I785847

(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 I785847 B**

(45)公告日： 中華民國 111 (2022) 年 12 月 01 日

(21)申請案號：110138325 (22)申請日： 中華民國 110 (2021) 年 10 月 15 日

(51)Int. Cl.： ***G16B30/00 (2019.01)*** ***G16B40/10 (2019.01)***
G06F17/18 (2006.01)

(71)申請人：國立陽明交通大學 (中華民國) NATIONAL YANG MING CHIAO TUNG UNIVERSITY (TW)
新竹市東區大學路 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) ；陳彥龍 CHEN, YEN-LUNG (TW) ；楊仲萱 YANG, CHUNG-HSUAN (TW)

(74)代理人：高玉駿；楊祺雄

(56)參考文獻：

TW	200422914A	TW	201931181A
CN	103336916A	CN	108256291A
US	2005/0209787A1	US	2014/0297196A1

審查人員：姚乃綺

申請專利範圍項數：11 項 圖式數：28 共 82 頁

(54)名稱

用於處理基因定序資料的資料處理系統

(57)摘要

一種資料處理系統可操作在用於處理與參考 DNA 序列有關的後綴字串資料的預處理模式，或者可操作在與待測 DNA 序列有關的短片段回貼模式、序列重組模式或變體識別模式，並包含可支援在該預處理模式和該序列重組模式中高速處理排序工作的多工排序引擎，以及可支援在該短片段回貼模式和該變體識別模式中的動態編程演算工作的動態編程處理引擎。因此，該資料處理系統能夠實現一種能夠加速並整合 DNA 定序資料分析處理以及大幅降低記憶體需求的系統單晶片。

A data processing system can be operated in a preprocessing mode for processing suffix string data related to a reference DNA sequence, or can be operated in a short-read mapping mode, a sequence assembly mode or a variant calling mode that are related to a DNA sequence to be tested. The data processing system includes a multiplexed sorting engine that can support high-speed processing of sorting tasks in the pre-processing mode and the sequence assembly mode, and a dynamic programming processing engine that can support dynamic programming calculations in the short-read mapping mode and the variant calling mode. Therefore, the data processing system can realize a system-on-chip that can accelerate and integrate DNA sequencing data analysis and processing with greatly reduced memory requirements.

指定代表圖：

TW I785847 B

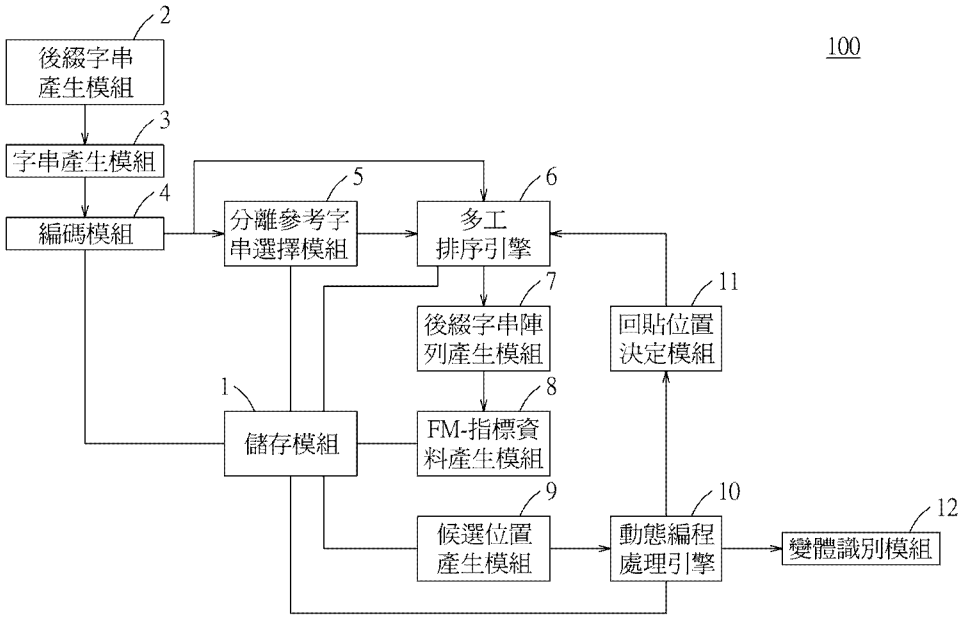

圖1

符號簡單說明：

100:資料處理系統

1:儲存模組

2:後綴字串產生模組

3:字串產生模組

4:編碼模組

5:分離參考字串選擇模組

6:多工排序引擎

7:後綴字串矩陣產生模組

8:FM-指標資料產生模組

9:候選位置產生模組

10:動態編程處理引擎

11:回貼位置決定模組

12:變體識別模組

專利案號：110138325

公告本

I785847

【發明摘要】

【中文發明名稱】 用於處理基因定序資料的資料處理系統

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

【中文】

一種資料處理系統可操作在用於處理與參考DNA序列有關的後綴字串資料的預處理模式，或者可操作在與待測DNA序列有關的短片段回貼模式、序列重組模式或變體識別模式，並包含可支援在該預處理模式和該序列重組模式中高速處理排序工作的多工排序引擎，以及可支援在該短片段回貼模式和該變體識別模式中的動態編程演算工作的動態編程處理引擎。因此，該資料處理系統能夠實現一種能夠加速並整合DNA定序資料分析處理以及大幅降低記憶體需求的系統單晶片。

【英文】

A data processing system can be operated in a preprocessing mode for processing suffix string data related to a reference DNA sequence, or can be operated in a short-read mapping mode, a sequence assembly mode or a variant calling mode that are related to a DNA sequence to be tested. The data processing system includes a multiplexed sorting engine that can support high-speed processing of sorting tasks in the pre-processing mode and the sequence assembly mode, and a dynamic programming processing engine that can support dynamic programming calculations

in the short-read mapping mode and the variant calling mode. Therefore, the data processing system can realize a system-on-chip that can accelerate and integrate DNA sequencing data analysis and processing with greatly reduced memory requirements.

【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…後綴字串產生模組

3…字串產生模組

4…編碼模組

5…分離參考字串選擇模組

6…多工排序引擎

7…後綴字串矩陣產生模組

8.. FM-指標資料產生模組

9…候選位置產生模組

10…動態編程處理引擎

11…回貼位置決定模組

12…變體識別模組

【發明說明書】

【中文發明名稱】 用於處理基因定序資料的資料處理系統

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

【技術領域】

【0001】 本發明是有關於一種資料處理系統，特別是指一種用於處理基因定序資料的資料處理系統。

【先前技術】

【0002】 次代定序(Next-Generation Sequencing，NGS)是目前最快的定序技術，其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。特別是，此技術可應用於產前嬰兒之非侵入式遺傳訊息分析、癌症識別、精準醫療診斷、生物與生醫科技、病毒測試、物種微演化分析等應用，於是相關DNA定序資料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】 因此，如何發展出一種能夠加速並整合DNA定序資料分析處理以及大幅降低記憶體需求的系統單晶片已成為目前重要

的議題之一。

【發明內容】

【0004】 因此，本發明的目的，即在提供一種用於處理基因定序資料的資料處理系統，其能克服現有技術的至少一缺點。

【0005】 於是，本發明所提供的一種資料處理系統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由四個分別代表四種不同含氮鹼基的字符A，C，G，T組成的(N-1)個字符之參考DNA序列以及一位在該參考DNA序列之後代表序列結束的字符$的參考序列的N個後綴字串、多個分別指示出該等N個字符在該參考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，以及多個擷取自一待測DNA序列的短片段。該資料處理系統可操作在與該參考DNA序列有關的一預處理模式，或可操作在與該待測DNA序列有關的一短片段回貼模式、一序列重組模式及一變體識別模式其中一者，並包含：一字串產生模組；一編碼模組，連接該字串產生模組；一分離參考字串選擇模組；一多工排序引擎，連接該分離參考字串選擇模組；一後綴字串矩陣產生模組；連接該多工排序模組；一FM-指標資料產生模組，連接該後綴字串矩陣生模組；一候選位置產生模組；一動態編程處理引擎，連接該候選位置產生模組；一回貼位置決定模組，連接該多工排序引擎和該動態編程處理引擎；

