

基因連續性分析應用於基因表現之資料探勘

施宗牧^a 莊麗月^b 谷德倫^c 張慧朗^d 張學偉^c 楊正宏^a

^a 高雄應用科技大學 電子工程系

^d 長庚大學長庚紀念醫院生物資訊中心

^c 高雄醫學大學 生物醫學暨環境生物學系

^b 義守大學 化學工程系

E-mail: gmsunqmu@gmail.com

摘要

基因連續性分析(Serial Analysis of Gene Expression, SAGE)是一種利用分子生物學方法從 cDNA 中獲得一連串短序列(tag)來分析基因表現量的技術。透過此種技術，生物學家們可以廣泛地分析各種現有的生物資料庫樣本的基因表現資料。有鑑於現有的 SAGE 分析平台提供的分析結果，只能提供單組比較而無法作跨組比較，無法完善地滿足需求。因此，我們根據 NCBI 上的 SAGEmap 所提供之資料庫，在本次研究中提出一個簡易且便利的分析工具軟體。透過本工具軟體，我們提供一個獨特地多群組分析方式，並且呈現一個明顯易懂的視覺化圖表，作為生物學家分析各種基因表現之分析平台。

關鍵字：Gene expression quantifying techniques, SAGE, tag, cDNA, restriction enzyme

Abstract

SAGE(Serial Analysis of Gene Expression) is a technique analyzing a series of tags from the cDNA for its gene expression. Application of such technology, biologists can analyze the relative gene expression of many samples in silico. To date, some visible platforms only offer the simple comparison without cross-libraries evaluation. Therefore, we propose a simple and convenient tool for these complex analyses among different kinds of libraries. This platform retrieves the databases from NCBI SAGEmap and provides the multi-groups comparison of gene expression relatively in the output of visualization platform. Therefore, this platform provides a useful tool for biologists' concern on gene expression in various SAGE libraries.

Keywords: Gene expression quantifying techniques,

SAGE, tag, cDNA, restriction enzyme

1、前言

基因表現資料的量化技術可以在正常與不正常的細胞型態中塑造出對轉錄(transcription)結果的分配與調控上的認知。基因連續性分析(Serial Analysis of Gene Expression, SAGE)是由 Dr. Victor Velculescu 等人所提出的一種基因表現資料量化技術[1]。它是從 cDNA 經由特定限制酶(restriction enzyme)辨認後獲得一連串短序列(tag)並計算這些短序列重複數目，分析其基因表現的技術[8]。原則上，透過此技術我們即可得知一個短序列所代表的一個基因[2]。藉由此技術，生物學家們可以廣泛地分析各種生物資料庫樣本的基因表現資料，觀察比較出正常組織樣本與腫瘤組織樣本中，哪些基因有較高或者較低的表現，進而得知特定短序列所代表之特定基因。但是，由於這類型的技術會產生龐大的實驗數據(包含雜訊)，因此如何去過濾篩選與安排整理出有用的資訊，也就是找出代表某個基因之獨一無二的短序列，便成為一個很重要的問題。

據我們目前所知，SAGE 分析平台有 SAGEmap [3]與 ACTG [4]等。SAGEmap 是由 Lash AE, Tolstoshev CM 等人所提出的一個線上 SAGE 操作平台。SAGEmap 提供了多個物種的大量現有可用的 SAGE 資料可供下載，並且提供生物樣本的比較分析功能，此分析功能是利用 SAGE 技術所產生之短序列與已知基因和序列作比較、定位與篩選以建立 SAGE 資料庫，獲得短序列在各種組織樣本與細胞表現的相對表現量資訊；然而，此功能僅限定於兩個群組間之比較分析，並且呈現冗長且無順序性的列表清單結果等諸多不便的缺點。ACTG 是由 Pedro A. F. Galante, Jeff Trimarchi 等人所提出的一個線上操作系統，此系統主要提供之功

