I636372

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I636372 B**

(45)公告日： 中華民國 107 (2018) 年 09 月 21 日

(21)申請案號：107100449 (22)申請日： 中華民國 107 (2018) 年 01 月 05 日

(51)Int. Cl.： *G06F19/00 (2018.01)* *G06F19/20 (2011.01)*
G06F19/22 (2011.01) *G06F9/30 (2018.01)*

(71)申請人：國立交通大學 (中華民國) NATIONAL CHIAO TUNG UNIVERSITY (TW)
新竹市大學路 1001 號
國立臺灣大學 (中華民國) NATIONAL TAIWAN UNIVERSITY (TW)
臺北市大安區羅斯福路四段一號

(72)發明人：洪瑞鴻 HUNG, JUI-HUNG (TW) ；楊家驤 YANG, CHIA-HSIANG (TW) ；吳易忠 WU, YI-CHUNG (TW)

(74)代理人：高玉駿；楊祺雄

(56)參考文獻：

TW	I451285	TW	I512517
US	2013/0031536A1	US	2013/0159666A1

審查人員：陳泰龍

申請專利範圍項數：21 項 圖式數：10 共 51 頁

(54)名稱

用於基因定序資料的資料處理方法及系統

DATA PROCESSING METHOD AND SYSTEM FOR GENE SEQUENCING DATA

(57)摘要

在一種資料處理方法及系統中，對於擷取自一參考 DNA 序列的後綴字串各自的前 K 個字符所得到的字串依序進行編碼、先升後降地的取樣、基於取樣結果的分群及排序等操作，以獲得一對應於該參考 DNA 字串的後綴字串陣列。一對應於該參考 DNA 字串的 FM-指標資料結構係根據該後綴字串陣列及其對應的指標而建立出，且包含一紀錄有該後綴字串陣列的第一欄字符的 F 表、一相關於該 F 表的 CNT 表、一紀錄有該後綴字串陣列的對應指標的 SA 表、一紀錄有該後綴字串陣列的最後欄字符的 L 表、及一相關於該 L 表的 OCC 表、並被搜尋可獲得一目標字串的搜尋結果。

In a data processing method and system, for a plurality of strings, each having first K characters extracted from a respective one of all suffix strings of a reference DNA sequence, encoding、up-sampling、down-sampling, and grouping are conducted in order so as to obtain a suffix string array corresponding to the reference DNA strings. An FM-index data structure corresponding to the reference DNA string is established based on the suffix string array, and on indexes related to the suffix string array. The FM-index data structure includes an F table recorded with characters in a first column of the suffix string array, a CNT table associated with the F table, an SA table recorded with the indexes related to the suffix string array, an L table recorded with characters in a last column of the suffix string array, and an OCC table associated with the L table. The FM-index data structure can be searched to obtain a search result for a target string.

指定代表圖：

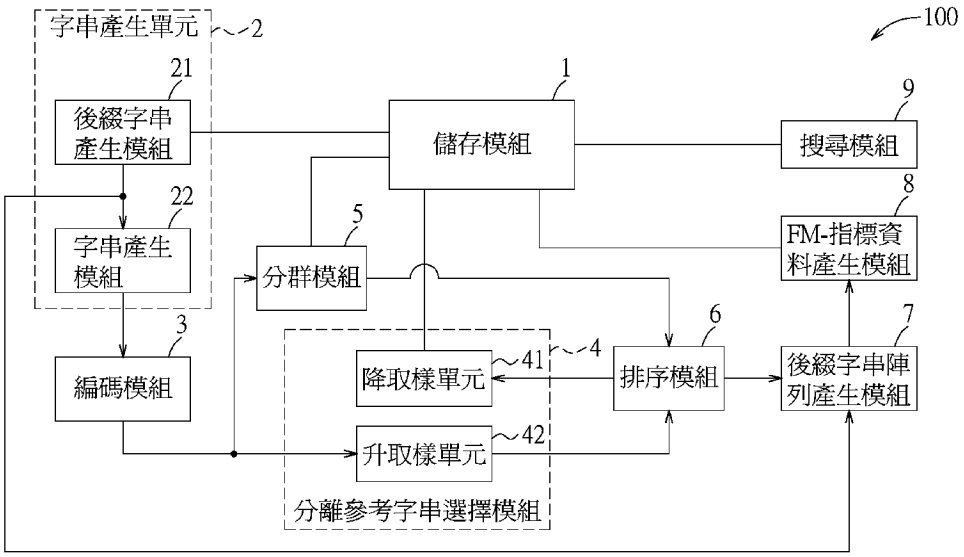

圖1

符號簡單說明：

100・・・資料處理系統

1・・・儲存模組

2・・・字串產生單元

21・・・後綴字串產生模組

22・・・字串產生模組

3・・・編碼模組

4・・・分離參考字串選擇模組

41・・・升取樣單元

42・・・降取樣單元

5・・・分群模組

6・・・排序模組

7・・・後綴字串陣列產生模組

8・・・FM-指標資料產生模組

9・・・搜尋模組

【發明說明書】

【中文發明名稱】　用於基因定序資料的資料處理方法及系統

【英文發明名稱】Data Processing Method and System for Gene Sequencing Data

【技術領域】

【0001】本發明是有關於一種資料處理方法及系統，特別是指一種用於基因定序資料的資料處理方法及系統。

【先前技術】

【0002】DNA定序就是確定出DNA分子中含氮鹼基的精確順序的過程。目前，DNA定序及DNA資料分析已成為確定出遺傳疾病的確切原因以及發展相關治療的不可或缺的工具，並且定序已成為遺傳學研究與生物醫藥應用的分析程序。因現實需求的驅使所發展出的平行定序，亦稱為次代定序(Next-Generation Sequencing，NGS)是目前最快的定序技術，且其花費約數小時來定序一整個人類DNA，並能以一大量平行的方式來定序多個短片段，以便達到相較於基於桑格(Sanger)定序的第一代DNA定序技術更高處理量的等級大小。NGS的應用範圍是廣大的且仍在擴大中，且此技術促進了許多相關於生物醫藥科學領域的快速發展。因為DNA定序的成長量已呈指數級增長，後續的資料處理及分析將極為耗時。

【0003】已發展出用於NGS資料分析的演算法，諸如揭露於Journal of Molecular Biology. Vol. 215, no. 3, pp. 403-410, May 1990的”Basic Local Alignment Search Tool”(以下簡稱BLAST)，以及揭露於PLoS ONE, vol. 4, no. 11, p.e7767, 2009的”An Alignment Tool for Large Scale Genome Resequencing”(以下簡稱BFAST)，仍耗費大量的處理時間而且並不適用於短片段基因比對。為解決此問題，在NGS資料分析的軟體封包使用了含有Burrow-Wheeler Transform (BWT)及其附屬資料結構的FM(Ferragina及Manzini)-index，以搜尋任意的短DNA序列。然而，雖然FM-index能以相對較快的比對速度進行基因比對，但在比對過程中，仍需佔用非常大的記憶體儲存空間。

【0004】因此，對於NGS資料分析及處理，如何兼顧快速DNA比對以提高比對量，以及降低所需的記憶體儲存空間遂成為一重要課題。

【發明內容】

【0005】因此，本發明的目的，即在提供一種用於基因定序資料的資料處理方法及系統，其能克服習知技藝的缺點。

【0006】於是，所提供的本發明資料處理方法係適用於處理基因定序資料並藉由一資料處理系統來實施。該基因定序資料包含相關於一具有連續的N個字符的參考DNA序列的N個後綴字串、及N個

分別指示出於該等N個字符在該參考DNA序列中的對應位置且分別指派該等N個後綴字串的指標。該等N個字符其中的前(N-1)個字符係由至少四個分別代表四種不同含氮鹼基的字符A，C，G，T所組成且其中的最後一個字符為一代表序列結束的字符$。該資料處理方法包含以下步驟：

【0007】(A)對於該等N後綴字串其中的每一者，擷取該後綴字串的前K個字符，以獲得N個分別對應於該等N個後綴字串的字串，其中N>K；

【0008】(B)依照一將該等字符，$，A，C，G，T分別以五個彼此不同且具有遞增數值的數字碼來表示的編碼方式，編碼該等N個字串，以產生N個具有一數字碼形式且分別對應於該等N個指標的編碼字串；

【0009】(C)以一先升後降的取樣方式，自該等N個編碼字串選出P個依照編碼值從小到大排列且分別作為第一至第P分離參考字串的字串；

【0010】(D)利用一根據該第一至第P分離參考字串所建立的分群規則，將該等N個編碼字串劃分成第一至第(P+1)群，其中該第r分離參考字串係歸屬於該第(r+1)群的編碼字串，且1≤r≤P；

【0011】(E)將每一群的編碼字串依照編碼值從小到大排序，以獲得該等N個編碼字串依照編碼值從小到大的排序結果；

【0012】(F)根據該排序結果，將該等N個後綴字串排序，以獲得一對應於該參考DNA字串的後綴字串陣列；

【0013】(G)根據該後綴字串陣列及該等指標，建立一對應於該參考DNA字串的FM-指標資料結構，該FM-指標資料結構包含一CNT表、一SA表、一F表、一L表及一OCC表，該F表係依序紀錄有該後綴字串陣列的該第一字符欄中的N個第一字符，該L表係依序紀錄有該後綴字串陣列的一最後字符欄的N個最後字符，該CNT表係依序紀錄有該表F中出現該等字符A，C，G，T各自的起始列位址之前一列位址，該SA表係依序紀錄有該後綴字串陣列中第一至第N個後綴字串所對應的指標，該OCC表紀錄有在對應於該表L的每一列位址，該等N個最後字符中已出現該等字符A，C，G，T其中每一者的累計次數；及

【0014】(H)根據該FM-指標資料結構、及一包含至少兩個選自該等字符A，C，G，T的字符的目標字串，利用一相關於後進搜尋方式的預定指標演算法，搜尋該FM-指標資料結構中的資料，以獲得一對應於該目標字串的搜尋結果，該搜尋結果指示出該參考DNA序列不存在該目標字串或者包含至少一個指示出該參考DNA序列中存在有該目標字串的所有位置的目標指標。