及一變體識別模組，連接該動態編程處理引擎。

【0006】 當該資料處理系統操作在該預處理模式時：該字串產生模組擷取該等N個後綴字串其中的每一者的前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；該編碼模組利用一將該等字符$，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，將該等N個後綴字串編碼以產生N個分別對應於該等N個指標且具有一數字碼形式的編碼字串，並將該參考DNA序列和該等短片段以相同的編碼方式編碼以產生對應於該參考DNA序列的參考編碼字串和多個分別對應於該等短片段的待測編碼字串；該分離參考字串選擇模組以一升取樣方式從該等N個編碼字串選出P×Q個編碼字串提供給該多工排序引擎其中P代表分離參考字串的數量且Q代表取樣倍數，以使該多工排序引擎依照編碼值將該P×Q個編碼字串排序，然後以一降取樣方式從該排序的P×Q個編碼字串選出P個依照編碼值從小到大排列的編碼字串分別作為第一至第P分離參考字串；該多工排序引擎操作來根據根據該分離參考字串選擇模組選出的該第一至第P分離參考字串將該編碼模組產生的該N個編碼字串分成(P+1)群、並將該(P+1)群其中每一群的編碼字串依照編碼值從小到大排序，以獲得該N個編碼字串依照編碼值從小到大的排序結果；該後綴字串矩陣產生模組根據來自該多工排序引擎的該排序結果，產生一對應於該參考DNA序

列的後綴字串矩陣；及該FM-指標資料產生模組根據來自該後綴字串矩陣產生模組的該後綴字串矩陣及該等指標，建立一對應於該參考DNA序列的FM-指標資料結構，其中該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矩陣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矩陣的一最後字符欄的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矩陣中第一至第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。

【0007】 當該資料處理系統操作在該短片段回貼模式時；該候選位置產生模組將該等短片段其中每一者分割成多個小片段，然後根據該FM-指標資料產生模組產生的該FM-指標資料結構，對於每一小片段，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，以獲得一個或多個代表該小片段在該待測DNA序列中的候選位置的指標；該動態編程處理引擎操作來根據來自該候選位置產生模組對於每一短片段的該等小片段所獲得的所有指標，執行每一短片段與該參考DNA序列中在每一候選位置擷取的對應參考片段的相似度演算，以獲得對應於該候選位置的相

似度分數；及該回貼位置決定模組將根據該動態編程處理引擎對於每一短片段所獲得的所有相似度分數中的最高者對應的指標所代表的候選位置決定為該短片段的回貼位置。

【0008】 當該資料處理系統操作在該序列重組模式時，該多工排序引擎操作來根據與該等短片段對應的回貼位置以及該編碼模組產生的該參考編碼字串和該等待測編碼字串，重組出有關於該待測DNA序列的一個或多個編碼序列組合，該(等)編碼序列組合各自代表一對應的半倍體序列。

【0009】 當該資料處理系統操作在該變體識別模式時；該動態編程處理引擎操作來執行該參考DNA序列和每一半倍體序列的相似度演算，以產生對應於該半倍體序列的一相似度分數矩陣表、及一與分數來源方向有關的方向矩陣表；及對於每一半倍體序列，該變體識別模組根據該動態編程處理引擎產生對應於該半倍體序列的該相似度分數矩陣表和該方向矩陣表，從該相似度分數矩陣表確認其中出現最高分數的位置，然後從該方向矩陣表獲得達到該位置的方向軌跡，且至少根據該方向軌跡識別出存在於該倍半體序列中的每一變體的位置並推估出對應於每一變體的突變類型。

【0010】 本發明之功效在於：由於使用了擷取自該等後綴字串的前數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，因此可以有效降低排序時的複雜度並大量降低在建立該FM-指

標資料結構期間所需的記憶體使用量。此外，該多工排序引擎和該動態編程處理引擎各自可以在不同模式中操作使用，藉此實現硬體共用優點。另外，該多工排序引擎包含大量彼此串接的排序單元，其適於支援如需高速處理資料的排序和比對操作；而該動態編程處理引擎可以被實施成一維架構的運算電路架構，相較於傳統的二維運算單元，可以大幅減少電路面積。 因此，該資料處理系統能夠實現一種能夠加速並整合DNA定序資料分析處理以及大幅降低記憶體需求的系統單晶片。

【圖式簡單說明】

【0011】 本發明之其他的特徵及功效，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，示例性地說明本發明實施例的資料處理系統；

圖 2 示例性地說明該實施例的一後綴字串產生模組根據一參考序列所產生的後綴字串及其所對應的指標；

圖 3 示例性地說明該實施例的一字串產生模組根據圖 2 的後綴字串產生的字串；

圖 4 示例性地說明該實施例的一後綴字串矩陣產生模組所產生對應於圖 2 的後綴字串的一後綴字串矩陣及其所對應的指標；

圖 5 示例性地說明該實施例的一 FM-指標資料產生模組所產生

一對應於圖 2 的後綴字串的 FM-指標資料結構；

圖 6 示例性地說明該實施例的一儲存模組中儲存圖 5 所示的 FM-指標資料結構的一部分；。

圖 7 是一示意圖，說明該實施例的一多工排序引擎的架構；

圖 8 是一示意圖，繪示出該多工排序引擎中的每一排序單元所具有的輸入端與輸出端；

圖 9 是一電路圖，示例性地說明每一排序單元的組成元件以連續三個排序單元之間的連接關係；

圖 10 是一示意圖，示例性地繪示出該實施例的一動態編程處理引擎的架構；

圖 11 是一電路圖，示例地繪示出該動態編程處理引擎所含的每一處理單元的組成；

圖 12 是一等效電路圖，說明該多工排序引擎如何執行字串排序操作；

圖 13 是一等效電路圖，說明該多工排序引擎如何執行字串分群操作；

圖 14 示例性地說明該實施例的一動態編程處理引擎如何執行動態編程演算來獲得一相似度分數矩陣表；

圖 15 至圖 21 是等效電路圖，示例性地說明該多工排序引擎如何建立一短片段的德布魯因建表；

圖 22 至圖 24 是等效電路圖，示例性地說明該多工排序引擎如何重組出一短片段的編碼序列；

圖 25 是一示意圖，示例性地說明該多個回貼的短片段、及在重組過程中的序列；

圖 26 是一示意圖，示例性地說明該動態編程處理引擎所獲得的一相似度分數矩陣表和一方向矩陣表；

圖 27 是一示意圖，示例性地說明該實施例中使用有關基因變異的生物模型；及

圖 28 示例性地說明該動態編程處理引擎的每一運算單元分別操作在單點突變、插入突變和刪除突變之可能性演算時的等效電路圖。

【實施方式】

【0012】 在本發明被詳細描述之前，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。

【0013】 參閱圖1，所繪示的本發明實施例的資料處理系統100係用於處理與一參考DNA序列(例如但不限於人類DNA序列)和一待測DNA序列有關的基因定序資料。在本實施例中，該參考DNA序列具有(N-1)個字符，其係由至少四個分別代表四種不同含氮鹼基(例如分別為腺嘌呤、胞嘧啶、鳥嘌呤及胸腺嘧啶)的字符A，C，G，T所組成，而最後一個字符為一代表序列結束的字符$。然而，值得注意的是，在實際使用時，該參考DNA序列亦可含有一個或多

個異於該等字符A，C，G，T的字符，此(等)字符用來表示尚未被確認的含氮鹼基。該基因定序資料例如包含相關於一具有該參考DNA序列和一位在該參考DNA序列之後代表序列結束的字符$的參考序列(其具有N個字符)的N個後綴字串、多個分別指示出該等N個字符在該參考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，以及多個擷取自該待測DNA序列的短片段(Short Reads)。該資料處理系統100可包含：一儲存模組1；一後綴字串產生模組2；一連接該後綴字串產生模組2的字串產生模組3；一連接該儲存模組1和該字串產生模組3的編碼模組4；一連接該儲存模組1和該編碼模組的4分離參考字串選擇模組5；一連接該儲存模組1、該編碼模組4和該分離參考字串選擇模組5的多工排序引擎6；一連接該多工排序引擎6的後綴字串陣列產生模組7；一連接該儲存模組1和該後綴字串陣列產生模組7的FM-指標資料產生模組8；一連接該儲存模組1的候選位置產生模組9；一連接該儲存模組1和該候選位置產生模組9的動態編程處理引擎10；一連接該多工排序引擎6和該動態編程處理引擎10的回貼位置決定模組11；及一連接該動態編程處理引擎的變體識別模組12。

【0014】 該儲存模組1是用來儲存該參考DNA序列和該等N個指標、該等短片段，以及在該資料處理系統100操作期間所產生的相關資料(將詳細說明於下文中)。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序列可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序列的該等N個字符(其包含該參考DNA序列的(N-1)個字符和一位在最後的字符$)與該等N個指標的關係，其中N=11，且該等十一個字符及該等十一個指標如以下表1所示：