能是藉由使用者輸入一組大量的短序列進行查詢，以獲得短序列所對應(多種基因資料庫)的基因資訊結果；ACTG 雖然可以便利地檢索短序列所對應之基因，但是依然無法做多個組織樣本間的比較分析。有鑑於此，我們提出一個能夠大量比較多個樣本的一個基因表現資料探勘系統。透過本系統獨特的多群組分析方式與明顯易懂的視覺化結果呈現，生物學家可以更輕易地觀察出短序列在多個組織樣本間之表現情形。以期發現表現量相對地高或低的 tag 資訊，更進一步地找出可能的治病或致病基因。

2、研究方法

本研究提出一個在個人電腦上的 SAGE 資訊分析平台，提供使用者自由地挑選多組有興趣的樣本(library)以進行分析。分析的結果中，包含了各種 tag 在不同樣本或由多種樣本所組成之群組(group)中的表現量。為了使生物學家更便於分析，我們以圖形量表之形式來呈現這些分析結果，並且提供各種 tag 所對應的基因名稱資訊。詳細的系統架構、資料庫架構與群組分析說明如下。

2.1、系統架構

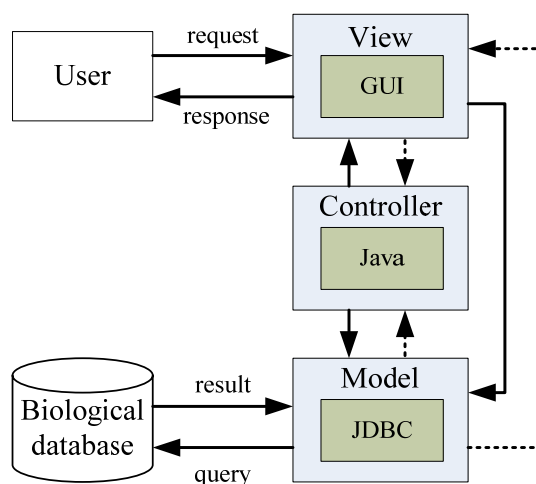


Figure 1. 本系統之 MVC 架構圖

本系統以 JAVA 程式語言為基礎，採用 MVC (Model-View-Controller)之架構將資料、程式邏輯與呈現外觀分開設計[7]。MVC 整體架構如 Figure 1 所示，我們設計(1)一個 GUI (Graphical user interface)來呈現使用者可操作之介面與結果資訊，(2)以 Java 為核心來處理使用者的分析需求與程式流程導向控制以及(3)

JDBC (Java Database Connectivity)作為底層程式邏輯存取生物資料庫之橋樑。

如 Figure 2 所示，本系統主要可分成 Input model、Output model、Group analysis model、Query model 與資料庫等五部份。詳述如下：

1.輸入模組(Input model)

在使用本系統進行群組分析(Group analysis)前，使用者需先設定欲分析之每個群組的相關參數，參數內容包含選擇要分析之 organism、library 與 tag 的篩選標準 tpm (tag per million)等部分。tag 資訊檢索部份，使用者則需於設定欲查詢之 tag、organism 與 anchor 後，即可得知 tag 的相關資訊。

2.群組分析模組(Group analysis model)

群組分析模組部分主要包含：(1)群組參數儲存單元(Parameter setting)，(2)群組儲存池(Group pool)，(3)與單一或多個群組分析單元(Single group analyzer, Multi group analyzer)。當完成群組參數設定後，參數儲存單元將儲存群組之參數設定值。接著，群組儲存池會根據各個群組之參數值作為篩選 tag 之標準來載入群組資訊。最後，挑選單一或多個群組以進行交叉比較分析。

3.檢索模組(Query model)

根據使用者之需求設定，至 SAGE database 進行檢索。

4.輸出模組(Output model)

本系統輸出之結果可分為(1)群組分析的圖表結果與(2) tag 檢索結果資訊。群組分析的圖表，包含了各種短序列在多種樣本之間的表現量或在多個群組間的表現量比較圖表資訊。tag 檢索，則會列出此 tag 在何種物種之樣本出現，表現量為何等資訊。

5.資料庫(Database)

(1)SAGE database:從 SAGEmap 中取得之 SAGE data；包含 tag、organism、library name、tag count、total tag count in the library、anchor 等資訊。

(2)UniGene mapping database:從 SAGEmap 中取得 tag 與 UniGene 的對應資料；包含 SAGE tag、organism、UniGene cluster number、UniGene cluster name 與 Tag-to-gene mapping reliability score 等資訊。