【0015】於是，所提供的本發明資料處理系統係適用於處理基因定序資料，該基因定序資料包含相關於一具有連續的N個字符的參

考DNA序列的N個後綴字串、及多個分別指示出該等N個字符在該參考DNA序列中的對應位置且分別指派給該等N個後綴字串的指標，該等N個字符其中的前(N-1)個字符係由四個分別代表四種不同含氮鹼基的字符A，C，G，T所組成且其中的最後一個字符為一代表序列結束的字符$。該資料處理系統包含一字串產生模組、一編碼模組、一分離參考字串選擇模組、一分群模組、一排序模組、一後綴字串陣列產生模組、一FM-指標資料產生模組、及一搜尋模組。

【0016】該字串產生模組，對於該等N後綴字串其中的每一者，擷取該後綴字串的前K個字符，以產生N個分別對應於該等N個後綴字串的字串，其中N>K。

【0017】該編碼模組電連接該字串產生模組以接收該等N個字串，並依照一將該等字符$，A，C，G，T，分別以五個彼此不同且具有遞增數值的數字碼來表示的編碼方式，編碼該等N個字串，以產生N個具有一數字碼形式且分別對應於該等N個指標的編碼字串。

【0018】該分離參考字串選擇模組電連接該編碼模組以接收該等N個編碼字串，並以一先升後降的取樣方式，自該等N個編碼字串選出P個依照編碼值從小到大排列且分別作為第一至第P分離參考字串的字串。

【0019】該分群模組電連接該編碼模組以接收該等N個編碼字串，並利用一根據該第一至第P分離參考字串所建立的分群規則，將該等N個編碼字串劃分成第一至第(P+1)群，其中該第r分離參考字串係歸屬於該第(r+1)群的編碼字串，且1≤r≤P。

【0020】該排序模組電連接該分群模組以接收該第一至第(P+1)群的編碼字串，並將每一群的編碼字串依照編碼值從小到大排序，以獲得該等N個編碼字串依照編碼值從小到大的排序結果。

【0021】該後綴字串陣列產生模組電連接該排序模組及該後綴字串產生模組，並根據來自於該排序模組的該排序結果，將來自於該後綴字串產生模組的該等N個後綴字串排序，以獲得一對應於該參考DNA字串的後綴字串陣列。

【0022】該FM-指標資料產生模組電連接該後綴字串陣列產生模組以接收該後綴字串陣列，並根據該後綴字串陣列產生模組所產生的該後綴字串陣列、及該等指標，建立一對應於該參考DNA字串的FM-指標資料結構。該FM-指標資料結構包含一CNT表、一SA表、一F表、一L表及一OCC表。該F表係依序紀錄有該後綴字串陣列的該第一字符欄中的N個第一字符，該L表係依序紀錄有該後綴字串陣列的一最後字符欄的N個最後字符，該CNT表係依序紀錄有該表F中出現該等字符A，C，G，T各自的起始列位址之前一列位址，該SA表係依序紀錄有該後綴字串陣列中第一至第N個後綴字串

所對應的指標，並且該OCC表紀錄有在對應於該表L的每一列位址，該等N個最後字符中已出現該等字符A，C，G，T其中每一者的累計次數。

【0023】該搜尋模組係根據該FM-指標資料結構、及一包含至少兩個選自該等字符A，C，G，T的字符的目標字串，利用一相關於後進搜尋方式的指標演算法，搜尋該FM-指標資料結構中的資料，以獲得一對應於該目標字串的搜尋結果。該搜尋結果指示出該參考DNA序列不存在該目標字串，或者包含至少一個指示出該參考DNA序列中存在有該目標字串的所有位置的目標指標。

【0024】本發明功效在於：由於使用了擷取自該等後綴字串的前K個字符而產生的該等字串來進行後續的編碼、分群及排序操作，因此可大量降低在建立該FM-指標資料結構期間所需的記憶體使用量。此外，由於使用了先升後降的取樣方式來選出該第一至第P分離參考字串，如此可有效且相對均勻地分群該等編碼字串，因而有效降低排序的複雜度。因此，當本發明係應用來處理人類DNA序列時，能在相對較低硬體成本及相對較少的執行時間下有效達成目標字串的搜尋。

【圖式簡單說明】

【0025】本發明的其他的特徵及功效，將於參照圖式的實施方式

中清楚地呈現，其中：

圖 1 是一方塊圖，說明本發明資料處理系統的一實施例；

圖 2 示例地說明該實施例根據一參考 DNA 序列所產生的後綴字串及其所對應的指標；

圖 3 示例地說明該實施例根據圖 2 的後綴字串所產生的字串及其所對應的指標；

圖 4 是一電路方塊圖，說明該實施例的一排序模組的結構；

圖 5 是一電路方塊圖；說明該實施例的該排序模組的三個彼此串接的排序元件；

圖 6 示例地說明該實施例的一分群模組將圖 3 的該等字串分群後所獲得的一分群結果；

圖 7 示例地說明該實施例的該排序模組將圖 6 所示的分群結果排序後所獲得的一排序結果；；

圖 8 示例地說明該實施例的一後綴字串陣列產生模組所產生的一後綴字串陣列及其所對應的指標；

圖 9 示例地說明該實施例的一 FM-指標資料產生模組所產生一 FM-指標資料結構；及

圖 10 示例地說明該實施例中被該 FM-指標資料產生模組儲存於一儲存模組的一部份的該 FM-指標資料結構。

【實施方式】

【0026】在本發明被詳細描述的前，應當注意在以下的說明內容中，類似的元件是以相同的編號來表示。

【0027】參閱圖1，所繪示的本發明資料處理系統100的一實施例係適於應用來處理一參考DNA序列，例如人類DNA序列，但不在此限。在本實施例中，該參考DNA序列具有連續的N個字符，其中的前(N-1)個字符係由至少四個分別代表四種不同含氮鹼基的字符A，C，G，T(例如分別為腺嘌呤、胞啼啶、鳥嘌呤及胸腺啼啶)所組成，且其中的最後一個字符為一代表序列結束的字符$。值得注意的是，在實際使用時，該參考DNA序列的前(N-1)個字符中可含有一個或多個異於該等字符A，C，G，T的字符，此(等)字符代表尚未被確認的含氮鹼基。該資料處理系統100包含一儲存模組1、一電連接該儲存模組1的字串產生單元2、一電連接該儲存模組1及該字串產生單元2的編碼模組3、一電連接該編碼模組3及該儲存模組1的分離參考字串選擇模組4、一電連接該儲存模組1及該編碼模組3的分群模組5、一電連接該分離參考字串選擇模組4及該分群模組5的排序模組6、一電連接該排序模組6及該後綴字串產生模組21的後綴字串陣列產生模組7、一電連接該後綴字串陣列產生模組7及該儲存模組1的FM-指標資料產生模組8、及一電連接該儲存模組1的搜尋模組9。

【0028】該儲存模組1係儲存有該參考DNA序列、及N個分別指示出於該等N個字符在該參考DNA序列中的對應位置的指標。在本實施例中，例如以0至(N-1)作為該等N個分別指派給該等N個字符的指標，但不在此限。由於實際應用時作為該參考DNA序列的人體DNA序列含有約三十億個含氮鹼基，為方便說明，以下列舉一簡單例子來說明該參考DNA序列的該等N個含氮鹼基(以字符來表示)與該等N個指標的關係，其中N=8，且該等八個字符及該等八個指標如以下表1所示。

表1

指標	0	1	2	3	4	5	6	7
含氮鹼基	C	A	T	G	C	A	A	$

【0029】該字串產生單元2包含一電連接該儲存模組1的後綴字串產生模組21、及一電連接該後綴字串產生模組21與該編碼模組3的字串產生模組22。

【0030】該後綴字串產生模組21根據該儲存模組1所儲存的該參考DNA序列及該等指標，從該參考DNA序列的左側第一個字符開始，依序產生分別對應於該等N個字符的該等N個後綴字串，並將作為該等指標的0至(N-1)依序指派給該等N個後綴字串。舉例來說，當沿用表1的例子(即，該參考DNA序列例如為“CATGCAA$”)時，該後綴字串產生模組21所產生的該等後綴字串及其所對應的該

等指標係如圖2所示。在本實施例中，該等後綴字串及其所對應的該等指標共同構成相關於該參考DNA序列的基因定序資料。該後綴字串產生模組21還將該基因定序資料輸出至該字串產生模組22。

【0031】當該字串產生模組22接收到來自該後綴字串產生模組21的該基因定序資料時，該字串產生模組22，對於該等N後綴字串其中的每一者，擷取該後綴字串的前K個字符，以產生N個分別對應於該等N個後綴字串的字串，其中N>K。舉例來說，若利用圖2所示的例子且當K=4的情況下，該字串產生模組22所產生的該等八個字串及其所對應的指標係如圖3所示。值得注意的是，前例係為了方便說明才採用N=8及K=4。值得注意的是，在實際應用時，由於N≈3×109並且配合該儲存模組1的規格，例如K=16，故N係遠大於K，藉此可在後續處理期間大幅降低對於記憶體儲存容量的需求。該字串產生模組22將該等N個字串輸出至該編碼模組3。

【0032】該編碼模組3在接收到來自該該字串產生模組22的該等N個字串時，依照一將該等字符$，A，C，G，T，分別以五個彼此不同且具有遞增數值的數字碼來表示的編碼方式，編碼該等N個字串，以產生N個具有一數字碼形式且分別對應於該等N個指標的編碼字串。例如，該等字符$，A，C，G，T分別被編碼成000、001、010、011及100的數字碼，但不在此限。