表1

指標	0	1	2	3	4	5	6	7	8	9	10
字符	C	A	T	G	A	A	A	G	G	A	$

【0015】 該後綴字串產生模組2是用來產生與該參考序列有關的後綴字串。

【0016】 該字串產生模組3是用來產生從該後綴字串產生模組2所產生的每一後綴字串擷取出前K個字符的對應字串。

【0017】 該編碼模組4是用來對該字串產生模組3所產生的字串以及該儲存模組1儲存的該參考DNA序列和該等短片段進行編碼。具體而言，該編碼模組3可以依照一將該等字符$，A，C，G，T，分別以五個彼此不同且具有遞增數值的數字碼來表示的編碼方式，來編碼由該字串產生模組3所產生的每一字串，以產生具有一數字碼形式的對應編碼字串以產生N個具有一數字碼形式且分別對應於該等N個指標的編碼字串。例如，針對每一字串，該等字符$，A，

C，G，T可分別被編碼成000、001、010、011及100的數字碼，而針對每一短片段(其不含有字符$)及該參考DNA序列，該等字符A，C，G，T可分別被編碼成00、01、10及11，但不以此例為限。

【0018】 該分離參考字串選擇模組5是用來從該編碼模組4針對所有字串的編碼結果選出適當的分離參考字串，並將所有選出的分離參考字串儲存於該儲存模組1。

【0019】 再參閱圖7，該多工排序引擎6可包含多個彼此串接的排序單元61、及一耦接該等排序單元61的加法器62。

【0020】 再參閱圖8與圖9，每一排序單元61具有一用於接收來自外部的待處理資料的第一資料輸入端data_in、一用於接收來自前一級的排序單元(圖為示)的輸出資料的第二資料輸入端data_pre、一用於接收來自該前一級的排序單元的一第一控制信號的第一控制輸入端EN_pre、一用於接收來自外部的一第二控制信號的第二控制輸入端mode、一用於輸出資料給下一級的排序單元(圖未示)的第一輸出端data_out、一用於輸出提供給該下一級的排序單元的第一控制信號的第二輸出端EN、一第三輸出端result和一第四輸出端target。簡言之，對於每一排序單元61(第一級的排序單元除外)而言，該第二資料輸入端data_pre耦接該前一級的排序單元61的第一輸出端data_out，該第一控制輸入端EN_pre耦接該前一級的排序單元61的第二輸出端EN，該第一輸出端data_out耦接

該後一級的排序單元61的第二資料輸入端data_pre，該第二輸出端EN耦接該後一級的該第一控制輸入端EN_pre(見圖9)；而對於第一級排序單元61的該第二資料輸入端data_pre和該第一控制輸入端EN_pre可在不同的操作模式下提供適當的資料及控制信號。此外，所有排序單元61同步接收來自外部的輸入資料及該第二控制信號。在本實施例中，該加法器62具有多個輸入端(其分別耦接該等排序單元61的該等第三輸出端result，圖未示出)、及一輸出端。

【0021】 如圖9所示，每一排序單元61可包含一暫存器611、一比較器612、一第一2×1多工器613、一3×1多工器614、一第二2×1多工器615、一反閘616及一及閘617。該暫存器611具有一用於接收一時脈信號的時脈輸入端、一用於接收資料的輸入端、及一耦接該排序單元61的該第一輸出端data_out的輸出端(用於輸出該暫存器611所暫存的資料(以Q_i來表示))。該比較器612具有一耦接該排序單元61的該第一資料輸入端data_in的第一輸入端、一耦接該暫存器611的該輸出端的第二輸入端、及一耦接該排序單元61的該第二輸出端EN和該第三輸出端result的輸出端，並且當該第二輸入端接收的信號邏輯值大於或等於該第一輸入端接收的信號的邏輯值時，該比較器612在該輸出端輸出邏輯1的信號，反之，則輸出邏輯0的信號。該第一2×1多工器613具有一耦接該排序單元61的該第一資料輸入端data_in的第一輸入端、一耦接該排序單元61的該第二

資料輸入端data_pre的第二輸入端、一耦接該排序單元61的該第一控制輸入端EN_pre的控制端、及一輸出端，並且當該控制端接收一邏輯0的信號時，該第一輸入端連接該輸出端，而當該控制端接收一邏輯1的信號時，該第二輸入端連接該輸出端。該3×1多工器614具有一耦接該前一級的排序單元61的第一輸出端data_out的第一輸入端(用於接收來自該前一級的排序單元61的輸出資料(以Q_{i-1}來表示))、一耦接後一級的排序單元61的第一輸出端data_out的第二輸入端(用於接收來自該前一級的排序單元61的輸出資料(以Q_{i+1}來表示))、一耦接該第一2×1多工器的該輸出端的第三輸入端、一作為該排序單元61的該第二控制輸入端mode的控制端、及一輸出端，並且根據該控制端所接收的一控制信號來使該第一至第三輸入端其中一者與該輸出端連接或使該第一至第三輸入端與該輸出端之間呈高阻抗。該第二2×1多工器615具有一耦接該暫存器611的該輸出端的第一輸入端、一耦接該3×1多工器614的該輸出端的第二輸入端、一耦接該比較器612的輸出端的控制端、及一耦接該暫存器611的該輸入端的輸出端，並且當該控制端接收一邏輯0的信號時，該第一輸入端連接該輸出端，而當該控制端接收一邏輯1的信號時，該第二輸入端連接該輸出端。該反閘616具有一耦接該排序單元61的該第一控制輸入端的輸入端、及一輸出端。該及閘617具有一耦接該反閘616的該輸出端的第一輸入端、一耦接該比

較器612的該輸出端的第二輸入端、及一作為該排序單元61的該第四輸出端target的輸出端。

【0022】 請注意，該多工排序引擎6中的該等排序單元61是回應於相同的時脈信號(提供給暫存器611)和相同的第二控制信號(提供給該3×1多工器614)來運作，並且該時脈信號和該第二控制信號可由外部的一控制電路(圖未示)根據該資料處理系統100所處的操作模式而產生。

【0023】 再參閱圖10與圖11，在本實施例中，該動態編程處理引擎10可包含多個大致呈陣列排列的運算單元101、及一用於儲存該等運算單元101之運算結果的緩衝器102。如圖11所示，每一運算單元101可以是已知的Smith-Waterman運算單元，其包含三個信號輸入端(用於接收如$H_{(i-1,j-1)}$，$H_{(i-1,j)}$，$H_{(i,j-1)}$等輸入信號)、四個參數輸入端(用於接收如T1，T2，T3，S等參數)、一個控制信號端(用於接收如mode的控制信號)及一個輸出端(用於輸出如$H_{(i,j)}$的輸出信號)，其中該等信號輸入端分別耦接上方、左方及左上方運算單元的輸出端。如圖11所示，每一運算單元101可包含三個加法器、一線性整流單元(ReLU)、一比較器組件(max)和一2×1多工器，並可操作來進行如以下式1的運算：

$$H(i,j) = max\begin{cases}(H(i-1,j-1) + T1) + S \\ (H(i-1,j) + T2) + S \\ (H(i,j-1) + T3) + S \\ 0\end{cases} \quad (式1)$$

其中T1，T2，T3和S為參數。由於此Smith-Waterman運算單元具有已知的電路結構，且並非本實施例的主要特徵，故在此省略其組件的詳細操作而不再贅述。

【0024】 在本實施例中，該資料處理系統100可以操作在與該參考DNA序列有關的一預處理(Preprocessing)模式，或者可以操作在與該待測DNA序列有關的一短片段回貼(Short-Read Mapping)模式、一序列重組(Sequence Assembly)模式及一變體識別(Variant Calling)模式其中一者。以下，將針對該資料處理系統100操作在上述每一模式時，進一步示例性地說明相關組件各自的詳細操作或處理。

【0025】 當該資料處理系統100操作在該預處理模式時，首先，該後綴字串產生模組20根據該參考序列及該等指標(其可從外部輸入或從該儲存模組1讀取)，從該參考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，該參考序列例如為“CATGAAAGGA$”)時，該後綴字串產生模組2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。

【0026】 接著，該字串產生模組3擷取來自該後綴字串產生模組2的該等N個後綴字串其中的每一者的前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，若沿用圖2的示例且K=4的情況下，該字串產生模組3所產生的該等十一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=11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。

