40   0.010
TC251134 B1003B09.10 [Oryza sativa (japonica cultivar-group)]          38   0.039
TC284026 Oryza sativa (japonica cultivar-group) cDNA clone:002-1...    36   0.15
TC277397 UP|Q6SZS7 (Q6SZS7) Cytosolic NADP malic enzyme, complete      36   0.15
TC261619 Oryza sativa (japonica cultivar-group) cDNA clone:J0330...    36   0.15
TC266761 Oryza sativa (japonica cultivar-group) cDNA clone:J0230...    34   0.60

>TC275121 GB|AAP47473.1|32331891|AY164899 RTNLB20 {Oryza sativa (japonica
            cultivar-group);} , complete
          Length = 1876

 Score =  389 bits (196), Expect = e-108
 Identities = 199/200 (99%)
 Strand = Plus / Minus
Query: 19   ggtacacctgagaactcgcaaacccgacaaaagacgccacccttccgaatcgagtcccct 78
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1078 ggtacacctgagaactcgcaaacccgacaaaagacgccacccttccgaatcgagtcccct 1019


Query: 79   tgcgcaacttattctagttcttctttggcttcgccgggcccctgggtatcttgctcaaaa 138
            |||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
Sbjct: 1018 tgcgcaacttattctagttcttctttggcttcaccgggcccctgggtatcttgctcaaaa 959


Query: 139  ccttggcatccagcacctcatagtgcttgcaggcctcagagtgagcccttccagcaaagt 198
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 958  ccttggcatccagcacctcatagtgcttgcaggcctcagagtgagcccttccagcaaagt 899


Query: 199  ggtcgaccttatcctggtac 218
            ||||||||||||||||||||
Sbjct: 898  ggtcgaccttatcctggtac 879


>TC261271 similar to GB|BAC22440.1|24414198|AP005452 alcohol dehydrogenase-like
            protein {Oryza sativa (japonica cultivar-group);} ,
            partial (72%)
          Length = 1202

 Score = 46.1 bits (23), Expect = 2e-04
 Identities = 23/23 (100%)
 Strand = Plus / Minus


Query: 215  gtacctgcccgggcggccgctcg 237
            |||||||||||||||||||||||
Sbjct: 1052 gtacctgcccgggcggccgctcg 1030
```

圖 4. 水稻 EST 功能分析自動化系統的輸出文字檔內容範例：比對 TIGR 之 TC 資料庫的相似性搜尋結果。

Fig. 4. An example of output text file from rice EST automated functional analysis system: results of similarity search to annotated TC database in TIGR.

Excel 和 HTML 兩種檔案的輸出結果完全相同，皆包含三部份內容，如圖 5 所示，此處僅以網頁形式檔爲例說明：(1) 各查詢序列 (query) 所比對到相似 TC 的 id 及其定義、該 TC id 連結到 GO 功能分類系統所得之 GO id 及其註解粗分類、再經由該 GO id 連結到 MIPS_FunCat 功能分類系統所得之 MIPS id 及其註解細分類，這些分析結果皆被整合在同一個表格內，以利相互參照 (圖 5-1)。以圖中第一條查詢序列 contig1 爲例，在內設 E value = 1 之條件下，比對到的所有相似 TC 有 TC275121、TC261271 等 (因版面限制，圖中僅列出部份結果)，由這些相似 TC 之 id 所對應到的 GO 和 MIPS 分類代碼，可知該 EST 單一序列與何種功能類別之基因表現有關。畫面中也提供超連結功能，讓使用者藉由點選 TC、GO 及 MIPS id，就能直接連結到其相關網站以瀏覽該 id 的詳細資訊。(2) 依據所有查詢序列比對到的 GO 分類代碼，統計在 GO 功能分類系統下所佔表現量及其比例 (圖 5-2)。以圖中實例而言，這些序列比對到分子功能 (molecular function, F)、生物程序 (biological process, P) 及細胞成分 (cellular component, C) 之功能類別表現量各爲 98、91 及 85，各