【0033】該分離參考字串選擇模組4在接收到來自該編碼模組3的該等N個編碼字串時，以一先升後降的取樣方式，自該等N個編碼字串選出P個依照編碼值從小到大排列且分別作為第一至第P分離參考字串的字串，並將所產生的該第一至第P分離參考字串儲存於該儲存模組1。在本實施例中，該分離參考字串選擇模組4包含一電連接該編碼模組3及該排序模組6的升取樣單元41、及一電連接該排序模組6及該儲存模組1的降取樣單元42。更明確地，該升取樣單元41接收來自於該編碼模組3的該等N個編碼字串，且自該等N個編碼字串任意取出P×Q個字串，並將該等P×Q個字串輸出至該排序模組5。例如，該升取樣單元41係以每次任意取出P個字串的方式連續執行Q次而獲得該等P×Q個字串。於是，該排序模組6在接收到來自於該升取樣單元41的該等P×Q個字串時，還將該等P×Q個字串依照編碼值從小到大排序。然後，該降取樣單元43將來自於該排序模組5且已被排序的該等P×Q個字串，從其中具有最小編碼值的一字串開始，且以每Q個字串取出一個字串的方式，取出該第一至第P分離參考字串，並且將該第一至第P分離參考字串，以數字碼的形式，儲存於該儲存模組1。舉例來說，當沿用圖3所示的該等字串的情況時，例如指標分別為0及2的字串，即CATG及TGCA，所對應的編碼字串被選為該第一及第二分離參考字串。值得注意的是，由於使用了先升後降的取樣方式，於是可有效確保該分離參考

字串選擇模組4所選出的該第一至第P個分離參考字串分布更加均勻，藉此可降低在後續將要實施的分群及排序操作上的複雜度。

【0034】參閱圖4及圖5，在本實施例中，該排序模組6包含C個彼此串接的排序元件61、及一A×1多工器62，並且每一排序元件61包含一暫存器611、一比較器612及一2×1多工器(以下簡稱多工器)613。接著，就每一排序元件61的組成及其操作進一步詳細說明。該暫存器611(例如一D型正反器，但不在此限)具有一輸入端、一控制端及一輸出端，且在該輸出端暫時保留當前資料(例如以Qi來表示)。該比較器612具有一用於接收待排序資料(Qin)的第一輸入端、一耦接該暫存器61的該輸出端的第二輸入端、及一耦接該暫存器的該控制端的輸出端。該多工器613具有一用於接收該待排序資料Qin的第一輸入端、一耦接前一個排序元件61的該暫存器611的該輸出端的第二輸入端、一耦接該前一個排序元件61的該比較器612的該輸出端的控制端、及一耦接該暫存器的該輸入端的輸出端。在操作時，例如，就位於圖5中間的排序元件61而言，該比較器612在比較出該待排序資料Qin在數值上大於該暫存器611所暫存(保留)的該當前資料Qi時，經由其本身的該輸出端將一例如具有高邏輯準位的致能信號輸出至該暫存器611的該控制端以及後一個排序元件61(即位於圖5右側的排序元件61)的該多工器613的該控制端，以致該後一個排序元件61的該多工器613回應於該致能信號

而選擇將來自該暫存器611的該當前資料Qin輸出，同時該暫存器611回應於該致能信號將來自於該多工器613的該輸出端的資料暫時保留在其本身的該輸出端。應注意的是，該多工器613所輸出的該資料需視來自於該前一個排序元件61(即位於圖5左側的排序元件61)的該比較器612的該輸出端的一信號而定，更明確地說，當該信號為一致能信號時，該多工器613將來自於該前一個排序元件61的該暫存器611的該輸出端的當前資料Qi-1作為該資料而輸出，而當信號為一禁能信號時，該多工器613將該待排序資料Qin作為該資料而輸出。另一方面，該比較器612在比較出該待排序資料Qin在數值上小於該暫存器611所暫存的該當前資料Qi時，經由其本身的該輸出端將一例如具有低邏輯準位的禁能信號輸出至該暫存器611的該控制端以及該後一個排序元件61的該多工器613的該控制端，以致該後一個排序元件61的該多工器613回應於該禁能信號而選擇將該待排序資料Qin輸出，同時該暫存器611回應於該禁能信號維持不變(即，仍保留該當前資料Qi)。於是，在此配置下，具有越小數值的待排序資料容易被優先輸出而達到排序的目的。

【0035】該A×1多工器62具有A個輸入端D0~DA-1、一用於接收一相關於待排序資料的總筆數D的控制信號的控制端、及一輸出端，例如其中的第m個輸入端(即Dm-1)係耦接第(m×B)個排序元件61的該暫存器611的該輸出端，且m=1,2,..,A，但不在此限。該

A×1多工器62係根據該控制信號而可操作在下述的情況：當D≦B，建立該第一個輸入端D0與該輸出端的電連接；當n×B<D≦(n+1)×B，建立該第(n+1)個輸入端Dn與該輸出端的電連接，其中n=1,2,...，A-1；及當C<D<2C，建立該第A個輸入端DA-1與該輸出端電連接。特別注意的是，當C<D<2C時，該排序模組在執行完該等D筆待排序資料的(第一次)排序時，還須將最先輸出的(D-C)筆資料，也就是溢流(overflow)的資料再執行一次(第二次)排序操作。於是，藉由將第二次排操作所獲得的該等(D-C)筆資料的排序結果結合第一次排序操作所獲得的該等C筆資料的排序結果而獲得該等D筆待排序資料的排序結果。值得注意的是，由於使用了該A×1多工器62，特別是對於小於C之筆數的待排序資料可適應性將該排序模組6的輸出延遲(Output Latency)降至最低。

【0036】該分群模組5接收該編碼模組3所產生的該等N個編碼字串，並利用一根據該儲存模組1所儲存的該第一至第P分離參考字串所建立的分群規則，將該等N個編碼字串劃分成第一至第(P+1)群。在本實施例中，其中該第r分離參考字串係歸屬於該第(r+1)群的編碼字串，且r=1,2,...,P。對於該等N個編碼字串其中每一者，該分群規則包含：當該編碼字串的編碼值係小於該第一分離參考字串的編碼值時，該編碼字串係歸屬於該第一群；當該編碼字串的編碼值係不小於該第P分離參考字串的編碼值時，該編碼字串係歸屬

於該第(P+1)群；及當該編碼字串的編碼值係不小於該第(j-1)分離參考字串的編碼值但小於該第j分離參考字串的編碼值時，該編碼字串係歸屬於該第j群，其中j=2,3,...,P。舉例來說，在沿用圖3所示的該等字串並且指標分別為0及2的字串，即“CATG”及“TGCA”，所對應的編碼字串被選為該第一及第二分離參考字串的情況下，該等八個編碼短片字串被分成三群，也就是圖6所繪示出對應於該等三群的編碼字串的三群(字符形式)的字串，其中第一群與第二群均有對應的三個字串，而第三群只有兩個對應的字串，並且該第一分離參考字串(即“CATG”)係歸屬於第二群且該第二分離參考字串(即“TGCA”)係歸屬於第三群。

【0037】此外，值得注意的是，該分群模組5在執行每個編碼字串的分群操作時，係利用一二元搜尋方式來相對快速地決定出該編碼字串所應歸屬的一群。

【0038】該排序模組6接收來自該分群模組5的該第一至第(P+1)群的編碼字串，並將每一群的編碼字串依照編碼值從小到大排序，以獲得該等N個編碼字串依照編碼值從小到大的排序結果。舉例來說，在沿用圖6所之示分群結果的情況下，該排序模組6所獲得的排序結果係如圖7所示。值得注意的是，由於該排序模組6是以逐群的方式進行排序操作，因此可相對大幅降低該等N個編碼字串在排序上的複雜度。

【0039】該後綴字串陣列產生模組7接收來自於該排序模組6的(該等N個編碼字串的)該排序結果以及來自於該後綴字串產生模組21的該等N個後綴字串，並根據該排序結果，將該等N個後綴字串排序，以獲得一對應於該參考DNA字串的後綴字串陣列。舉例來說，在沿用圖2所示的該等後綴字串及圖7所示的排序果的情況下，該後綴字串陣列產生模組7所獲得的後綴字串陣列以及其所對應的該等指標係如圖8所示。

【0040】該FM-指標資料產生模組8接收來自於該後綴字串陣列產生模組7的該後綴字串陣列，並根據該後綴字串陣列、及該等指標，建立一對應於該參考DNA字串的FM-指標資料結構。在本實施例中，該FM-指標資料結構包含一CNT表、一SA表、一F表、一L表及一OCC表。該F表係依序紀錄有該後綴字串陣列的該第一字符欄中的N個第一字符，該L表係依序紀錄有該後綴字串陣列的一最後字符欄的N個最後字符，該CNT表係依序紀錄有該表F中出現該等字符A，C，G，T各自的起始列位址之前一列位址，該SA表係依序紀錄有該後綴字串陣列中第一至第N個後綴字串所對應的指標，並且該OCC表紀錄有在對應於該表L的每一列位址，該等N個最後字符中已出現該等字符A，C，G，T其中每一者的累計次數。舉例來說，在沿用圖8的情況下，該FM-指標資料產生模組8所建立的FM-指標資料結構係如圖9所示。

【0041】值得注意的是，選擇上，若該儲存模組1並無儲存容量的限制時，該FM-指標資料產生模組8可將該FM-指標資料結構完整地儲存於該儲存模組1。或者，為了降低該儲存模組1對於該FM-指標資料結構中的資料所需的儲存空間，較佳地，該FM-指標資料產生模組8可僅將一部份的該FM-指標資料結構儲存於該儲存模組1。由於該CNT表係根據該FM表所紀錄的內容而產生，且該OCC表係根據該L表所紀錄的內容而產生以及該SA表係與該OCC表相關聯，所以該部分的FM-指標資料結構可至少由該CNT表、該L表、一部分的該SA表、及一部分的該OCC表所構成。在本實施例中，例如，該FM-指標資料產生模組8係藉由自該SA表以每T1列(row)取其中的第一列的一第一下取樣方式來產生該部分的SA表，並且藉由自該OCC表以每T2列取其中的第一列的一第二取樣方式產生該部分的OCC表，但不在此限。舉例來說，在沿用圖9所示的FM-指標資料結構的情況下，當T1=T2=4時，該部分的FM-指標資料結構係如圖10所示。如此，在實際應用於人體DNA序列時，相較於習知技藝以儲存整個FM-指標資料結構的方式，可大幅降低用於儲存對應的FM-指標資料結構的必要資料所需的儲存空間。