【0027】 然後，該編碼模組4利用上述的編碼方式，將來自該字串產生模組3的該等N個字串編碼以產生N個分別對應於該等N個指標且具有一數字碼形式的編碼字串，另一方面，該編碼模組4以相同編碼方式將該等短片段與該參考DNA序列進行編碼以產生多個分別對應於該等短片段的待測編碼字串和一對應於該參考DNA序列的參考編碼字串，並將產生的該等待測編碼字串和該參考編碼字儲存於該儲存模組1。

【0028】 接著，該分離參考字串選擇模組5先以一升取樣方式從該等N個編碼字串選出P×Q個編碼字串提供給該多工排序引擎6其中P代表分離參考字串的數量且Q代表取樣倍數，以使該多工排序引擎6依照編碼值將該P×Q個編碼字串排序，然後以一降取樣方式從該多工排序引擎6輸出的已排序的P×Q個編碼字串選出P個依照編碼值從小到大排列的編碼字串分別作為第一至第P分離參考字串，並將該一至第P分離參考字串儲存於該儲存模組1。值得注意的

是，當該多工排序引擎6操作來對於該P×Q個編碼字串排序時，在此情況下，每一排序單元61會操作成如圖12的等效電路(其中每一排序單元61的該3×1多工器614將使其第三輸入端和輸出端保持連接)。在此配置下，當該第一資料輸入端data_in依序接收該P×Q個編碼字串時，具有越小編碼值的編碼字串越容易被優先輸出而達到排序的目的。於是，經過數個時脈週期後，該多工排序引擎6會最先輸出最小編碼值的編碼字串，而最後輸出最大編碼值的編碼字串。舉例來說，當沿用圖3所示的該等字串的情況時，指標分別為0及5的字串，即CATG及AAGG，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。

【0029】 接著，該多工排序引擎6操作來根據該儲存模組1儲存的該第一至第P分離參考字串將該編碼模組4產生的該N個編碼字串分成(P+1)群、並將該(P+1)群其中每一群的編碼字串依照編碼值從小到大排序，以獲得該N個編碼字串依照編碼值從小到大的排序結果。更具體地，該多工排序引擎6會先將該第一至第P分離參考字串分別紀錄/儲存於其中的P個排序單元61的暫存器611，接著使此P個排序單元61的每一者操作成如圖13所示的一等效電路(其中

該3×1多工器614由於內部的高阻抗而不運作，致使該第二2×1多工器615亦不運作)。在此情況下，該P個排序單元61的第一資料輸入端data_in會依序接收到該N個編碼字串，並對應於每一次接收到的編碼字串，該多工排序引擎6根據該加法器9(見圖7)的輸出值來決定本次的編碼字串被分到的一群。舉例來說，在沿用上例的情況下，若該加法器7的輸出值為2時，本次的編碼字串將被分到第一群；若該加法器的輸出值為1時，本次的編碼字串將被分到第二群；若該加法器的輸出值為0時，本次的編碼字串將被分到第三群。然後，該多工排序引擎6依照如圖12的操作方式並以第一、第二、第三群的順序將每一群的編碼字串排序，最後便可獲得編碼值從小到大的N個排序的編碼字串的排序結果。值得注意的是，由於該多工排序引擎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。

【0030】 接著，該後綴字串陣列產生模組7根據來自該多工排序引擎6的該排序結果(即，已排序的N個編碼字串)，產生一對應於該參考DNA序列的後綴字串陣列。舉例來說，在沿用圖2所示的該等後綴字串的情況下，該後綴字串陣列產生模組7根據對應於圖3的該十一個編碼字串的排序結果所獲得的後綴字串陣列以及其所對應的該等指標係如圖4所示。

【0031】 最後，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的該後綴字串陣列及該等指標，並據以建立一對應於該參考DNA序列的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，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4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5所示。

【0032】 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F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R1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R2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5所示的FM-指標資料結構的情況下，當R1=R2=3時，該部分的FM-指標資料結構係如圖6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。

【0033】 當該資料處理系統100操作在該短片段回貼模式時，首先，該候選位置產生模組9將該儲存模組1儲存的每一短片段分割成多個小片段(Seeds)，然後根據儲存於該儲存模組1的該(完整或部分的)FM-指標資料結構，對於每一小片段，利用一相關於後進搜尋方式的指標演算法搜尋該(完整的)FM-指標資料結構中的資料，以獲得一個或多個代表該小片段在該待測DNA序列中的候選位置的指標。在本實施例中，若所欲搜尋的小片段被表示為“$S_1S_2..S_M$”，該指標演算法可由以下式2、式3及式4來實現：

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

$$index_{min}[i] = CNT[S[i]] + OCC[index_{min}[i-1]-1, S[i]] + 1 \quad (式3)$$

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

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$index_{min}[i]$及$index_{max}[i]$分別代表在第i次迭代搜尋運算中與該目標字符可能所在的最小指標及最大指標有關的列位址，並且其初始值分別被定義為$index_{min}[0]=0$及$index_{max}[0]=N-1$。

【0034】 請注意，在該儲存模組1僅儲存了例如圖6所示的該部份的FM-指標資料結構的情況下，該候選位置產生模組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，並重新獲得該F表。更明確地說，該候選位置產生模組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，該候選位置產生模組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可由以下式5及式6來實現：

$$OCC[n.s]=OCC_D[\left\lfloor\frac{n}{T2},s\right\rfloor]+L[\left\langle\left\lfloor\frac{n}{T2}\right\rfloor+1,n\right\rangle,s] \quad (式5)$$

$$SA[n]=SA_D[CNT[L[n]]+OCC[n,L[n]]]+1 \quad (式6)$$

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，CNT代表該CNT表，SA_D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R2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。

【0035】 舉例來說，若沿用圖4所示的FM-指標資料結構，對於如“CATG”的一短片段，該候選位置產生模組9可獲得從“CATG”分成的兩個小片段，即，第一小片段“CA”和第二小片段“TG”。首先，對於第一小片段“CA”，該候選位置產生模組9利用上述式2而獲得S[1]=A(即，第1次迭代搜尋運算的目標字符)，且利用上述式3及式4並查找圖4中的該CNT表及該OCC表來執行第1次迭代搜尋運算，以獲得$index_{min}[1]$及$index_{max}[1]$。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10的資料，因此OCC[-1,A]被預設為0，此外$index_{min}[0]=0$及$index_{max}[0]=10$。於是，$index_{min}[1] = CNT[A] + OCC[index_{min}[0] - 1, A] + 1 = 0 + 0 + 1 = 1$，且$index_{max}[1] = CNT[A] + OCC[index_{max}[0], A] = 0 + 5 = 5$。然後，在第2次迭代搜尋運算中，同樣地，該候選位置產生模組9利用上述式2而獲得S[2]=C(即，第2次迭代搜尋運算的目標字符)，且利用上述式3及式4並查找圖4中的該CNT表及該OCC表來執行第2次迭代搜尋運算，以獲得$index_{min}[2]$及$index_{max}[2]$。於是，$index_{min}[2] = CNT[C] + OCC[index_{min}[1] - 1, C] + 1 = 5 + 0 + 1 = 6$，且$index_{max}[2] = CNT[C] + OCC[index_{max}[1], C] = 5 + 1 = 6$。最後，透過查找圖4中的該SA表可獲得代表第一小片段“CA”在該待測DNA序列的候選位置

的指標，即，SA[6]=0。並且以相似於搜尋該第一小片段“CA”的指標的演算方式，可獲得代表第二小片段“TG”在該待測DNA序列的候選位置的指標(即，2)。

【0036】 因此，重複執行上述演算，該候選位置產生模組9可以獲得對應於其他短片段的小片段的指標。請注意，將每一短片段先分割成小片段後在進行搜尋的好處可以有效避免因存在於短片段的變異而無法搜尋到回貼位置。

【0037】 然後，該動態編程處理引擎10操作來根據來自該候選位置產生模組9對於每一短片段的該等小片段所獲得的所有指標，執行每一短片段與該參考DNA序列中在每一候選位置擷取的對應參考片段的相似度演算，以獲得對應於該候選位置的相似度分數。更具體地，該動態編程處理引擎10利用動態編程演算法，且根據來自該該候選位置產生模組9對於每一短片段的該等小片段獲得的所有指標，將每一短片段和該參考DNA序列中在與分割自該短片段的每一小片段對應的每一候選位置所擷取的對應參考片段進行字符比對，並根據字符比對結果執行作為該相似度演算的Smith-Waterman演算(如上式1所示)。特別要說明的是，該短片段和該對應參考片段的相似度可以一個二維矩陣(Matrix)的形式來表示，此矩陣的每一元素(element)可以存放一代表相似度的分數(分數越高代表相似程度越高，分數越低代表相似程度越低)，每一元素的分