List of TIGR-GO-MIPS Hits

Query	TC id	TC term	GO id	GO term	Type	MIPS id
Contig1	TC275121	RTNLB20 {Oryza sativa (japonica cultivar-group);}, complete	GO:0005783	endoplasmic reticulum	C	MIPS_funcat:70.07
Contig1	TC261271	homologue to (Q8LIC3) Putative sex determination protein tasselseed 2, partial (93%)	GO:0016491	oxidoreductase activity	F	
Contig1	TC256026	(Q9AWN4) P0504D03.14 protein (OSJNBa0054L14.23 protein), complete	GO:0006810	transport	P	MIPS_funcat:20
Contig1	TC256026	(Q9AWN4) P0504D03.14 protein (OSJNBa0054L14.23 protein), complete	GO:0006839	mitochondrial transport	P	MIPS_funcat:20.09.04
Contig1	TC256026	(Q9AWN4) P0504D03.14 protein (OSJNBa0054L14.23 protein), complete	GO:0005488	binding	F	
Contig1	TC258321	(O10247) Nonstructural protein 1 (Fragment), partial (4%)				
Contig1	TC248959	similar to (O82660) Photosystem II stability/assembly factor HCF136 chloroplast precursor, partial (77%)	GO:0030095	chloroplast photosystem II	C	

圖 5-1. 水稻 EST 功能分析自動化系統輸出的 HTML 網頁檔內容範例：從相似 TC 連結對應到的 GO 及 MIPS 代碼之相關資訊。

Fig. 5-1. An example of HTML output file from rice EST automated functional analysis system: Hits information of GO and MIPS ids linked from similar TC sequences in TIGR.

Statistics of Gene Functional Representation based on GO category

Molecular Function (F)	Hit	%
	98	35.76
Category		
Antioxidant activity	0	0
Binding	75	10.59
Catalytic activity	65	9.180
Chaperone regulator activity	0	0
Enzyme regulator activity	6	0.847
Molecular function unknown	39	5.508
Motor activity	2	0.282
Nutrient reservoir activity	0	0
Obsolete molecular function	13	1.836
Protein tag	0	0
Signal transducer activity	3	0.423
Structural molecule activity	8	1.129
Transcription regulator activity	8	1.129
Translation regulator activity	3	0.423
Transporter activity	12	1.694
Triplet codon-amino acid adaptor activity	0	0

Biological Process (P)	Hit	%
	91	33.21
Category		
Behavior	1	0.141
Biological process unknown	32	4.519
Cellular process	82	11.58
Development	21	
Obsolete biological process	1	0.141
Physiological process	85	12.00
Regulation of biological process	22	3.107
Viral life cycle	0	0

Cellular Component (C)	Hit	%
	85	31.02
Category		
Cell	78	11.01
Cellular component unknown	63	8.898
Extracellular matrix	0	0
Extracellular region	0	0
Obsolete cellular component	0	0
Organelle	72	10.16
Protein complex	38	5.367
Virion	0	0

圖 5-2. 水稻 EST 功能分析自動化系統輸出的 HTML 網頁檔內容範例：GO 功能分類系統下各功能類別之基因表現量及其比例。

Fig. 5-2. An example of HTML output file from rice EST automated functional analysis system: Frequency and percentage of gene representation of functional classification based on GO functional category.