2.2、資料庫架構

本系統採用之資料庫主要以 SAGEmap 平台所提供之資料為主。我們主要採用了兩種資料：(1)SAGEmap 所提供的 SAGE 資料，其中涵括 *Homo sapiens*、*Mus musculus*、*Rattus norvegicus*、*Arabidopsis thaliana*、*Medicago truncatula*、*Bos taurus*、*Meleagris gallopavo*、*Pinus taeda*、*Zea mays*、*Caenorhabditis*

elegans、*Sus scrofa*、*Lentinula edodes*、*Gallus gallus*、*Bombyx mori*、*Magnaporthe grisea* 70-15、*Oryza sativa* (japonica cultivar-group)、*Drosophila melanogaster* 與 *Musa acuminata* 等多個物種的 SAGE 資料[5]。(2) SAGEmap 從 UniGene 上獲得的具有可信度的各種物種間之 tag 對照基因資料[6]。我們根據上述兩種資料採用 MySQL 建置資料庫。

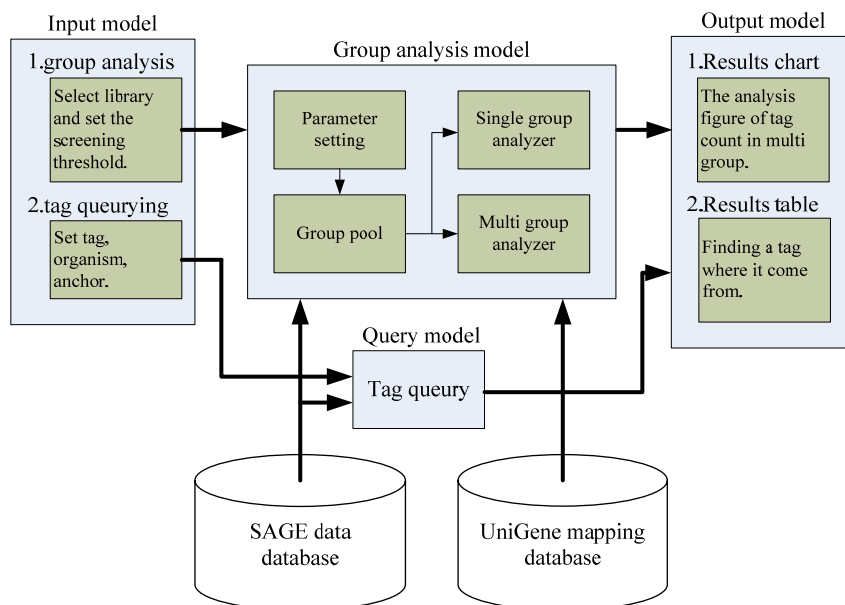


Figure 2. 系統架構圖

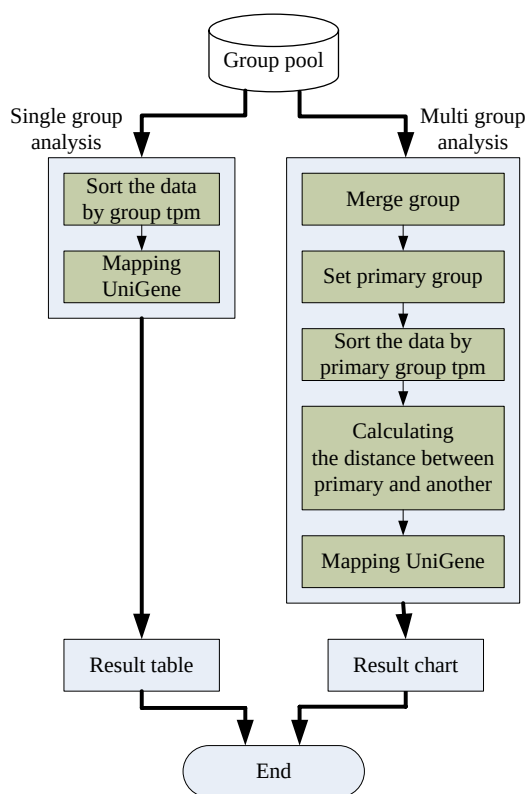


Figure 3. 群組分析流程圖

2.3、群組分析

我們所挑選的每一個群組，其主要包括多個 SAGE 的短序列，每個短序列於群組中之表現量(group tpm)，以及短序列從何種物種擷取與擷取之 anchor 為何等多種資訊。本系統所提供之群組分析功能可分為二種，分別為單一群組分析與多群組分析。兩種功能如 Figure 3 所示，其分析步驟如下：