數都是根據字符比對結果以及在其上方、左方或左上的元素的分數並透過上述式1的演算而獲得。在式1的演算中，T1=T2=T3=0，且當比對的字符相同時，$S=S_m$(其為一大於零的正整數)，而當比對的字符不同時，$S=S_p$(其為一小於零的負整數)。分數的計算是從矩陣的左上角的元素開始，並往右下方向逐層進行直到整個矩陣內的元素的分數都計算出，以獲得該短片段對應於該候選位置的一相似度分數矩陣表。該相似度分數矩陣表可被儲存於該緩衝器102(見圖10)，並且其中的最高相似度分數代表該短片段和該對應參考片段的相似程度，並作為對應於該候選位置的相似度分數。

【0038】 舉例來說，參閱圖14，沿用上述短片段“CATG”的示例，該動態編程處理引擎10將該短片段“CATG”與該參考DNA序列對應於指標”0”(其為針對該第一小片段所獲得的指標)所代表的候選位置擷取的對應參考片段“CATG”進行每一字符的動態比對，並利用上述式1來演算出每一運算單元101所儲存的分數值。在本例中，式1中的$S_p=5$且$S_m=-2$，但不在此限。於是，在經過一個運算週期(1 cycle)後，由於該短片段的第一字符”C”相同於該對應參考片段的第一字符”C”，所以圖10中的運算單元101_{11}所儲存的分數為5；在經過兩個運算週期(2 cycles)後，由於該短片段的第二字符”A”不同於該對應參考片段的第一字符”C”，所以圖10中的運算單元101_{12}所儲存的分數為3(=5-2)，同時由於該短片段的第一字

符”C”不同於該對應參考片段的第二字符”A”，所以圖10中的運算單元101_{21}所儲存的分數為3(=5-2)；在經過三個運算週期(3 cycles)後，由於該短片段的第三字符”T”不同於該對應參考片段的第一字符”C”，所以圖10中的運算單元101_{13}所儲存的分數為1(=3-2)，由於該短片段的第二字符”A”相同於該對應參考片段的第二字符”A”，所以圖10中的運算單元101_{21}所儲存的分數為10(=5+5)，由於該短片段的第一字符”C”不同於該對應參考片段的第三字符”T”，所以圖10中的運算單元101_{31}所儲存的分數為1(=3-1)；同理，在經過七個運算週期(7 cycles)後，圖10中的運算單元101_{11}~101_{44}所儲存的分數(見圖14)構成該短片段“CATG”對應於該指標”0”所代表的候選位置的相似度分數矩陣表，其中的最高相似度分數(即，該運算單元101_{44}所儲存的分數)作為對應於該候選位置(即，該指標”0”)的相似度分數。此外，對於該短片段“CATG”，仍須將其與該參考DNA序列對應於指標”2”(其為針對該第二小片段所獲得的指標)所代表的候選位置擷取的對應參考片段(同樣為“CATG”)進行每一字符的動態比對，以便獲得對應於該指標”2”的相似度分數。由於該參考DNA序列對應於該指標”2”擷取的對應參考片段相同於對應於該指標”0” 擷取的對應參考片段，因此對應於該指標”2”的相似度分數亦為20。

【0039】 然後，該回貼位置決定模組11將根據該動態編程處理

引擎10的緩衝器102所儲存對於每一短片段所獲得的所有相似度分數中的最高者對應的指標所代表的候選位置決定為該短片段的回貼位置。如此，該回貼位置決定模組11可獲得多個分別對應於該等短片段的回貼位置。

【0040】 當該資料處理系統100操作在該序列重組模式時，該多工排序引擎6操作來根據該儲存模組1所儲存與該等等短片段對應的該等待測編碼字串和對應於該參考DNA序列的該參考編碼字串，以及來自於該回貼位置決定模組11的該等短片段各自的回貼位置，重組出有關於該待測DNA序列的一個或多個編碼序列組合。該(等)編碼序列組合各自代表一對應的半倍體序列(Haplotype Sequence)，且該(等)半倍體序列包含該參考DNA序列。更明確地說，若該待測DNA序列未出現有任何的變體，則對應於該等短片段的該等待測編碼字串與該參考編碼字串僅會重組出單一個編碼序列組合，其所代表的半倍體序列就是該參考DNA序列。在本實施例中，為了更有效率地重組出該編碼序列組合，必須先獲得對應於該參考DNA序列與該等短片段其中每一者的德布魯因(de Bruijn)建表。

【0041】 以下，將參閱圖15至圖18示例性地詳細說明該多工排序引擎6如何建立該參考DNA序列或每一短片段的德布魯因(de Bruijn)建表以及如何利用對應於該參考DNA序列和該等短片段

的德布魯因建表重組出該(等)編碼序列組合。

【0042】 首先，該多工排序引擎6透過對於每一排序單元61的該第一2×1多工器614、該3×1多工器614和該第二2×1多工器615的控制使該排序單元61的該暫存器611儲存了一與一具有(k+1)個相同字符(含氮鹼基)的片段對應且具有相對最大編碼值的參考子編碼序列。舉例來說，如圖15所示(僅示出第1級至第3級的排序單元)，每一排序單元61的暫存器611儲存的參考子編碼序列為”11111111”，其對應於具有例如4(即，k=3)個相同字符”T”的片段”TTTT”。請注意，為了容易理解，以下將第1~3級的排序單元61的暫存器611輸出的資料分別以Q_1、Q_2及Q_3來表示，並以字符的形式來表示Q_1、Q_2及Q_3的資料的內容(即，在圖15的情況下，$Q_1=Q_2=Q_3$=TTTT)，然而，實際上在運作時，暫存器611所儲存的資料為數位編碼(即，”11111111”)。此外，只有第1級的排序單元61的第一2×1多工器613根據一邏輯0的第一控制信號而保持其第一輸入端與該輸出端連接，而每一排序單元61的該3×1多工器614根據該第二控制信號維持該第三輸入端與該輸出端的連接，如圖15所示。

【0043】 然後，該多工排序引擎6使每一排序單元61的該第一資料輸入端data_in依序接收對應於每一短片段的待測編碼字串(或對應於該參考DNA序列的參考編碼字串)的所有與連續(k+1)個字

符有關的子編碼序列，以便將該待測編碼字串(或該參考編碼字串)的每一子編碼序列紀錄在該等排序單元61其中一個對應的排序單元61的該暫存器611中，以完成與該短片段(或該參考編碼字串)有關的德布魯因建表。舉例來說，仍沿用上例，亦即在每一排序單元61的暫存器611已儲存有”TTTT”的資料的情況下，若一短片段為”ACAATT”(亦可被視為一德布魯因序列)時，首先，如圖16所示，該多工排序引擎6使每一排序單元61的該第一資料輸入端data_in接收與該短片段的前4個字符”ACAA”(其可代表第一個4-mer)對應的子編碼序列，於是，每一排序單元61的比較器612會將接收到且對應於”ACAA”的子編碼序列與對應於”TTTT”的參考子編碼序列進行比較，若該參考子編碼序列之值大於接收到的子編碼序列之值時，該比較器612會輸出邏輯1的控制信號給該第二2×1多工器615，否則，該比較器612會輸出邏輯0的控制信號給該第二2×1多工器615。因此，經過一個時脈週期後，第1級的排序單元61的暫存器611所儲存的資料會更新為對應於”ACAA”的子編碼序列，而其他排序單元61的暫存器611所儲存的資料保持不變(即，仍為對應於”TTTT”的參考子編碼序列)，如圖17圖所示。接著，如圖18所示，當每一排序單元61的該第一資料輸入端data_in接收與”CAAT”(其可代表第二個4-mer)對應的子編碼序列，每一排序單元61的比較器612會將接收到且對應於”CAAT”的子編碼序列