佔 35.76、33.21 和 31.02%。GO 分類系統除了三個層次的粗註解外，其下層仍有其自有的更細分功能，由於 GO 作爲基因功能註解有逐漸增加之趨勢，故本系統也同時再提供其第二層之細分功能類別的統計資料，以供參考。(3) 依據所有查詢序列比對到的 MIPS 分類代碼，統計在 MIPS 分類系統下各功能類別所佔表現量及其比例 (圖 5-3)。MIPS 分類目錄是一個階層系統，但一般皆以第一層的主功能類別來探討基因表現量之差異性，因此本系統僅就主功能類別進行統計。以圖 5-3 之實例而言，藉由 MIPS 分類系統，可知該 cDNA 基因庫表現出代謝 (metabolism) 功能的基因表現量佔 14.368%、能量 (energy) 功能佔 2.586%等。

Statistics of Gene Functional Representation based on MIPS category

MIPS	Hit	%
Metabolism	50	14.368
Energy	9	2.586
Storage protein	0	0
Cell cycle and DNA processing	5	1.437
Transcription	3	0.862
Protein synthesis	0	0
Protein fate (folding, modification, destination)	16	4.598
Protein with binding function or cofactor requirement (structureal or catalytic)	17	4.885
Regulation of metabolism and protein fucntion	0	0
Cellular transport, transport facilities and transport routes	53	15.23
Cellular communication / signal transduction mechanism	53	15.23
Cell rescue, defense and virulence	6	1.724
Interaction with the environment	2	0.575
Systemic interaction with environemnt	0	0
Transposable elements, viral and plasmid proteins	0	0
Cell fate	4	1.149
Development (Systemic)	1	0.287
Biogenesis of cellualr components	49	14.08
Cell type differentiation	0	0
Tissue differentiation	0	0
Organ differentiation	0	0
Subcellular localization	41	11.782
Cell type localization	0	0
Tissue localization	0	0
Organ differentiation	0	0
Classification not yet clear-cut	0	0
Unclassified proteins	39	11.207

圖 5-3. 水稻 EST 功能分析自動化系統輸出的 HTML 網頁檔內容範例：MIPS 功能分類系統下各功能類別之基因表現量及其比例。

Fig. 5-3. An example of HTML output file from rice EST automated functional analysis system: Frequency and percentage of gene representation of functional classification based on MIPS functional category.

除網頁形式的輸出檔案外，本研究還提供同樣內容的 Excel 輸出檔之目的，是方便使用者可將分析結果另用 Excel 軟體來開啓檢視，此有利於使用者在 Window 個人電腦上做進一步的其他各種統計分析。

系統效率評估

就系統分析的效率而言，Chiu *et al.* (2007) 直接將共 155 條受褐飛蝨誘導之水稻 EST 單一序列，提交到 TIGR 網站之 TC 資料庫進行相似性序列的蒐集比對，然後逐一用人工方式檢視，篩選出每條查詢序列所比對到的相似序列及其註解，再從其註解敘述中擷取出適當的關鍵字，分別輸入到 GO 及 MIPS 等網站找出該功能屬於何種類別，最後利用 Excel 試算表就每條查詢序列所對應到的各類別功能進行表現量個數和百分比統計，前後共花了大約 3 個月的時間。而使用本研究之 EST 功能分析自動化系統，約僅 30 分鐘左右的時間即完成所有序列的功能註解、分類及統計之全程分析工作。至於 32,165 條受稻熱病菌誘導之水稻 EST 的功能註解，由於數量非常龐大，是絕不可能經由人工方式來分析，但透過本自動化系統不用兩天即已完成。由此可知，本研究所建立之 EST 功能分析自動化系統確可節省大量資料之處理時間。

討　論

生物資訊學計算的最大特色就是要面對巨量資料之處理，惟有藉助於資料庫和操作自動化的設計技術，才能提高資料分析的效率與準確度效率。基此，本研究針對水稻 EST 之功能註解、分類及表現量統計分析，研發了一套自動化系統。該系統雖建立在 Linux 分析平台上，但使用者可在 Window 操作環境下透過網頁瀏覽器，以一個非常簡便的操作方法，無需面對繁瑣的參數和命令，也無需關注該程式如何工作，任務一旦提交將立即執行，具有簡單易用的優點，大大提高了工作效率。在分析完成後，系統會主動將將結果透過電子郵件提供超連結的檔案開啓和下載，讓使用者不必擔心分析過程中可能因網路中斷或個人電腦當機等意外而造成執行中斷，保證了研究的順利進行，達到事半功倍的效果。輸出結果的內容，同時提供了文字、Excel 和 HTML 之三種不同形式的檔案格式，更有利於後續的分析。更重要的是能夠一次快速且準確地完成 EST 註解、分類及表現量統計的完整分析工作；就本研究現有伺服器的硬體配備而言，雖只算是低階規格，但一次處理 155 條的序列只要約 0.5 小時即可完成，而處理 3 萬多條的大量資料，亦能於兩天內完成，節省了人工處理大量資料所需要的人力和時間。未來 Linux 分析平台的硬體配備若能予以提升，以加速執行速度，所需的時間當可更少。