【0042】該搜尋模組9係配置來判斷該參考DNA序列是否存在有一目標字串，並在判斷出該參考DNA序列存在有一個或多個該目

標字串時，決定出該(等)目標字串於該參考DNA序列中的位置，其中該目標字串包含至少兩個選自該等字符A，C，G，T的字符。應注意的是，在執行該目標字串的判斷操作之前，該搜尋模組9須獲得完整的該FM-指標資料結構。在本實施例中，若該儲存模組1已完整儲存有該FM-指標資料結構時，則該搜尋模組9可相對省時地直接開始進行該判斷操作。然而，若該儲存模組1僅儲存有該部分的FM-指標資料結構時，則該搜尋模組9必須將該部分的SA表及該部分的OCC表重建回完整的該SA表及該OCC表，且重新獲得該F表。更明確地說，該搜尋模組9可簡單地根據該儲存模組1所儲存的該CNT表而重新獲得該F表。此外，該搜尋模組9根據該儲存模組1所儲存的該部分的該SA表及該部分的OCC表，且利用一FM-指標資料重建演算法，獲得完整的該SA表及該OCC表，藉此獲得完整的該FM-指標資料結構。在本實施例中，該FM-指標資料重建演算法被表示成以下兩式：

【0043】 $$OCC[n.s]=OCC_D[\left\lfloor \frac{n}{T2},s \right\rfloor]+L[\left\langle \left\lfloor \frac{n}{T2} \right\rfloor+1,n \right\rangle,s] \quad (式1)$$

【0044】SA[n]=SAD[CNT[L[n]]+OCC[n,L[n]]]+1 (式2)

【0045】其中，n代表列位址，s代表字符，OCCD代表該部分的OCC表，L代表該L表，OCC代表該OCC表，CNT代表該CNT表，SAD代表該部分的SA表，以及SA代表該SA表。如此，該搜尋模

組9可根據該部分的OCC表且利用式1、該L表及T2重建出完整的該OCC表，並且可根據該部分的SA表及已重建的該OCC表且利用式2重建出完整的該SA表。

【0046】然後，該搜尋模組9根據自該儲存模組1所讀取出或經由本身重建出的該FM-指標資料結構、及該目標字串，利用一相關於後進搜尋方式的指標演算法，搜尋該FM-指標資料結構中的資料，以獲得一對應於該目標字串的搜尋結果。值得注意的是，若該參考DNA序列實際上不存在有該目標字串時，該搜尋結果將指示出該參考DNA序列不存在該目標字串，或者，若該參考DNA序列實際上存在有該目標字串時，該搜尋結果將包含至少一個指示出該參考DNA序列中存在有該目標字串的所有位置的目標指標。以下將示例地詳細說明該搜尋模組9如何利用該指標演算法來獲得對應於該目標字串的搜尋結果。

【0047】在本實施例中，當該目標字串被表示為“S1S2..SM”且1≤M<N時，該預定指標演算法被表示成

$$S[i] = S_{(M-i)+1},\ i = 1,2,....,M \qquad (式3)$$

$$index_{\min}[i] = CNT[S[i]] + OCC[index_{\min}[i-1]-1, S[i]] + 1 \qquad (式4)$$

$$index_{\max}[i] = CNT[S[i]] + OCC[index_{\max}[i-1], S[i]] \qquad (式5)$$

【0048】其中S[i]代表在第i次迭代搜尋運算中所欲搜尋的目標字符，及indexmin[i]及indexmax[i]分別代表在第i次迭代搜尋運算中可能存在有該目標字符所對應的最小指標及最大指標，並且其

$$index_{\min}[0]=0$$

初始值分別被定義為 $index_{\max}[0]=N-1$。在此情況下，該搜尋模組9所執行的該搜尋操作包含以下步驟。

【0049】首先，該搜尋模組9利用式3而獲得S[1]=SM(即，第1次迭代搜尋運算的目標字符)，且利用式3及式4並查找該CNT表及該OCC表來執行第1次迭代搜尋運算，以獲得indexmin[1]及indexmax[1]。值得注意的是，在第1次迭代搜尋運算中，由於該OCC表僅紀錄有列位址0至N-1的資料，因此OCC[-1,SM]被預設為0。接著，該搜尋模組9判定indexmin[1]是否大於indexmax[1]。若判定結果為肯定時，該搜尋模組9判定出該參考DNA序列不存在任何該目標字串並同時產生指示出該參考DNA序列不存在該目標字串的該搜尋結果。相反地，若該搜尋模組9判定出該indexmin[1]不大於indexmax[1]時，該搜尋模組9令i=1+1=2，也就是說，並將繼續執行第2次迭代搜尋運算。

【0050】同樣地，在第2次迭代搜尋運算中，該搜尋模組9利用式3而獲得S[2]=SM-1(即，第2次迭代搜尋運算的目標字符)，並利用式4及式5且查找該CNT表及該OCC表，以獲得indexmin[2]及indexmax[2]，並在判定出indexmin[2]不大於indexmax[2]時，令i=2+1。否則，該搜尋模組9同樣地產生指示出該參考DNA序列不存在該目標字串的該搜尋結果。

【0051】之後，該搜尋模組9重複執行相似於上述步驟且依照上述判定邏輯，在前(M-1)次迭代搜尋運算後均未發生indexmin不大於indexmax的情況下，直到執行完第M次迭代搜尋運算後而獲得indexmin[M]及indexmax[M]，並接著判定indexmin[M]與indexmax[M]在大小關係。更明確地，當該搜尋模組9判定出indexmin[M]大於indexmax[M]時，同樣地，該搜尋模組9產生指示出該參考DNA序列不存在該目標字串的該搜尋結果。當該搜尋模組9判定出indexmin[M]等於indexmax[M]時，該搜尋模組9根據indexmin[M]或indexmax[M]查找該SA表，以產生僅包含有一個指示出該參考DNA序列中存在有該目標字串的唯一位置的目標指標的該搜尋結果，其中SA[indexmin[M]]或SA[indexmax[M]]代表該SA表中在列位址indexmin[M]或indexmax[M]所紀錄的指標並作為該目標指標。當判定出indexmin[M]小於indexmax[M]時，根據indexmin[M]至indexmax[M]查找該SA表，以產生包含有R個指示出該參考DNA序列中存在有該目標字串的R個不同位置的目標指標的該搜尋結果，其中R=indexmax[M]-indexmin[M]+1，且SA[indexmin[M]]至SA[indexmax[M]]代表該SA表中從列位址indexmin[M]至列位址indexmax[M]所紀錄的R個指標並作為該等R個目標指標。

【0052】以下，舉一簡單例子來詳細說明有關該搜尋模組9實際上

如何搜尋該FM-指標資料結構來獲得搜尋結果的細節。在此例中，圖9所示的FM-指標資料結構被使用，該目標字串為“CA”，M=2，indexmin[0]=0且indexmax[0]=7。

【0053】首先，由於是以後進搜群方式，在第一次迭代搜尋操作中，

【0054】S[1]=A，

【0055】indexmin[1]=CNT[A]+OCC[indexmin[0]-1,A]+1=0+0+1=1，

【0056】indexmax[1]=CNT[A]+OCC[indexmax[0],A]=0+3=3。

【0057】由於indexmax[1]大於indexmin[1]，繼續執行第二次迭代搜尋操作。

【0058】然後，在第二次迭代搜尋操作中，

【0059】S[2]=C，

【0060】indexmin[2]=CNT[C]+OCC[indexmin[1]-1,C]+1

【0061】=3+OCC[0,C]+1=3+0+1=4，

【0062】indexmax[2]=CNT[C]+OCC[indexmax[1],C]

【0063】=3+OCC[3,C]=3+2=5。

【0064】於是，R=(indexmax[2]-indexmin[2])+1=2，且SA[indexmin[2]]至SA[indexmax[2]]作為兩個目標指標，即，SA[4]=4，SA[5]=0。最後，該搜尋結果指示出包含分別為4與0的兩個目標指標。換言之，從圖2所示的該參考DNA序列

(“CATGCAA$”)可看出，確實在對應於指標0及4(目標指標)的位置存在有該目標字串。

【0065】綜上所述，由於本發明資料處理系統100使用了擷取自該等後綴字串的前K個字符所產生的該等字串來進行後續的編碼、分群及排序操作，特別是在N遠大於K的情況下，因此可大幅降低在建立該FM-指標資料結構期間所需的記憶體使用量。此外，由於使用了先升後降的取樣方式來選出該第一至第P分離參考字串，如此可有效且相對均勻地分群該等編碼字串，因而有效降低排序的複雜度。因此，當本發明係應用來處理人類DNA序列時，能在相對較低硬體成本及相對較少的執行時間下有效達成目標字串的搜尋，故確實能達成本發明之功效。

【0066】惟以上所述者，僅為本發明的實施例而已，當不能以此限定本發明實施的範圍，凡是依本發明申請專利範圍及專利說明書內容所作的簡單的等效變化與修飾，皆仍屬本發明專利涵蓋的範圍內。

【符號說明】

【0067】

100……資料處理系統

1……儲存模組

2……字串產生單元

61……排序元件

611……暫存器

612……比較器

21……後綴字串產生模組

22……字串產生模組

3……編碼模組

4……分離參考字串選擇模組

41……升取樣單元

42……降取樣單元

5……分群模組

6……排序模組

613……2×1 多工器

62……A×1 多工器

7……後綴字串陣列產生模組

8……FM-指標資料產生模組

9……搜尋模組

Q_i,Q_{i-1},Q_{i+1}……當前資料

Qin……待排序資料

$D_0 \sim D_{A-1}$ 輸入端

專利案號：107100449

公告本

I636372

【發明摘要】

【中文發明名稱】 用於基因定序資料的資料處理方法及系統

【英文發明名稱】Data Processing Method and System for Gene Sequencing Data

【中文】

在一種資料處理方法及系統中，對於擷取自一參考DNA序列的後綴字串各自的前K個字符所得到的字串依序進行編碼、先升後降地的取樣、基於取樣結果的分群及排序等操作，以獲得一對應於該參考DNA字串的後綴字串陣列。一對應於該參考DNA字串的FM-指標資料結構係根據該後綴字串陣列及其對應的指標而建立出，且包含一紀錄有該後綴字串陣列的第一欄字符的F表、一相關於該F表的CNT表、一紀錄有該後綴字串陣列的對應指標的SA表、一紀錄有該後綴字串陣列的最後欄字符的L表、及一相關於該L表的OCC表、並被搜尋可獲得一目標字串的搜尋結果。

【英文】

In a data processing method and system, for a plurality of strings, each having first K characters extracted from a respective one of all suffix strings of a reference DNA sequence, encoding、up-sampling、down-sampling, and grouping are conducted in order so as to obtain a suffix string array corresponding to the reference DNA strings. An FM-index data structure corresponding to the reference DNA string is established based on the suffix string array, and on indexes related to the suffix string array. The FM-index data structure includes an F table recorded with characters in a first column of the suffix string array, a CNT table associated with the F table, an SA table recorded with the indexes related to the suffix string array, an L table recorded with characters in

a last column of the suffix string array, and an OCC table associated with the L table. The FM-index data structure can be searched to obtain a search result for a target string.