與其暫存器611所儲存的資料進行比較。於是，經過一個時脈週期後，第1級的排序單元61的暫存器611所儲存的資料保持不變(即，仍為對應於”ACAA”的子編碼序列)，第2級的排序單元61的暫存器611所儲存的資料被更新為對應於”CAAT”的子編碼序列而其他排序單元61的暫存器611所儲存的資料保持不變(即，仍為對應於”TTTT”的參考子編碼序列)，如圖19所示。繼而，如圖20所示，當每一排序單元61的該第一資料輸入端data_in接收與”AATT”(其可代表第三個4-mer)對應的子編碼序列時，每一排序單元61的比較器612會將接收到且對應於”AATT”的子編碼序列與其暫存器611所儲存的資料進行比較。於是，經過一個時脈週期後，第1級的排序單元61的暫存器611所儲存的資料會更新為對應於”AATT”的子編碼序列，第2級的排序單元61的暫存器611所儲存的資料被更新為對應於”ACAA”的子編碼序列，第3級的排序單元61的暫存器611所儲存的資料會更新為對應於”CAAT”的子編碼而其他排序單元61的暫存器611所儲存的資料保持不變(即，仍為對應於”TTTT”的參考子編碼序列)，如圖21所示。至此，透過將對應該短片段 ”ACAATT”的所有子編碼序列均儲存於對應的排序單元61中而建立出與”ACAATT”有關的所有4-mer的德布魯因建表。

【0044】 在該短片段”ACAATT”的德布魯因建表建立之後，若

後續有需要重組出對應於該短片段”ACAATT”的編碼序列時，如圖22所示，該多工排序引擎6可使每一排序單元61的該第一資料輸入端data_in接收對應於”ACA”(可視為第一個3-mer)的子編碼字串，此外，不同於圖15，該多工排序引擎6將使每一排序單元61的第一2×1多工器613和該3×1多工器不運作，並且該比較器612僅將該暫存器61所儲存之子編碼序列對應前3個字符的部分與接收的子編碼字串進行比較，於是，僅第2級的排序單元61的第四輸出端target會輸出邏輯1的信號，而第1、3級的排序單元61的第四輸出端target會輸出邏輯0信號，因此將第2級的排序單元6的暫存器61所儲存的對應於”ACAA”的子編碼序列被輸出作為與該短片段”ACAATT”有關的一編碼序列。接著，如圖23所示，該多工排序引擎6會使每一排序單元61的該第一資料輸入端data_in接收對應於”ACAA”的後3個字符，即”CAA”(可視為第二個3-mer)的子編碼字串，於是，僅第3級的排序單元61的第四輸出端target會輸出邏輯1的信號，而第1、2級的排序單元61的第四輸出端target會輸出邏輯0信號，因此將第3級的排序單元6的暫存器61所儲存的對應於”CAAT”的子編碼序列被輸出，並根據輸出的子編碼序列來擴展該編碼序列，亦即從”ACAA”擴展為”ACAAT”。然後，如圖24所示，該多工排序引擎6會使每一排序單元61的該第一資料輸入端data_in接收對應於”CAAT”的後3個字符，即”AAT”(可視為第三

個3-mer)的子編碼字串，於是，僅第1級的排序單元61的第四輸出端target會輸出邏輯1的信號，而第2、3級的排序單元61的第四輸出端target會輸出邏輯0信號，因此將第1級的排序單元6的暫存器61所儲存的對應於”AATT”的子編碼序列被輸出，並根據輸出的子編碼序列來進一步擴展該編碼序列，亦即從”ACAAT”擴展為”ACAATT”，如此便獲得了有關於該短片段的重組編碼序列。

【0045】 在依照如以上示例的方式完成該參考DNA序列以及所有短片段的德布魯因建表後，該多工排序引擎6將進行以下操作以重組出有關於該待測DNA序列的一個或多個編碼序列組合。

【0046】 首先，該多工排序引擎6使每一排序單元61的該第一資料輸入端data_in先接收與該等短片段其中一個具有最小回貼位置的短片段的前k個字符(可稱之為k-mer)對應的子編碼字串，根據在該等排序單元61的第四輸出端的輸出結果(邏輯0或邏輯1之信號)來決定要被輸出的子編碼序列(亦即，將輸出邏輯1之信號的排序單元61中的暫存器61所儲存的子編碼序列輸出)並將其作為與該待測DNA序列有關的一編碼序列，然後在每一排序單元61的該第一資料輸入端data_in再一次接收前一次輸出的子編碼序列中與其對應的(k+1)個字符中的後k個字符(即，下一個k-mer)所對應的子編碼字串，以便據以決定本次要輸出的子編碼序列，並根據本次輸出的子編碼序列擴展該編碼序列，並重複執行上述操作直到獲得該(等)

編碼序列組合。該多工排序引擎6還將該(等)編碼序列組合儲存於該儲存模組1。在實際使用時，只需將每一編碼序列組合透過對應於編碼方式解碼後即可獲得一對應的半倍體序列。

【0047】 以下，將參閱圖25進一步示例性地詳細說明該多工排序引擎6如何重組一個編碼序列組合。在此示例中，圖25繪示出該參考DNA序列、及對應於不同回貼位置的該等短片段(以下簡稱為Read 1、Read 2、Read 3、Read 4及Read 5來表示)，其中該等短片段的以回貼位置從小到大的排列順序為Read 3→Read 4→Read 1→Read 2→Read 5。該多工排序引擎6可利用如圖22~圖24所描述的方式先從Read 3開始重組，接著完成Read 4的重組時可獲得如圖25所示的序列。請注意，由於Read 4出現有例如單點突變(Single Nucleotide Polymorphism，以下簡稱SNP)所導致的變體 (即，如加畫有陰影之位置所指示)，因此圖25所示的序列僅代表在重組過程中的一個部份的序列。此外，Read 4及Read 5各自亦出現有如SNP變體(即，如加畫有陰影之處所指示)。於是，當繼續完成Read 1、Read 2和Read 5的重組後，應可獲得相關於該待測DNA序列的多個半倍體序列(圖未示出)。

【0048】 在獲得所有半倍體序列之後，該資料處理系統100可操作在該變體識別(Variant Calling)模式，以識別出每一半倍體序列中出現有變體的位置並且推估出每一變體所述的突變類型。

【0049】 在該變體識別模式下，首先，該動態編程處理引擎10操作來執行該參考DNA序列和每一半倍體序列的相似度演算，以產生對應於該半倍體序列的一相似度分數矩陣表、及一與分數來源方向有關的方向矩陣表。更具體地，對於每一半倍體序列，該動態編程處理引擎10利用動態編程將該參考DNA序列與該半倍體序列進行字符比對，並根據對應於該半倍體序列的編碼序列組合、該參考編碼字串和字符比對結果執行作為該相似度演算的Smith-Waterman演算(如上式1所示)。同樣地，該半倍體序列和該參考DNA序列的相似度可以一個二維矩陣的形式來表示，此二維矩陣的每一元素(element)可以存放一代表相似度的分數(分數越高代表相似程度越高，分數越低代表相似程度越低)，每一元素的分數都是根據字符比對結果以及在其上方、左方或左上的元素的分數並透過上述式1的演算而獲得。在式1的演算中，相似地，T1=T2=T3=0，且當比對的字符相同時，S=Sm(其為一大於零的正整數，例如，5)，而當比對的字符不同時，S=Sp(其為一小於零的負整數，例如，-2)。分數的計算是從矩陣的左上角的元素開始，並往右下方向逐層進行，直到整個矩陣內的元素的分數都計算出。如此，不僅可獲得該半倍體序列和該參考DNA序列的該相似度分數矩陣表，此外，還獲得在Smith-Waterman演算過程中紀錄了每一元素之分數的分數來源方向的該方向矩陣表。該動態編程處理引擎10將獲得的對應於

每一半倍體序列的該相似度分數矩陣表和該方向矩陣表儲存於該緩衝器102(見圖10)中。

【0050】 以下，將參閱圖26來示例地詳細說明該動態編程處理引擎10如何獲得該相似度分數矩陣表和該方向矩陣表。在此示例中，該參考DNA序列(以a來表示)例如為”GTACGT”，而該半倍體序列(以b來表示)例如為”GTAATC”。請注意，為了方便說明，所以此示例中的該參考DNA序列a和該半倍體序列的長度相當短，然而在實際使用時，二者的長度須配合該緩衝器102所配置規格，例如為300個字符長度。於是，經過動態比對該參考DNA序列a與該半倍體序列b的每一字符以及Smith-Waterman演算後所獲得的相似度分數矩陣表和方向矩陣表係分別顯示於圖26中的左表和右表。例如，當比對該參考DNA序列a的第一個字符”G”與該半倍體序列b的第一個字符”G”時，由於二者相同，所以在該相似度分數矩陣表的左上角的元素的分數為5(=0+5)，且在該方向矩陣表中的對應元素的分數來源方向是以符號”↘”來表示；當比對該參考DNA序列a的第二個字符”T”與該半倍體序列b的第一個字符”G”時，由於二者不同，所以該相似度分數矩陣表的第一列(row)中的第二個元素的分數為3(=5-2)，且在該方向矩陣表中的對應元素的分數來源方向是以符號”→”來表示；當比對該參考DNA序列a的第一個字符”G”與該半倍體序列b的第二個字符”T”時，由於二者不同，所以該相似