在構築系統之前，本研究必須先到網路上下載 TIGR 網站之 TGI 的 TC 資料庫、GO 分類資料庫和 MIPS 分類目錄到本地伺服器，並爲確保本地資料庫的資料能維持最新，隨時留意版本更新，以不定期重新下載新的序列及目錄資料，是必要的工作。由於本系統建構之初期目標，在協助水稻基因組研究人員快速進行水稻 EST 的功能註解分析工作，因此目前僅自 TGI 中下載水稻 TC 至本地資料庫，但其實 TGI 尚蒐集了其他各種動植物和微生物等的 TC，未來只要再從 TGI 下載其他物種的 TC，則本系統將可輕易地擴展運用到其他物種之 EST 功能註解分析工作。目前本所正評估該系統之技術授權申請案，故暫不對外開放使用，僅公開給本所區域網路內部人員利用。

誌　謝

本研究承蒙行政院國家科學委員會補助經費 (NSC92-2313-B-055-006, NSC93-2313-B-055-002)，特此誌謝。

引用文獻 (Literature cited)

Adams, M. D., M. Dubnick, A. R. Kerlavage, R. Moreno, J. M. Kelley, T. R. Utterback, J. W. Nagle, C. Fields, and J. C. Venter. 1992. Sequence identification of 2,375 human brain genes. Nature 355:632–634.

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656.

Adams, M. D., A. R. Kerlavage, R. D. Fleischmann, R. A. Fuldner, C. J. Bult, N. H. Lee, E. F. Kirkness, K. G. Weinstock, J. D. Gocayne, O. White, G. Sutton, J. A. Blake, R. C. Brandon, M. W. Chiu, R. A. Clayton, R. T. Cline, M. D. Cotton, J. Earle-Hughes, L. D. Fine, L.M. FitzGerald, W. M. FitzHungh, J. L. Fritchman, N. S. M. Geoghagen, A. Glodek, C. L. Gnehm, M. C. Hanna, E. Hedblom, P. S. Hinkle Jr., J. M. Jelley, K. M. Klimek, J. C. Kelley, L. I. Liu, S. M. Marmaros, J. M. Merrick, R. F. Moreno-Palanques, L. A. McDonald, D. T. Nguyen, S. M. Pellegrino, C. A. Phillips, S. E. Ryder, J. L. Scott, D. M. Saudek, R. Shirley, K. V. Small, T. A. Spriggs, T. R. Utterback, J. F. Weidman, Y. Li, R. Barthlow, D. P. Bednarik, L. Cao, M. A. Cepeda, T. A. Coleman, E. J. Collins, D. Dimke, P. Feng, A. Ferrie, C. Fischer, G. A. Hastings, W. W. He, J. S. Hu, K. A. Huddleston, J. M. Greene, J. Gruber, P. Hudson, A. Kim, D. L. Kozak, C. Kunsch, H. Ji, P. S. Meissner, H. Olsen, L. Raymond, Y. F. Wei, J. Wing, C. Xu, G. L. Yu, S. M. Ruben, P. J. Dillon, M. R. Fannon, C. A. Rosen, W. A. Haseltine, C. Fields, C. M. Fraser, and J. C. Venter. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377 (6457 Suppl.):3–174.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Altschul, S. F. and W. Gish. 1996. Local alignment statistics. Methods Enzymol. 266:460–480.

Andreas, R., A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Güldener, G. Mannhaupt, M. Münsterkötter, and H. W. Mewes. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32:5539–5545.

Bonaldo, M. F., G. Lennon, and M. B. Soares. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6:791–806.