【指定代表圖】：圖（1）。

【代表圖之符號簡單說明】

100……資料處理系統

1……儲存模組

2……字串產生單元

21……後綴字串產生模組

22……字串產生模組

3……編碼模組

4……分離參考字串選擇模組

41……升取樣單元

42……降取樣單元

5……分群模組

6……排序模組

7……後綴字串陣列產生模組

8……FM-指標資料產生模組

9……搜尋模組

【發明申請專利範圍】

【第1項】 一種資料處理方法，適於處理基因定序資料且藉由一資料處理系統來實施，該基因定序資料包含相關於一具有連續的N個字符的參考DNA序列的N個後綴字串、及N個分別指示出於該等N個字符在該參考DNA序列中的對應位置且分別指派給該等N個後綴字串的指標，該等N個字符其中的前(N-1)個字符係由至少四個分別代表四種不同含氮鹼基的字符A，C，G，T所組成且其中的最後一個字符為一代表序列結束的字符$，該資料處理方法包含以下步驟：

(A)對於該等N後綴字串其中的每一者，擷取該後綴字串的前K個字符，以獲得N個分別對應於該等N個後綴字串的字串，其中N>K；

(B)依照一將該等字符，$，A，C，G，T分別以五個彼此不同且具有遞增數值的數字碼來表示的編碼方式，編碼該等N個字串，以產生N個具有一數字碼形式且分別對應於該等N個指標的編碼字串；

(C)以一先升後降的取樣方式，自該等N個編碼字串選出P個依照編碼值從小到大排列且分別作為第一至第P分離參考字串的字串；

(D)利用一根據該第一至第P分離參考字串所建立的分群規則，將該等N個編碼字串劃分成第一至第(P+1)群，其中該第r分離參考字串係歸屬於該第(r+1)群的編碼字串，且r=1,2,...,P；

(E)將每一群的編碼字串依照編碼值從小到大排序，

以獲得該等N個編碼字串依照編碼值從小到大的排序結果；

(F)根據該排序結果，將該等N個後綴字串排序，以獲得一對應於該參考DNA字串的後綴字串陣列；

(G)根據該後綴字串陣列及該等指標，建立一對應於該參考DNA字串的FM-指標資料結構，該FM-指標資料結構包含一CNT表、一SA表、一F表、一L表及一OCC表，該F表係依序紀錄有該後綴字串陣列的該第一字符欄中的N個第一字符，該L表係依序紀錄有該後綴字串陣列的一最後字符欄的N個最後字符，該CNT表係依序紀錄有該表F中出現該等字符A，C，G，T各自的起始列位址之前一列位址，該SA表係依序紀錄有該後綴字串陣列中第一至第N個後綴字串所對應的指標，該OCC表紀錄有在對應於該表L的每一列位址，該等N個最後字符中已出現該等字符A，C，G，T其中每一者的累計次數；及

(H)根據該FM-指標資料結構、及一包含至少兩個選自該等字符A，C，G，T的字符的目標字串，利用一相關於後進搜尋方式的指標演算法，搜尋該FM-指標資料結構中的資料，以獲得一對應於該目標字串的搜尋結果，該搜尋結果指示出該參考DNA序列不存在該目標字串或者包含至少一個指示出該參考DNA序列中存在有該目標字串的所有位置的目標指標。

【第2項】 如請求項1所述的資料處理方法，在步驟(A)之前，還包含以下步驟：

(I)根據該參考DNA序列及該等指標，從該參考DNA序列的左側第一個字符開始，依序產生分別對應於該等N個字符的該等N個後綴字串，並將作為該等指標的0至(N-1)依序指派給該等N個後綴字串。

【第3項】 如請求項1所述的資料處理方法，其中，步驟(C)包含以下子步驟：

(C1)自該等N個編碼字串任意取出P×Q個字串；

(C2)將該等P×Q個字串依照編碼值從小到大排序；

(C3)自排序的該等P×Q個字串，從具有最小編碼值的字串開始以每Q個字串取出一個字串的方式，取出該第一至第P分離參考字串。

【第4項】 如請求項3所述的資料處理方法，其中，在子步驟(C1)中，該資料處理系統係以每次任意取出P個字串的方式連續執行Q次而獲得該等P×Q個字串。

【第5項】 如請求項1所述的資料處理方法，其中，在步驟(D)中，對於該等N個編碼字串其中每一者，該分群規則包含：

當該編碼字串的編碼值係小於該第一分離參考字串的編碼值時，該編碼字串係歸屬於該第一群；

當該編碼字串的編碼值係不小於該第P分離參考字串的編碼值時，該編碼字串係歸屬於該第(P+1)群；及

當該編碼字串的編碼值係不小於該第(j-1)分離參考字串的編碼值但小於該第j分離參考字串的編碼值時，該編碼字串係歸屬於該第j群，其中$j=2,3,\ldots,P$。

【第6項】 如請求項5所述的資料處理方法，其中，在步驟(D)中，該

資料處理系統在執行該等N個編碼字串其中每一者的分群操作時，係利用一二元搜尋方式來決定出該編碼字串所應歸屬的一群。

【第7項】 如請求項1所述的資料處理方法，其中，在步驟(H)中，當該目標字串被表示為“$S_1S_2..S_M$”且$1 \le M < N$時，該預定指標演算法被表示成

$$S[i] = S_{(M-i)+1},\ i = 1,2,....,M$$

$$index_{\min}[i] = CNT[S[i]] + OCC[index_{\min}[i-1]-1, S[i]]+1$$

$$index_{\max}[i] = CNT[S[i]] + OCC[index_{\max}[i-1], S[i]]$$

，其中，S[i]代表在第i次迭代搜尋運算中所欲搜尋的目標字符，以及$index_{min}[i]$及$index_{max}[i]$分別代表在第i次迭代搜尋運算中可能存在有該目標字符所對應的最小指標及最大指標，並且其初始值分別被定義為 $index_{\min}[0] = 0$，$index_{\max}[0] = N-1$，並且步驟(H)還包含以下子步驟：

利用該預定指標演算法並查找該CNT表及該OCC表來執行第i次迭代搜尋運算，以獲得$index_{min}[i]$及$index_{max}[i]$；

當判定出$index_{min}[i]$不大於$index_{max}[i]$時，令i=i+1，重複執行子步驟(H1)，直到獲得$index_{min}[M]$及$index_{max}[M]$；

當判定出$index_{min}[i]$大於$index_{max}[i]$或者$index_{min}[M]$大於$index_{max}[M]$時，產生指示出該參考DNA序列不存在該目標字串的該搜尋結果；

當判定出$index_{min}[M]$等於$index_{max}[M]$時，根據

$index_{min}[M]$或$index_{max}[M]$查找該SA表，以產生僅包含有一個指示出該參考DNA序列中存在有該目標字串的唯一位置的目標指標的該搜尋結果，其中$SA[index_{min}[M]]$或$SA[index_{max}[M]]$代表該SA表中在列位址$index_{min}[M]$或$index_{max}[M]$所紀錄的指標並作為該目標指標；及

當判定出$index_{min}[M]$小於$index_{max}[M]$時，根據$index_{min}[M]$至$index_{max}[M]$查找該SA表，以產生包含有R個指示出該參考DNA序列中存在有該目標字串的R個不同位置的目標指標的該搜尋結果，其中$R=index_{max}[M]-index_{min}[M]+1$，且$SA[index_{min}[M]]$至$SA[index_{max}[M]]$代表該SA表中從列位址$index_{min}[M]$至列位址$index_{max}[M]$所紀錄的R個指標並作為該等R個目標指標。

【第8項】一種資料處理系統，適於處理基因定序資料，該基因定序資料包含相關於一具有連續的N個字符的參考DNA序列的N個後綴字串、及多個分別指示出該等N個字符在該參考DNA序列中的對應位置且分別指派給該等N個後綴字串的指標，該等N個字符其中的前(N-1)個字符係由至少四個分別代表四種不同含氮鹼基的字符A，C，G，T所組成且其中的最後一個字符為一代表序列結束的字符$，該資料處理系統包含：

一字串產生模組，對於該等N後綴字串其中的每一者，擷取該後綴字串的前K個字符，以產生N個分別對應

於該等N個後綴字串的字串，其中N>K；

一編碼模組，電連接該字串產生模組以接收該等N個字串，並依照一將該等字符$，A，C，G，T，分別以五個彼此不同且具有遞增數值的數字碼來表示的編碼方式，編碼該等N個字串，以產生N個具有一數字碼形式且分別對應於該等N個指標的編碼字串；

一分離參考字串選擇模組，電連接該編碼模組以接收該等N個編碼字串，並以一先升後降的取樣方式，自該等N個編碼字串選出P個依照編碼值從小到大排列且分別作為第一至第P分離參考字串的字串；

一分群模組，電連接該編碼模組以接收該等N個編碼字串，並利用一根據該第一至第P分離參考字串所建立的分群規則，將該等N個編碼字串劃分成第一至第(P+1)群，其中該第r分離參考字串係歸屬於該第(r+1)群的編碼字串，且r=1,2,...,P；

一排序模組，電連接該分群模組以接收該第一至第(P+1)群的編碼字串，並將每一群的編碼字串依照編碼值從小到大排序，以獲得該等N個編碼字串依照編碼值從小到大的排序結果；

一後綴字串陣列產生模組，電連接該排序模組及該後綴字串產生模組，並根據來自於該排序模組的該排序結果，將來自於該後綴字串產生模組的該等N個後綴字串排序，以獲得一對應於該參考DNA字串的後綴字串陣列；