度分數矩陣表的第一行(column)中的第二個元素的分數亦為3(=5-2)，而在該方向矩陣表中的對應元素的分數來源方向是以符號”↓”；同理，可獲得如圖26所示的整個相似度分數矩陣表和整個方向矩陣表。請注意，使用符號”↘”， ”→”， ”↓”僅是為了方便說明，而實際上在該緩衝器102中所儲存的該方向矩陣表的資料內容是以不同的編碼來代表前述不同符號所代表方向。

【0051】 然後，對於每一半倍體序列而言，該變體識別模組12根據由該動態編程處理引擎10提供該緩衝器102(見圖10)儲存對應於該半倍體序列的該相似度分數矩陣表和該方向矩陣表，從該相似度分數矩陣中確認在該相似度分數矩陣表中出現最高分數的位置，然後從該方向矩陣表獲得達到該位置的方向軌跡，且至少根據該方向軌跡識別出存在於該半倍體序列中的每一變體的位置並推估出每一變體所屬的突變類型。具體而言，當該方向軌跡含有符號”→”時，則該變體識別模組12會識別出該符號”→”所在位置即為對應變體的位置並推估出該對應變體所屬的突變類型為刪除突變(Deletion Mutation，以下簡稱DM)，於是，該變體識別模組12還可對於具有DM之變體的半倍體序列以一特定形式進行校正；當該方向軌跡含有符號”↓”時，則該變體識別模組12會識別出該符號”↓”所在位置即為對應變體的位置並推估出該對應變體所屬的突變類型為插入突變(Insertion Mutation，以下簡稱IM)；而在該方向

軌跡全由符號”↘”所組成(即，不含有”→”且亦不含有”↓”)的情況下，該變體識別模組12可進一步根據該該相似度分數矩陣中從對應於該方向軌跡之分數中辨識出有比前一個分數更小的分數之位置即為對應變體的位置並推估出該對應變體所屬的突變類型為SNP。舉例來說，若根據圖26的示例情況，該相似度分數矩陣表中出現最高分數(即，23)的位置在第5列(row)中的最後(右)一個元素的位置(即，加畫有陰影的位置)，並且從該方向矩陣表所獲得的方向軌跡是由表中的粗黑色的(方向)箭頭符號所組成。由於從此方向軌跡往回搜尋可知在該參考DNA序列a的第4個字符(含氮鹼基)的位置出現有有符號”→”，此代表該半倍體序列b在第4個字符的位置出現有歸屬於刪除突變的變體(也就是說，推估出該待測DNA序列在第4個字符的位置發生了DM的基因變異)，於是，該變體識別模組12可進一步將該半倍體序列b(即，”GTAATC”)校正成”GTA-AT”以供後續輸出之用。

【0052】 此外，在該變體識別模式下，對於該待測DNA序列發生的每一變體，該動態編程處理引擎10還可操作來根據含有有該變體之位置的一個或多個相關短片段、具有該變體的一半倍體序列和該參考DNA序列(即，無變體的半倍體序列)，進行該變體導因於SNP、IM或DM的可能性(Likelihood)演算，以獲得對於該變體的一包含有該半倍體序列與該參考DNA序列其每一者相對於該(等)

相關短片段各自的可能性大小的矩陣結果；於是，該變體識別模組12根據該矩陣結果可進一步計算出包含該待測DNA序列的雙股DNA在該位置均沒有該變體的機率(即，待測者的雙親均無該變體的機率)、該雙股DNA在該位置均有該變體的機率(即，待測者的雙親均有該變體的機率)，以及該雙股DNA其中一者在該位置有該變體的機率(即，待測者的雙親其中一方有該變體的機率)。

【0053】 更明確地，根據如圖27所示有關SNP、IM和DM的已知生物模型，可定義出以下式7~式9：

$$V_S(i,j) = P(x_i, y_j) \cdot max \begin{cases} (1-2\delta) \cdot V_S(i-1, j-1) \\ (1-\varepsilon) \cdot V_I(i-1, j-1) \\ (1-\varepsilon) \cdot V_D(i-1, j-1) \end{cases} \quad (式7)$$

$$V_I(i,j) = P(x_i, \eta) \cdot max \begin{cases} \delta \cdot V_S(i-1, j) \\ \varepsilon \cdot V_D(i-1, j) \end{cases} \quad (式8)$$

$$V_D(i,j) = P(\eta, y_j) \cdot max \begin{cases} \delta \cdot V_S(i, j-1) \\ \varepsilon \cdot V_D(i, j-1) \end{cases} \quad (式9)$$

其中$P(x_i, y_j)$代表x序列相對於y序列發生SNP的可能性大小(即，x序列的第i個字符與y序列的第j個字符相符的可能性大小，$P(x_i, \eta)$代表x序列相對於y序列發生IM的可能性大小(即，x序列第i個字符對應到y序列的空位(empty base)的可能性大小)，$P(\eta, y_j)$代表x序列相對於y序列發生 DM的可能性大小(即，y序列第j個字母對應到x序列的空位的可能性大小)，$V_S(i,j)$代表x序列的第i個字符相對於y序列的第j個字符發生SNP的可能性大小，$V_I(i,j)$代表x序列的第i個字符相對於y序列的第j個字符發生IM的可能性大小，$V_D(i,j)$代表x序

列的第i個字符相對於y序列的第j個字符發生DM的可能性大小，且δ與ε均為預定參數。於是，將式7~式9取對數後分別可獲得以下式10~式12：

$$v_S(i,j) = p(x_i, y_j) + max\begin{cases} log(1-2\delta) + v_S(i-1, j-1) \\ log(1-\varepsilon) + v_I(i-1, j-1) \\ log(1-\varepsilon) + v_D(i-1, j-1) \end{cases} \quad (式10)$$

$$v_I(i,j) = p(x_i, \eta) + max\begin{cases} \log(\delta) + v_S(i-1, j) \\ \log(\varepsilon) + v_D(i-1, j) \end{cases} \quad (式11)$$

$$v_D(i,j) = p(\eta, y_j) + max\begin{cases} \log(\delta) + v_S(i, j-1) \\ \log(\varepsilon) + v_D(i, j-1) \end{cases} \quad (式12)$$

於是，當該動態編程處理引擎10操作來對於每一變體且根據相關短片段與相關半倍體序列分別進行SNP、IM和DM的可能性演算時，每一運算單元101可操作成如圖28所示且分別對應於SNP、IM、DM的等效電路，將此等效電路所輸出之每一值作為以10為底數的冪數即可獲得對應於該值的可能性大小。如此，對於該變體所演算出的SNP、IM和DM可能性結果可分別以V_S、V_I和V_D來代表，且各自具有呈矩陣排列的多個可能性大小之值，並從V_S、V_I和V_D其中在最後一列(row)出現有最大值的一者代表該變體所屬的突變類型且該最大值作為該相關半倍體序列相對於該相關短片段的可能性大小(此僅為對應於該變體之矩陣結果其中一個元素)，並重複上述運算操作直到完成對應於該變體的整個矩陣結果。

【0054】 最後，該變體識別模組12可將對應於該待測DNA序列且含有辨識出的所有變體各自的位置、推估出所有變體各自的突變

類型以及計算出對應於所有變體各自的相關機率的資訊紀錄作為完整的變異識別結果且以一合適的標準格式之紀錄檔案之形式向外輸出，以供相關人員運用和參考。特別一提的是，相關人員可根據此紀錄檔案中對應於每一變體的相關機率來進一步確認(辨識出的)該變體是基於實際發生突變所產生的真實變體，還是基於定序處理上的誤差或失誤而產生的。

【0055】 因此，當該資料處理系統100應用於人體三十億個含氮鹼基序列時，經過分段處理(例如每段的長度為300個含氮鹼基)後，再依照上述的預處理模式、短片段回貼模式、序列重組模式及變體識別模式等的操作後，完整的變異識別結果已被紀錄下來並可以一合適的標準格式輸出此紀錄檔案，以供後續如醫療院所或研究機構作之相關人員為判讀遺傳序列或潛在相關疾病的重要參考依據。此外，值得注意的是，本發明的資料處理系統100可被整合於一系統單晶片，並結合客製化的控制電路與指令傳輸電路等，能將待分析的資料直接儲存於一可攜式的紀錄媒體(例如SD卡)，在完成運算後將處理或分析結果直接儲存於該可攜式的紀錄媒體，藉此有利於相關人員的分析及資源共享。