1. 單一群組分析(Single group analysis)

- (1)根據群組中每個短序列表現量對所有資料排序。
- (2)建立群組中的每個短序列所對應至 UniGene 資訊(短序列所對應基因)。

2. 多群組分析(Multi group analysis)

- (1)將所有群組之資訊內容合併於一個短序列表單中。
- (2)設定中心群組。
- (3)以中心群組的每個短序列表現量作為排序依據，對整個短序列表單進行排序。

(4)計算中心群組與其他群組的短序列表現量距離。

(5)建立短序列表單中的每個短序列所對應至 UniGene 資訊(短序列所對應基因)。

3、結果與討論

本研究所發展之 SAGE 應用於基因表現資料之探勘系統介面如 Figure 4 所示，主要分為群組參數設定部份(group setting)與群組分析部份(group analysis)。在群組設定部份，使用者可以選擇系統所提供的 *Homo sapiens*、*Mus musculus*、*Rattus norvegicus*、*Arabidopsis thaliana*、*Medicago truncatula*、*Bos taurus*、*Meleagris gallopavo*、*Pinus taeda*、*Zea mays*、*Caenorhabditis elegans*、*Sus scrofa*、*Lentinula edodes*、*Gallus gallus*、*Bombyx mori*、*Magnaporthe grisea* 70-15、*Oryza sativa* (japonica cultivar-group)、*Drosophila melanogaster* 與 *Musa acuminata* 等 18 個物種的 SAGE 資料，接著設定一個群組包含哪些樣本(library，註解如右邊 Information 欄位所示)與篩選標準(tag per million)，最後設定自訂的群組名稱(group name)。在群組分析部份，使用者可以選擇一個或多個群組進行分析。我們在此以智慧人種(*Homo sapiens*)的前列腺組織(prostate tissue)的低表現量短序列之分析為例；首先，在群組參數設定部份選擇物種：*Homo sapiens*，載入該物種所包含之所有可選擇的生物樣本；接著，設定

正常的前列腺組織樣本-群組 1：prostate_normal，包含樣本 SAGE_Chen_Normal_Pr，tpm：1~100；有腫瘤的前列腺組織-群組 2：prostate_tumor，包含樣本 SAGE_Chen_Tumor_Pr，tpm：1~100。

在單一群組的分析中，群組 1 之分析結果畫面與 tag 分布情況各別如 Figure 5、Figure 6 所示；群組 2 之分析結果畫面與 tag 分布情況各別如 Figure 7、Figure 8 所示。其中群組 1 包含 19516 個不重複之 tag，群組 2 包含 21489 個不重複之 tag。群組中每個 tag 所對應之基因資訊，可在 Figure 5、Figure 7 其下面表格欄位(UniGene information)獲得。

在多群組之分析中，總共包含 34440 個不重複之 tag，分析結果畫面與 tag 分布情況各別如 Figure 9、Figure 10 與 Figure 11 所示。如 Figure 9 中每個 tag 所對應之基因資訊，可在其下面表格欄位(UniGene information)獲得。藉由 Figure 10 的結果呈現，生物學家可以群組 1 (prostate_normal)為中心(或者以群組 2 為中心，Figure 11)，與群組 2 之所有 tag 的表現量做一比較(如 Figure 11 中編號 5000 的 tag，在群組 1 的表現量較群組 2 高 17)；藉此發現單一獨特之 tag 在不同群組間之表現量為何，獲得表現量相對地高或低的 tag，更進一步地找出可能的治病或致病基因(如 Figure 11 中，表現量為 30、90 與 78 之 tag)。本系統與 SAGEmap 之比較結果則如 Table 1 所示。