Brenner, W. D., M. S. Katari, D. W. Stevenson, S. A. Rudd, A. W. Douglas, W. N. Moss, R. W. Twigg, S. J. Runko, G. M. Stellari, W. R. McCombie, and G. M. Coruzzi. 2005. EST analysis in Ginkgo biloba: an assessment of conserved developmental regulators and gymnosperm specific genes. BMC Genomics 6:143.

Chiu, Y. C., M. L. Wei, C. T. Lu, and H. Y. Lu. 2007. A new procedure of functional annotation and analysis in expressed sequence tag (EST). Crop Environ. Bioinform. 4:49–64. (in Chinese with English abstract)

DuBois, P. 2006. MySQL Cookbook, 2nd ed. O'Reilly Media Inc. Press, CA. 975 pp.

Ewing, R. M., A. B. Kahla, O. Poirot, F. Lopez, and S. Audic, and J. M. Claverie. 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res. 9:950–959.

Gene Ontology Consortium. 2006. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 34:D322–D326.

Haynes, P. A., S. Miller, T. Radabaugh, M. Galligan, L. Breci, J. Rohrbough, F. Hickman, and N. Merchant. 2006. The wildcat toolbox: a set of perl script utilities for use in peptide mass spectral database searching and proteomics experiments. J. Biomol. Tech. 17:97–102.

Jagota, A. 2004. Perl for Bioinformatics, 2nd ed. Bioinformatics By The Bay Press, CA. 82 pp.

Jan, A. A., B. J. Jungenous, and M. A. M. Goenen, 2004. POSA: perl objects for DNA sequencing data analysis. MBC Genomics 5:60–64.

Jongeneel, C. V. 2000. Searching the expressed sequence tag (EST) database: planning for genes. Brief. Bioinform. 1:76–92.

Kantety, R. V., M. L. Rota, D. E. Matthews, and M. E. Sorrells. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol. 48:501–510.

Lai, J., N. Dey, C. S. Kim, A. K. Bharti, S. Rudd, K. F. X. Mayer, B. A. Larkins, P. Becraft, and J. Messing. 2004. Characterization of the maize endosperm transcriptome and its comparison to the rice genome. Genome Res. 14:1932–1937.

Larry, W., T. Christiansen, and J. Orwant. 2000. Programming Perl, 3rd ed. O'Reilly Media Inc. Press, CA. 1104 pp.

Lu, H. Y., Y. C. Chiu, C. T. Lu, M. L. Wei, and C. T. Chen. 2007. Comparative Analysis of Expressed Sequence Tag (EST) in Rice Induced by *Magnaporthe grisea*. J. Taiwan Agric. Res. 56:261–280. (in Chinese with English abstract)

Mewes, H. W., D. Frishman, K. F. X. Mayer, M. Münsterkötter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stümpflenand. 2006. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 34:169–172.

Mewes, H. W., S. Dietmann, D. Frishman, R. Gregory, G. Mannhaupt, K. Mayer, M. Muensterkötter, A. Ruepp, M. Spannagl V. Stuempflen, and T. Rattei. 2008. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res. 36:196–201.

Polymeropoulos, M. H., H. Xiao, J. M. Sikela, M. Adams, J. C. Venter, and C. R. Merril. 1993. Chromosomal distribution of 320 genes from a brain cDNA library. Nat. Genet. 4:381–386.

Rounsley, S. D., A. Glodek, G. Sutton, M. D. Adams, C. R. Somerville, J. C. Venter, and A. R. Kerlavage. 1996. The construction of *Arabidopsis* expressed sequence tag assemblies. A new resource to facilitate gene identification. Plant Physiol. 112:1177–1183.

Rudd, S. 2003. Expressed sequence tags: alternative or complement to whole genome sequence? Trends Plant Sci. 8:321–329.

Schwartz, R. L., T. Phoenix, and B. Foy. 2005. Learning Perl, 4th ed. O'Reilly Media Inc. Press, CA. 304 pp.