一FM-指標資料產生模組，電連接該後綴字串陣列產

生模組以接收該後綴字串陣列，並根據該後綴字串陣列及該等指標，建立一對應於該參考DNA字串的FM-指標資料結構，該FM-指標資料結構包含一CNT表、一SA表、一F表、一L表及一OCC表，該F表係依序紀錄有該後綴字串陣列的該第一字符欄中的N個第一字符，該L表係依序紀錄有該後綴字串陣列的一最後字符欄的N個最後字符，該CNT表係依序紀錄有該表F中出現該等字符A，C，G，T各自的起始列位址之前一列位址，該SA表係依序紀錄有該後綴字串陣列中第一至第N個後綴字串所對應的指標，該OCC表紀錄有在對應於該表L的每一列位址，該等N個最後字符中已出現該等字符A，C，G，T其中每一者的累計次數；及

一搜尋模組，根據該FM-指標資料結構、及一包含至少兩個選自該等字符A，C，G，T的字符的目標字串，利用一相關於後進搜尋方式的指標演算法，搜尋該FM-指標資料結構中的資料，以獲得一對應於該目標字串的搜尋結果，該搜尋結果指示出該參考DNA序列不存在該目標字串或者包含至少一個指示出該參考DNA序列中存在有該目標字串的所有位置的目標指標。

【第9項】 如請求項8所述的資料處理系統，還包含：

一儲存模組，電連接該分離參考字串選擇模組及該分群模組，且儲存有該參考DNA序列及該等指標；

其中，該分離參考字串選擇模組還將所產生的該第一至第P分離參考字串儲存於該儲存模組；及

其中，該分群模組根據該儲存模組所儲存的該第一至第P分離參考字串來建立該分群規則。

【第10項】如請求項9所述的資料處理系統，還包含：

一後綴字串產生模組，電連接該儲存模組及該字串產生模組，且根據該儲存模組所儲存的該參考DNA序列及該等指標，從該參考DNA序列的左側第一個字符開始，依序產生分別對應於該等N個字符的該等N個後綴字串，並將作為該等指標的0至(N-1)依序指派給該等N個後綴字串，該等後綴字串及其所對應的該等指標共同構成該基因定序資料，該後綴字串產生模組還將該基因定序資料輸出至該字串產生模組，以致該字串產生模組根據該等N個後綴字串來產生該等N個字串。

【第11項】如請求項9所述的資料處理系統，其中，該分離參考字串選擇模組還連接該排序模組，並包含：

一升取樣單元，電連接該編碼模組及該排序模組以接收來自於該編碼模組的該等N個編碼字串，且自該等N個編碼字串任意取出P×Q個字串，並將該等P×Q個字串輸出至該排序模組，以致該排序模組在接收到來自於該升取樣單元的該等P×Q個字串時，還將該等P×Q個字串依照編碼值從小到大排序；及

一降取樣單元，電連接該排序模組及該儲存模組，並將來自於該排序模組且已被排序的該等P×Q個字串，從其中具有最小編碼值的一字串開始，且以每Q個字串取出一個字串的方式，取出該第一至第P分離參考字串，並且將

該第一至第P分離參考字串儲存於該儲存模組。

【第12項】如請求項11所述的資料處理系統，其中，該升取樣單元係以每次任意取出P個字串的方式連續執行Q次而獲得該等P×Q個字串。

【第13項】如請求項9所述的資料處理系統，其中，對於該等N個編碼字串其中每一者，該分群規則包含：

當該編碼字串的編碼值係小於該第一分離參考字串的編碼值時，該編碼字串係歸屬於該第一群；

當該編碼字串的編碼值係不小於該第P分離參考字串的編碼值時，該編碼字串係歸屬於該第(P+1)群；及

當該編碼字串的編碼值係不小於該第(j-1)分離參考字串的編碼值但小於該第j分離參考字串的編碼值時，該編碼字串係歸屬於該第j群，其中j=2,3,...,P。

【第14項】如請求項13所述的資料處理系統，其中，該分群模組在執行該等N個編碼字串其中每一者的分群操作時係利用一二元搜尋方式來決定出該編碼字串所應歸屬的一群。

【第15項】如請求項9所述的資料處理系統，其中：

該FM-指標資料產生模組還電連接該儲存模組，並將該FM-指標資料結構完整地儲存於該儲存模組；及

該搜尋模組電連接該儲存模組以讀取該儲存模組所儲存的該FM-指標資料結構中的資料。

【第16項】如請求項9所述的資料處理系統，其中：

該FM-指標資料產生模組還電連接該儲存模組，並將一部分的該FM-指標資料結構儲存於該儲存模組，該部分

的FM-指標資料結構係由該CNT表、該L表、一部分的該SA表、及一部分的該OCC表所構成；及

該搜尋模組電連接該儲存模組，並在執行該搜尋操作前，根據該儲存模組所儲存的該部分的FM-指標資料結構且利用一FM-指標資料重建演算法，獲得完整的該FM-指標資料結構。

【第17項】如請求項16所述的資料處理系統，其中：

該FM-指標資料產生模組係藉由自該SA表以每T1行取其中的第一行的一第一下取樣方式來產生該部分的SA表，並且藉由自該OCC表以每T2行取其中的第一行的一第二取樣方式產生該部分的OCC表；

該FM-指標資料重建演算法被表示成以下兩式

$$OCC[n.s] = OCC_D[\left\lfloor \frac{n}{T2}, s \right\rfloor] + L[\left\langle \left\lfloor \frac{n}{T2} \right\rfloor + 1, n \right\rangle, s] \qquad (式1)$$

$$SA[n]=SA_D[CNT[L[n]]+OCC[n,L[n]]]+1 \qquad (式2)$$

其中，n代表列位址，s代表字符，OCC_D代表該部分的OCC表，L代表該L表，OCC代表該OCC表，SA_D代表該部分的SA表，以及SA代表該SA表；及

該搜尋模組根據該部分的OCC表且利用式1、該L表及T2重建出完整的該OCC表，並根據該部分的SA表及該OCC表且利用式2重建出完整的該SA表。

【第18項】如請求項9所述的資料處理系統，其中：

當該目標字串被表示為“$S_1S_2..S_M$”且1≤M<N時，該預定指標演算法被表示成

$S[i]=S_{(M-i)+1},\ i=1,2,....,M$

$index_{\min}[i]=CNT[S[i]]+OCC[index_{\min}[i-1]-1,S[i]]+1$，其中S[i]代表在

$index_{\max}[i]=CNT[S[i]]+OCC[index_{\max}[i-1],S[i]]$

第i次迭代搜尋運算中所欲搜尋的目標字符，以及$index_{min}[i]$及$index_{max}[i]$分別代表在第i次迭代搜尋運算中可能存在有該目標字符所對應的最小指標及最大指標，並且其初始值分別被定義為 $index_{\min}[0]=0$ $index_{\max}[0]=N-1$；及

該搜尋模組所執行的該搜尋操作包含

利用該預定指標演算法並查找該CNT表及該OCC表來執行第i次迭代搜尋運算，以獲得$index_{min}[i]$及$index_{max}[i]$，

當判定出$index_{min}[i]$不大於$index_{max}[i]$時，令i=i+1，重複執行子步驟(H1)，直到獲得$index_{min}[M]$及$index_{max}[M]$，

當判定出$index_{min}[i]$大於$index_{max}[i]$或者$index_{min}[M]$大於$index_{max}[M]$時，產生指示出該參考DNA序列不存在該目標字串的該搜尋結果，

當判定出$index_{min}[M]$等於$index_{max}[M]$時，根據$index_{min}[M]$或$index_{max}[M]$查找該SA表，以產生僅包含有一個指示出該參考DNA序列中存在有該目標字串的唯一位置的目標指標的該搜尋結果，其中$SA[index_{min}[M]]$或$SA[index_{max}[M]]$代表該SA表中在列位址$index_{min}[M]$或$index_{max}[M]$所紀錄的指標並作為該目標指標，及

當判定出$index_{min}[M]$小於$index_{max}[M]$時，根據$index_{min}[M]$至$index_{max}[M]$查找該SA表，以產生包含有R個指示出該參考DNA序列中存在有該目標字串的R個不同位置的目標指標的該搜尋結果，其中$R=index_{max}[M]-index_{min}[M]+1$，且$SA[index_{min}[M]]$至$SA[index_{max}[M]]$代表該SA表中從列位址$index_{min}[M]$至列位址$index_{max}[M]$所紀錄的R個指標並作為該等R個目標指標。

【第19項】如請求項8所述的資料處理系統，其中，該排序模組包含C個彼此串接的排序元件，每一排序元件包含：

一暫存器，具有一輸入端、一控制端及一輸出端，且在該輸出端暫時保留當前資料；

一比較器，具有一用於接收待排序資料的第一輸入端、一耦接該暫存器的該輸出端的第二輸入端、及一耦接該暫存器的該控制端的輸出端；及

一2×1多工器，具有一用於接收該待排序資料的第一輸入端、一耦接前一個排序元件的該暫存器的該輸出端的第二輸入端、一耦接該前一個排序元件的該比較器的該輸出端的控制端、及一耦接該暫存器的該輸入端的輸出端；

對於每一排序元件，該比較器在比較出該待排序資料在數值上大於該暫存器所暫存的該當前資料時，經由其本身的該輸出端將一致能信號輸出至該暫存器的該控制端以及後一個排序元件的該2×1多工器的該控制端，以致該後一個排序元件的該2×1多工器回應於該致能信號而選擇

將來自該暫存器的該當前資料輸出，同時該暫存器回應於該致能信號將來自於該2×1多工器的該輸出端的資料暫時保留在其本身的該輸出端，而該比較器在比較出該待排序資料在數值上小於該暫存器所暫存的該當前資料時，經由其本身的該輸出端將一禁能信號輸出至該暫存器的該控制端以及該後一個排序元件的該多工器的該控制端，以致該後一個排序元件的該2×1多工器回應於該禁能信號而選擇將該待排序資料輸出，同時該暫存器回應於該禁能信號維持不變。