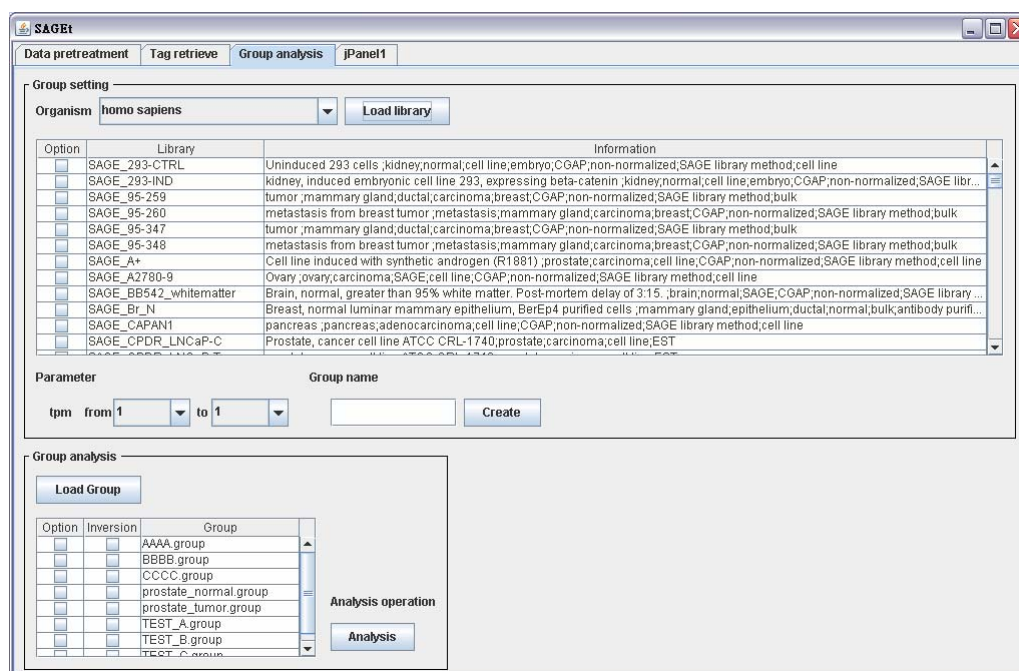


Figure 4. 系統介面

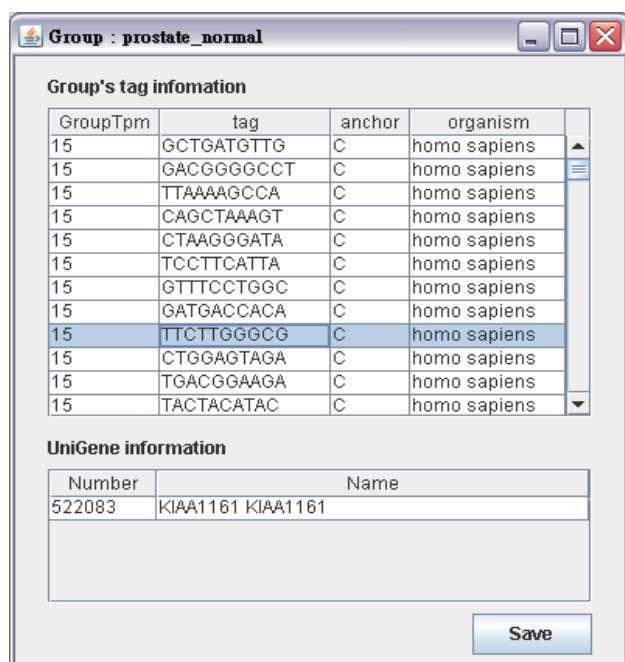


Figure 5. 群組 1 之分析結果畫面

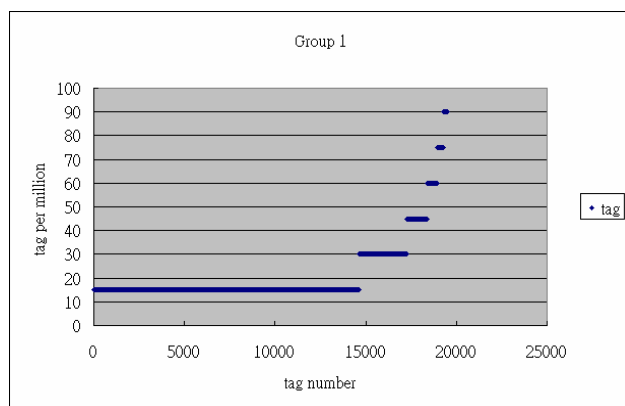


Figure 6. 群組 1 以分布圖表示之短序列分析結果

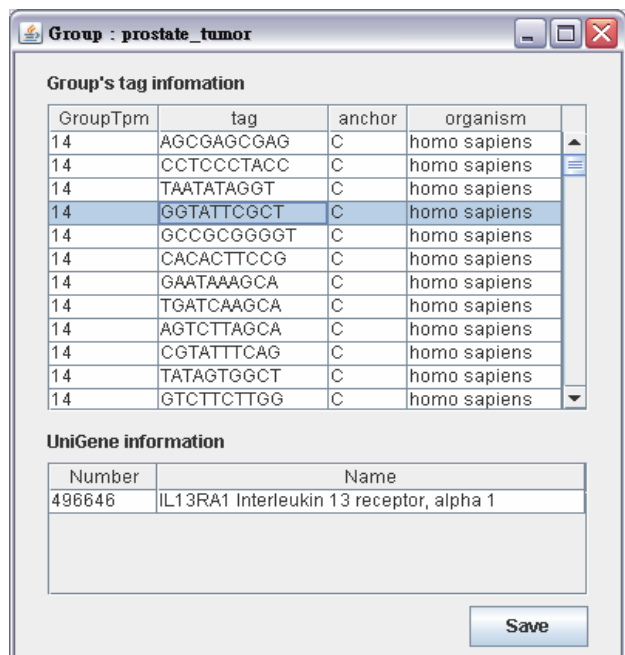


Figure 7. 群組 2(prostate tumor)之分析結果畫面

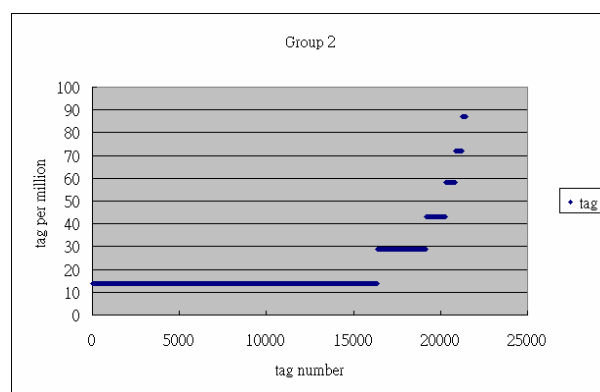


Figure 8. 群組 2 以分布圖表示之短序列分析結果

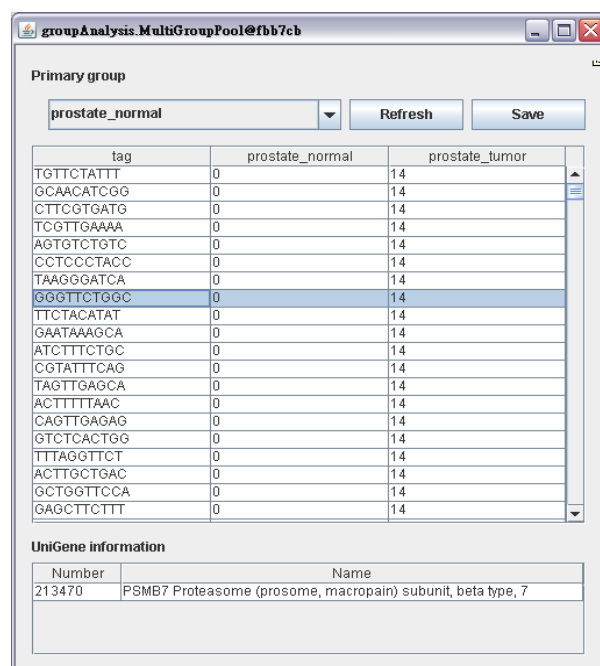


Figure 9. 多群組(群組 1 與群組 2)之分析結果畫面

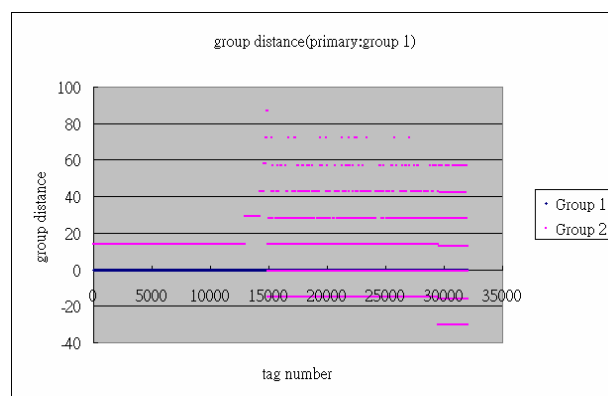


Figure 10. 多群組分析結果以分布圖表示之短序列分布圖(以群組 1 為中心)

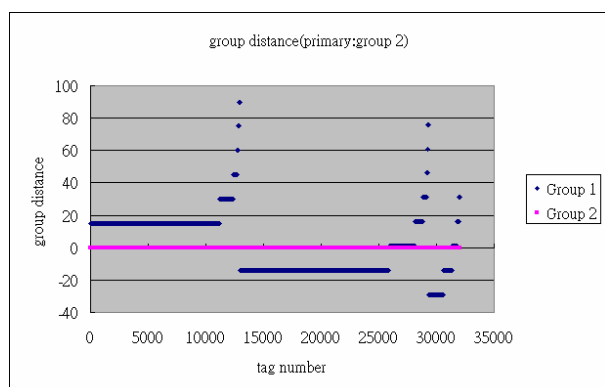


Figure 11. 多群組分析結果以分布圖表示之短序列分布圖(以群組 2 為中心)

Table 1. 本系統與 SAGEmap 比較表

	SAGEmap	本系統
分析型式	兩個群組進行分析。	一或多個群組進行分析。
呈現結果	不易觀察且冗長的 tag 比較串列表。	顯示 tag 表現量介於多個群組間之分布情形。

4、結論