Shim, K. S., S. K. Cho, J. U. Jeung, K. W. Jung, M. K. You, S. H. Ok, Y. S. Chung, K. H. Kang, H. G. Hwang, H. C. Choi, H. P. Moon, and J. S. Shin. 2004. Identification of fungal (*Magnaporthe grisa*) stress-induced genes in wild rice (*Oryza minuta*). Plant Cell Rep. 22:599–607.

Sreenivasuls, N., P. B. Kavi Kishor, R. K. Varshney, and L. Altschmied. 2002. Mining functional information from cereal genomes - the utility of expressed sequence tags. Curr. Sci. 83:965–973.

Trachtenberg, A. and D. Sklar. 2006. PHP Cookbook, 2nd ed. O'Reilly Media Inc. Press, CA. 810 pp.

Wei, M. L., C. T. Lu, and H. Y. Lu. 2005a. Research and development of expressed sequence tag (EST) in cereal crops. Sci. Agric. 53(1,2):15–21. (in Chinese)

Wei, M. L., H. Y. Lu, and C. T. Lu. 2005b. Bioinformatics collection and statistics of rice disease/pest resistance genomics research. Crop Environ. Bioinform. 2:295–306. (in Chinese with English abstract)

Williams, H. E. and D. Lane. 2004. Web Database Applications with PHP and MySQL, 2nd ed. O'Reilly Media Inc. Press, CA. 816 pp.

Wolfsberg, T. G. and D. Landsman. 2001. Expressed sequence tags (ESTs). p.283–301. *in*: Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2nd ed. (Andreas D. B. and B. F. F. Ouellette, eds.) John Wiley & Sons Inc. Press, NY.

Zhou. M., C. F. Tong, and J. S. Shi. 2007. Realization of perl-based automatic sequence homogeneous analysis. Biochnology 17:60–63. (in Chinese with English abstract)

Rice Expressed Sequence Tag (EST) Automated Functional Analysis System[1]

Hsiu-Ying Lu[2], Cheng-Tao Chen[2], Chun-Tang Lu[2,4], Yi-Chia Chiu[3], and Meng-Li Wei[2]

Abstract

Lu, H. Y., C. T. Chen, C. T. Lu, Y. C. Chiu, and M. L. Wei. 2008. Rice expressed sequence tag (EST) automated functional analysis system. J. Taiwan Agric. Res. 57:103–118.

Expressed sequence tag (EST) is the fragment of cDNA sequence. Direct study of EST is therefore helpful to obtain the gene expression information. The analytical procedures of EST functional annotation, classification and representation statistics are used to detect the differences in gene expression patterns among libraries and identify the functional categories involved. A vast number of rice (*Oryza sativa* L.) EST sequences in the public databases provide an important resource for functional annotation of rice genome. However, it's impossible to accomplish the entire process of functional analysis for large-scale EST sequences within a short time using manual intervention only. To accelerate the functional analysis for rice EST sequences, an automated system was constructed by using Perl and PHP scripts based on a local MySQL database of the Linux platform. A series of analysis including functional annotation, classification and representation statistics for EST sequences could be done automatically, which provides a rapid, accurate and efficient tool for large-scale rice EST functional analysis. The system could be used directly through a web browser on Microsoft Window operating system without logging into the Linux platform. The web-based input interface of automated system is very easy and convenient for use. After finishing the work, an email message with output results was created and sent to users. The automated system could also be easily expanded and implemented in the EST functional analysis of other species, which facilitates the EST analysis in molecular genetics laboratories.

Key words: Rice (*Oryza sativa* L.), Expressed sequence tag, Functional annotation and classification, Gene representation statistics, Automated system, Perl script.

---

1. Contribution No.2318 from Agricultural Research Institute, Council of Agriculture. Accepted: June 17, 2008.
2. Respectively, Senior Researcher, Project Assistant, Assistant Researcher and Assistant Researcher, Crop Science Division, ARI, Wufeng, Taichung, Taiwan, ROC.
3. Master, Graduate Institute of Biotechnology, Chaoyang University of Technology, Wufeng, Taichung, Taiwan, ROC.
4. Corresponding author, e-mail: tang@wufeng.tari.gov.tw; Fax: (04)23390528.