【第20項】如請求項19所述的資料處理系統，其中，該排序模組還包含一A×1多工器，該A×1多工器具有A個輸入端、一用於接收一相關於待排序資料的總筆數D的控制信號的控制端、及一輸出端，其中的第m個輸入端係耦接第(m×B)個排序元件的該暫存器的該輸出端，且m=1,2,..,A，該A×1多工器係根據該控制信號而操作如下：

當D≦B，建立該第一個輸入端與該輸出端的電連接；

當n×B<D≦(n+1)×B，建立該第(n+1)個輸入端與該輸出端的電連接，其中n=1,2,...，A-1；及

當C<D<2C，建立該第A個輸入端與該輸出端電連接。

【第21項】如請求項20所述的資料處理系統，其中，當C<D<2C時，該排序模組在執行完該等D筆待排序資料的排序時，還將最先輸出的(D-C)筆資料再執行一次排序操作。

【發明圖式】

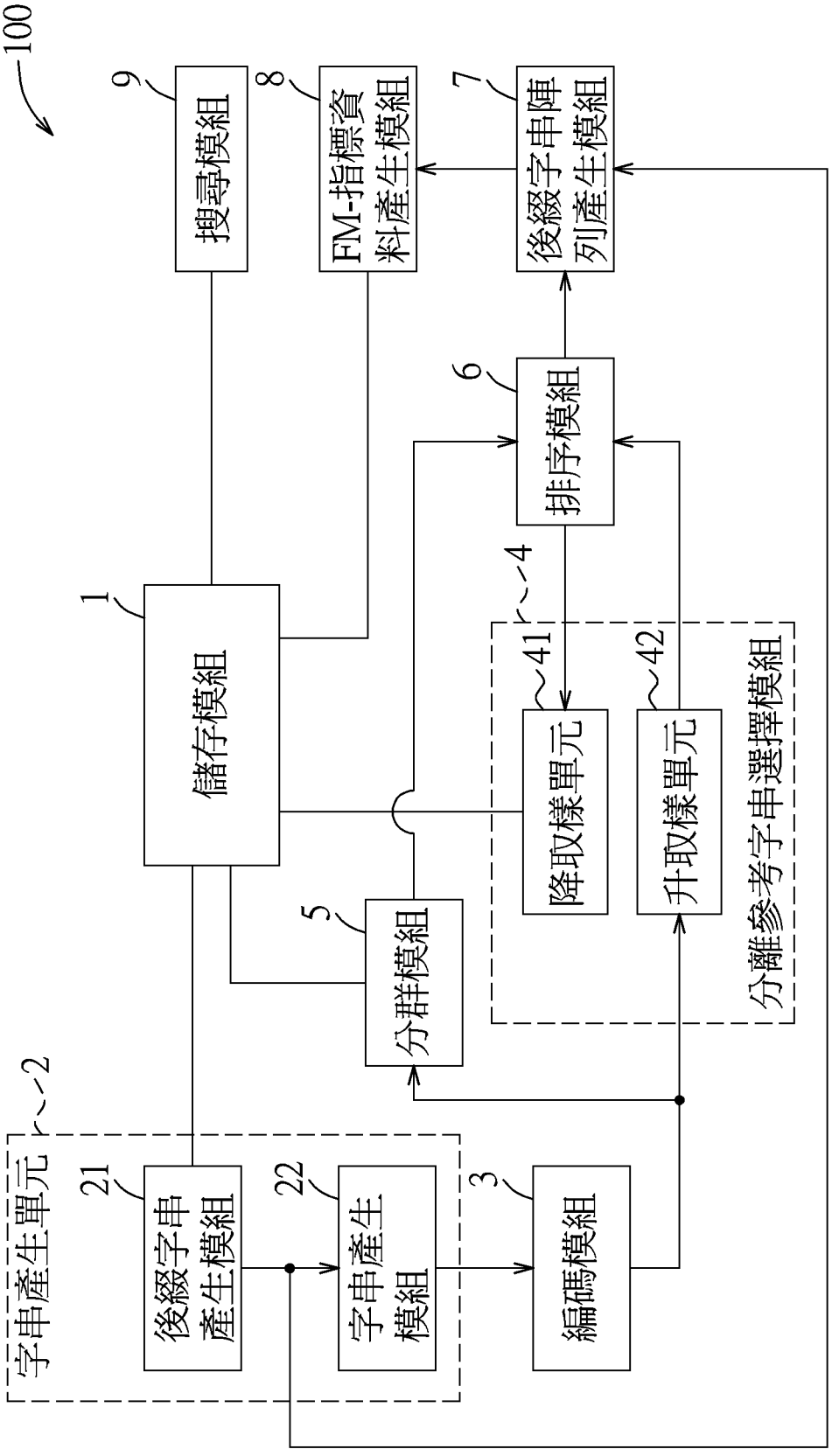
100
9
搜尋模組
8
FM-指標資料產生模組
7
後綴字串陣列產生模組
6
排序模組
1
儲存模組
4
41
降取樣單元
42
升取樣單元
分離參考字串選擇模組
5
分群模組
2
字串產生單元
21
後綴字串產生模組
22
字串產生模組
3
編碼模組

圖 1

參考DNA序列

C	A	T	G	C	A	A	$

指標	後綴字串							
0	C	A	T	G	C	A	A	$
1	A	T	G	C	A	A	$	C
2	T	G	C	A	A	$	C	A
3	G	C	A	A	$	C	A	T
4	C	A	A	$	C	A	T	G
5	A	A	$	C	A	T	G	C
6	A	$	C	A	T	G	C	A
7	$	C	A	T	G	C	A	A