在本研究中，我們提出了一個簡單且便利的 SAGE 分析工具。透過本文所提到之群組分析功能與一個簡單明瞭的結果圖表之呈現，生物學家可更輕易且便利地分析各種短序列在多種樣本之間的表現量，獲得單一獨特之 tag 在不同樣本間之表現量，發現表現量相對地高或低的 tag 資訊。如透過前列腺組織樣本的分析操作實例，我們可以輕易發現表現量差異頗大的三個部份(如 Figure 11，表現量分別為 30、90 與 78)，找出可能的治病或致病基因，成為生物學家分析各種基因在各種生體組織上的表現之利器。

參考文獻

- [1] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler, "Serial Analysis of Gene Expression", Science, 270:pp. 484-487, Oct. 1995.
- [2] C.H. Song and M. Wyse, "Painless Gene Expression Profiling : SAGE (Serial Analysis of Gene Expression)", The Science Creative Quarterly, Aug. 2004.
- [3] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D.

Schuler, R.L. Strausberg and G.J. Riggins, S.F. Altschul, "SAGEmap : a public gene expression resource", Genome Research, Vol. 10, Issue 7, pp. 1051-1060, Jul. 2000.

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

- [4] Pedro A. F. Galante, Jeff Trimarchi, Constance L. Cepko, Sandro J. de Souza, Lucila Ohno-Machado and Winston P. Kuo, "Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data", Bioinformatics, Applications note, Vol. 23 no. 7, pp. 903-905, Feb. 3, 2007.
<http://retina.med.harvard.edu/ACTG/>
- [5] SAGE data from SAGEmap in NCBI.
<ftp://ftp.ncbi.nlm.nih.gov/pub/sage/extr/>
- [6] Tag mapping data from SAGEmap in NCBI.
<ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/>
- [7] "Model-view-controller", Wikipedia.
<http://en.wikipedia.org/wiki/Model-view-controller>
- [8] "Serial Analysis of Gene Expression", Wikipedia.
http://en.wikipedia.org/wiki/Serial_Analysis_of_Gene_Expression