圖2

指標	字串			
0	C	A	T	G
1	A	T	G	C
2	T	G	C	A
3	G	C	A	A
4	C	A	A	$
5	A	A	$	C
6	A	$	C	A
7	$	C	A	T

圖3

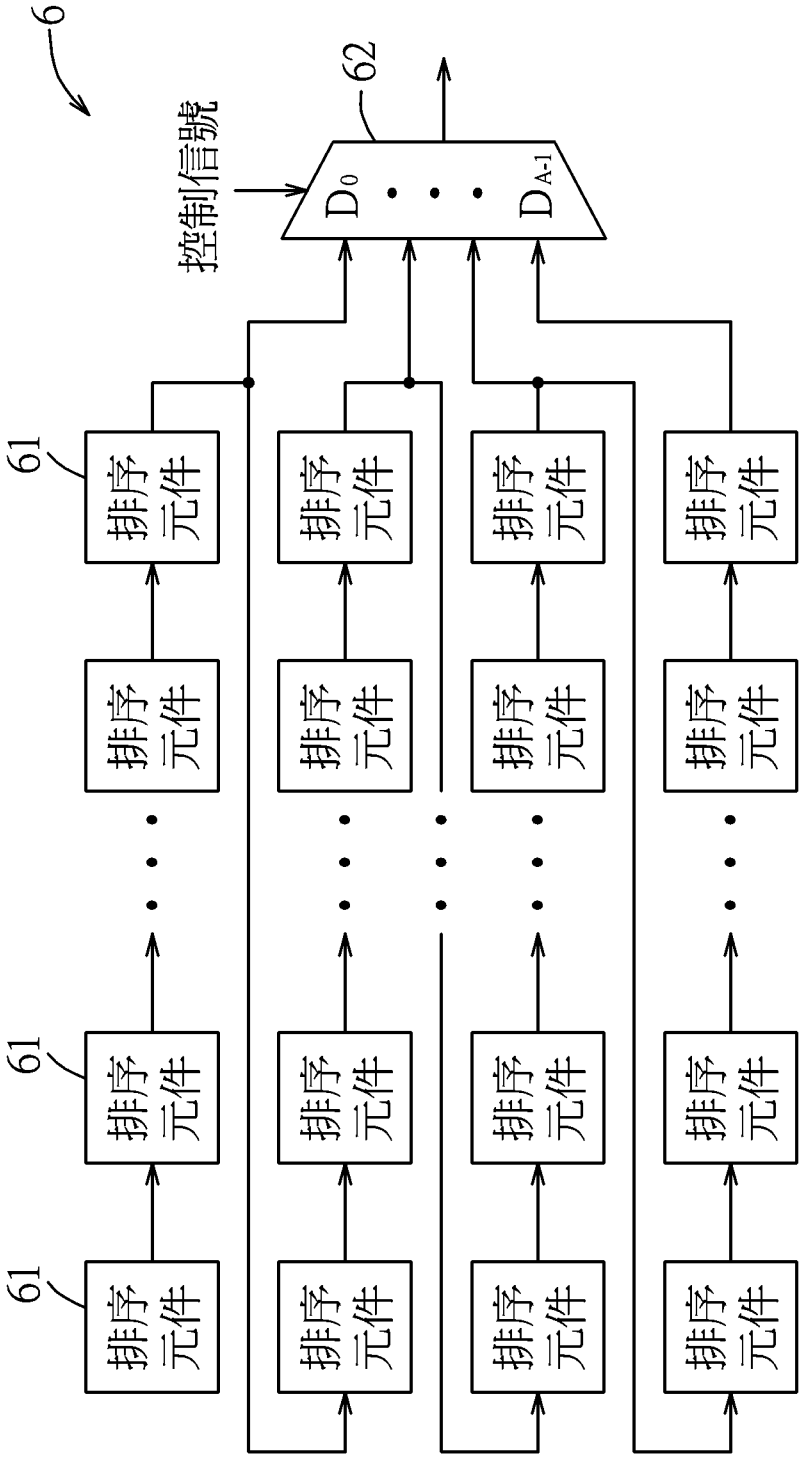

圖 4

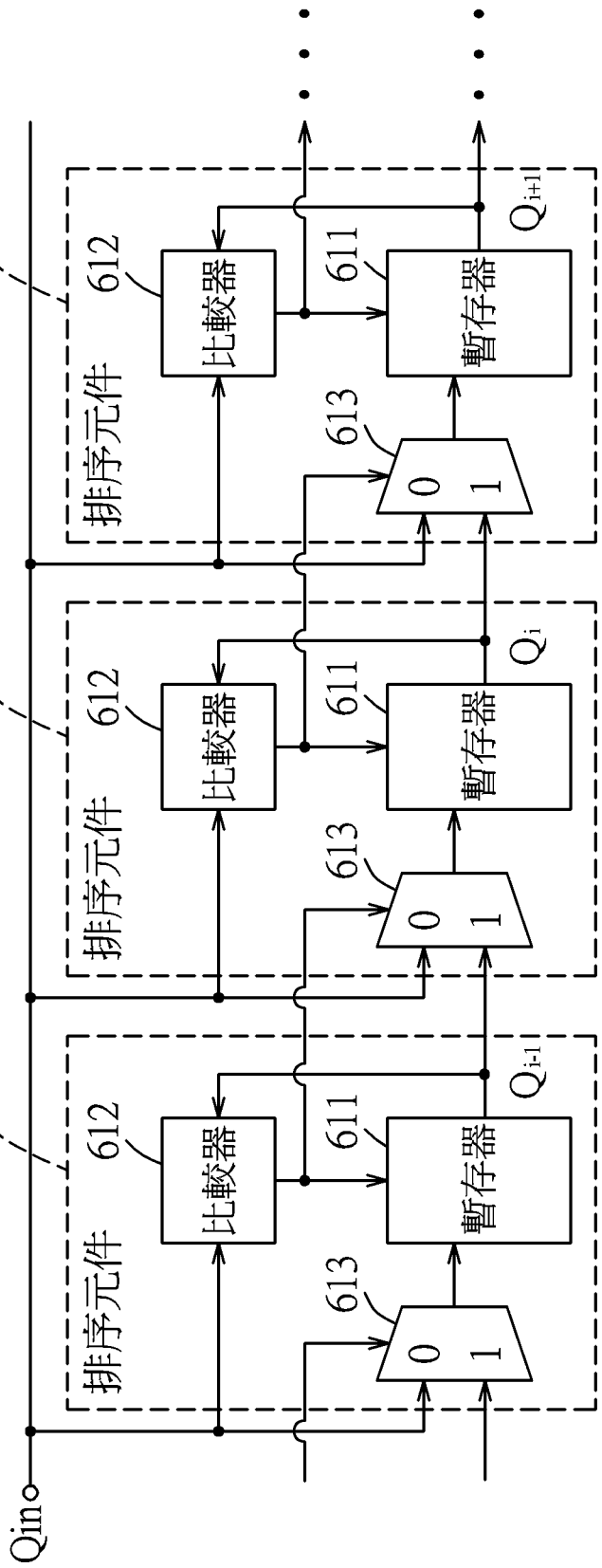

圖5

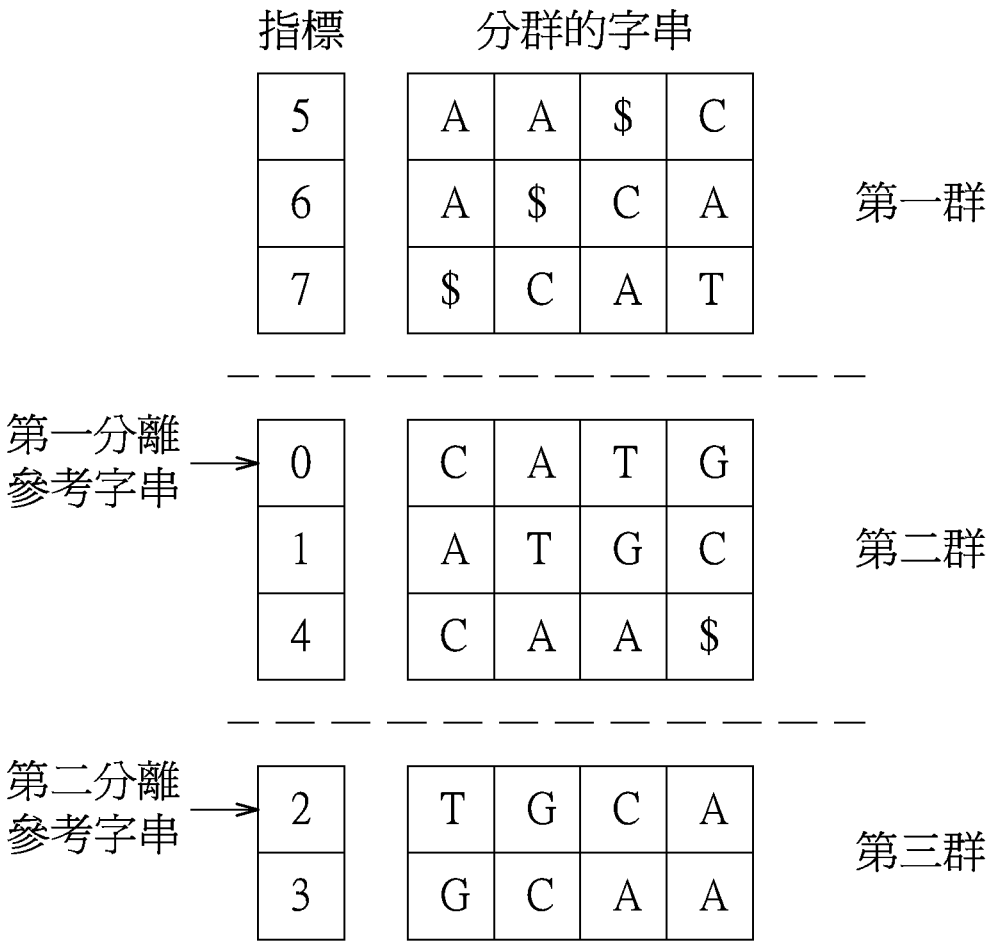
指標
分群的字串
5
A A $ C
6
A $ C A
第一群
7
$ C A T
第一分離參考字串
0
C A T G
1
A T G C
第二群
4
C A A $
第二分離參考字串
2
T G C A
3
G C A A
第三群

圖6

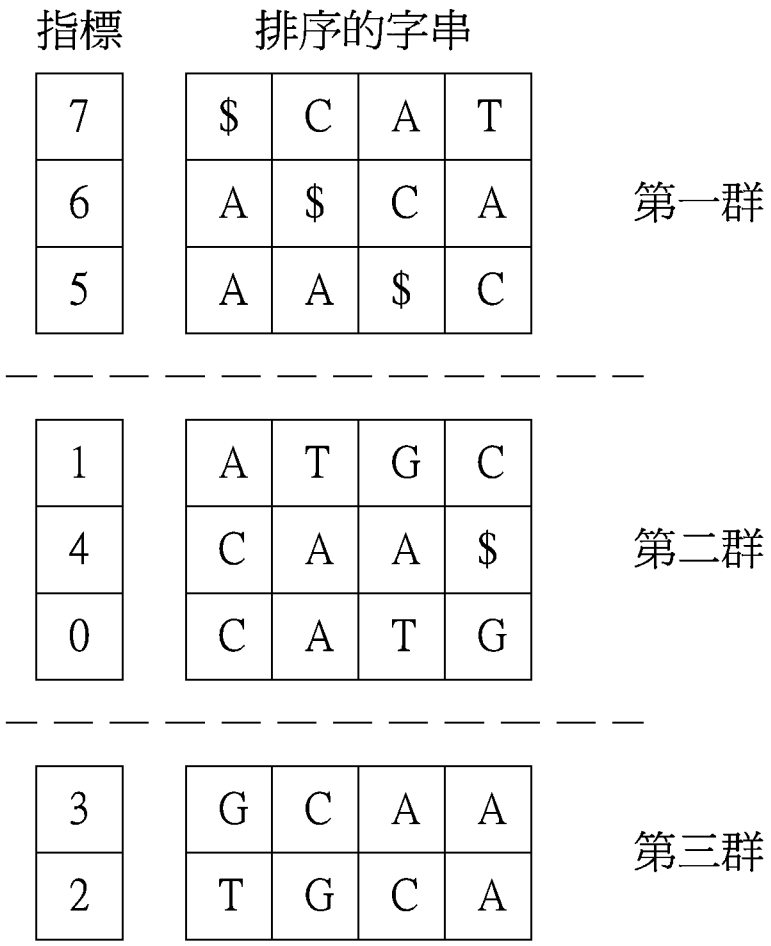
指標
排序的字串
7
$ C A T
6
A $ C A
5
A A $ C
第一群
1
A T G C
4
C A A $
0
C A T G
第二群
3
G C A A
2
T G C A
第三群

圖7

指標	後綴字串陣列							
7	$	C	A	T	G	C	A	A
6	A	$	C	A	T	G	C	A
5	A	A	$	C	A	T	G	C
1	A	T	G	C	A	A	$	C
4	C	A	A	$	C	A	T	G
0	C	A	T	G	C	A	A	$
3	G	C	A	A	$	C	A	T
2	T	G	C	A	A	$	C	A

圖 8

CNT表

字符	列位址
A	0
C	3
G	5
T	6

SA表

列位址	指標
0	7
1	6
2	5
3	1
4	4
5	0
6	3
7	2

F表

列位址	字符
0	$
1	A
2	A
3	A
4	C
5	C
6	G
7	T

L表

列位址	字符
0	A
1	A
2	C
3	C
4	G
5	$
6	T
7	A

OCC表

列位址	字符			
	A表	C表	G表	T表
0	1	0	0	0
1	2	0	0	0
2	2	1	0	0
3	2	2	0	0
4	2	2	1	0
5	2	2	1	0
6	2	2	1	1
7	3	2	1	1

圖9

CNT表

字符	列位址
A	0
C	3
G	5
T	6

部分的SA表

列位址	指標
0	7
4	4

部分的L表

列位址	字符
0	A
4	G

部分的OCC表

列位址	字符			
	A表	C表	G表	T表
0	1	0	0	0
4	2	2	1	0

圖 10

a last column of the suffix string array, and an OCC table associated with the L table. The FM-index data structure can be searched to obtain a search result for a target string.

【指定代表圖】：圖（1）。

【代表圖之符號簡單說明】

100……資料處理系統

1……儲存模組

2……字串產生單元

21……後綴字串產生模組

22……字串產生模組

3……編碼模組

4……分離參考字串選擇模組

41……升取樣單元

42……降取樣單元

5……分群模組

6……排序模組

7……後綴字串陣列產生模組

8……FM-指標資料產生模組

9……搜尋模組