【0056】 綜上所述，本發明的資料處理系統100確實能達成以下功效：

1. 在該預處理模式的操作中，僅使用後綴字串的前K個字

符的編碼字串作為排序的依據，此外，將後綴字串分群來排序以降低運算時間、複雜度和記憶體需求；

2. 在該可短片段回貼模式的操作中，利用FM-指標資料結構先進行小片段(Seed)的精確比對(exact match)以獲得候選位置後，再使用動態編程演算進行非精確比對(inexact match)之相似度計算來決定回貼位置；

3. 該多工排序引擎6可以支援在該預處理模式中的編碼字串分群和快速排序以及在該序列重組模式中的德布魯因建表和編碼序列重組，並且其所含的大量的平行排序單元61僅需一個電路時脈即可完成一次的運算，藉此實現大量的高速資料處理；及

4. 該動態編程處理引擎10支援該短片段回貼模式和該變體識別模式的操作，並可被設計成一維架構，藉此降低硬體複雜度並減少電路面積。

【0057】 惟以上所述者，僅為本發明之實施例而已，當不能以此限定本發明實施之範圍，凡是依本發明申請專利範圍及專利說明書內容所作之簡單的等效變化與修飾，皆仍屬本發明專利涵蓋之範圍內。

【符號說明】

【0058】

100…資料處理系統

1…儲存模組

2…後綴字串產生模組

3…字串產生模組

4…編碼模組

5…分離參考字串選擇模組

6…多工排序引擎

61…排序元件

611…暫存器

612…比較器

613…第一 2×1 多工器

614…3×1 多工器

615…第二 2×1 多工器

616…反閘

617…及閘

62…加法器

7…後綴字串矩陣產生模組

8.. FM-指標資料產生模組

9…候選位置產生模組

10…動態編程處理引擎

101,101_{11}~101_{44}…運算單元

102…緩衝器

11…回貼位置決定模組

12…變體識別模組

data_in…第一資料輸入端

data_pre…第二資料輸入端

EN_pre…第一控制輸入端

Mode…第二控制輸入端

data_out…第一輸出端

EN…第二輸出端

result…第三輸出端

target…第四輸出端

【發明申請專利範圍】

【第1項】 一種資料處理系統，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由四個分別代表四種不同含氮鹼基的字符A，C，G，T組成的(N-1)個字符之參考DNA序列以及一位在該參考DNA序列之後代表序列結束的字符$的參考序列的N個後綴字串、多個分別指示出該等N個字符在該參考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，以及多個擷取自一待測DNA序列的短片段，該資料處理系統可操作在與該參考DNA序列有關的一預處理模式，或可操作在與該待測DNA序列有關的一短片段回貼模式、一序列重組模式及一變體識別模式其中一者，並包含：

一字串產生模組；

一編碼模組，連接該字串產生模組；

一分離參考字串選擇模組；

一多工排序引擎，連接該分離參考字串選擇模組；

一後綴字串矩陣產生模組，連接該多工排序引擎；

一FM-指標資料產生模組，連接該後綴字串矩陣生模組；

一候選位置產生模組；

一動態編程處理引擎，連接該候選位置產生模組；

一回貼位置決定模組，連接該多工排序引擎和該動態編程處理引擎；及

一變體識別模組，連接該動態編程處理引擎；

其中，當該資料處理系統操作在該預處理模式時，該字串產生模組擷取該等N個後綴字串其中的每一者的前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，

該編碼模組利用一將該等字符$，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，將該等N個後綴字串編碼以產生N個分別對應於該等N個指標且具有一數字碼形式的編碼字串，並將該參考DNA序列和該等短片段以相同的編碼方式編碼以產生對應於該參考DNA序列的參考編碼字串和多個分別對應於該等短片段的待測編碼字串，

該分離參考字串選擇模組先以一升取樣方式從該等N個編碼字串選出P×Q個編碼字串提供給該多工排序引擎其中P代表分離參考字串的數量且Q代表取樣倍數，以使該多工排序引擎依照編碼值將該P×Q個編碼字串排序，然後以一降取樣方式從該排序的P×Q個編碼字串選出P個依照編碼值從小到大排列的編碼字串分別作為第一至第P分離參考字串，

該多工排序引擎操作來根據該分離參考字串選擇模組選出的該第一至第P分離參考字串將該編碼模組產生的該N個編碼字串分成(P+1)群、並將該(P+1)群其中每一群的編碼字串依照編碼值從小到大排序，以獲得該N個編碼字串依照編碼值從小到大的排序結果，

該後綴字串矩陣產生模組根據來自該多工排序

引擎的該排序結果，產生一對應於該參考序列的後綴字串矩陣，及

該FM-指標資料產生模組根據來自該後綴字串矩陣產生模組的該後綴字串矩陣及該等指標，建立一對應於該參考序列的FM-指標資料結構，其中該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矩陣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矩陣的一最後字符欄的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矩陣中第一至第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；

其中，當該資料處理系統操作在該短片段回貼模式時，該候選位置產生模組將該等短片段其中每一者分割成多個小片段，然後根據該FM-指標資料產生模組產生的該FM-指標資料結構，對於每一小片段，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，以獲得一個或多個代表該小片段在該待測DNA序列中的候選位置的指標，

該動態編程處理引擎操作來根據來自該候選位置產生模組對於每一短片段的該等小片段所獲得的所有指標，執行每一短片段與該參考DNA序列中在每一候選位

置擷取的對應參考片段的相似度演算，以獲得對應於該候選位置的相似度分數，及

該回貼位置決定模組將根據該動態編程處理引擎對於每一短片段所獲得的所有相似度分數中的最高者對應的指標所代表的候選位置決定為該短片段的回貼位置；

其中，當該資料處理系統操作在該序列重組模式時，該多工排序引擎操作來根據與該等短片段對應的回貼位置以及該編碼模組產生的該參考編碼字串和該等待測編碼字串，重組出有關於該待測DNA序列的一個或多個編碼序列組合，該(等)編碼序列組合各自代表一對應的半倍體序列；及

其中，當該資料處理系統操作在該變體識別模式時，該動態編程處理引擎操作來執行該參考DNA序列和每一半倍體序列的相似度演算，以產生對應於該半倍體序列的一相似度分數矩陣表、及一與分數來源方向有關的方向矩陣表，及

對於每一半倍體序列，該變體識別模組根據該動態編程處理引擎產生對應於該半倍體序列的該相似度分數矩陣表和該方向矩陣表，從該相似度分數矩陣表確認其中出現最高分數的位置，然後從該方向矩陣表獲得達到該位置的方向軌跡，且至少根據該方向軌跡識別出存在於該半倍體序列中的每一變體的位置並推估出對應於每一變體的突變類型。

【第2項】 如請求項1所述的資料處理系統，還包含：

一儲存模組，連接該分離參考字串選擇模組、該編碼模組、該多工排序引擎和該動態編程處理引擎，且用來儲存該參考DNA序列和該等指標、該等短片段、該分離參考字串選擇模組選出的該第一至第P分離參考字串、該編碼模組產生的該N個編碼字串、該等待測編碼字串和該參考編碼字串，以及該多工排序引擎重組出的該(等)編碼序列組合。

【第3項】 如請求項2所述的資料處理系統，其中，當該資料處理系統操作在該預處理模式時，該多工排序引擎根據讀取自該儲存模組儲存的該第一至第P分離參考字串及該N個編碼字串獲得對應於該(P+1)群的分群結果且將該分群結果儲存於該儲存模組，然後根據讀取自該儲存模組儲存的該分群結果獲得該排序結果。

【第4項】 如請求項2所述的資料處理系統，還包含：

一後綴字串產生模組，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，該後綴字串產生模組還將該等後綴字串及其所對應的該等指標輸出至該字串產生模組。

【第5項】 如請求項2所述的資料處理系統，其中：

該FM-指標資料產生模組還連接該儲存模組，並將該

FM-指標資料結構完整地儲存於該儲存模組；及

該候選位置產生模組連接該儲存模組，並且當該資料處理系統操作在該短片段回貼模式時讀取該儲存模組所儲存的該FM-指標資料結構中的資料。

【第6項】 如請求項2所述的資料處理系統，其中：

該FM-指標資料產生模組還連接該儲存模組，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

該候選位置產生模組連接該儲存模組，並且當該資料處理系統操作在該短片段回貼模式時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7項】 如請求項6所述的資料處理系統，其中，該多工排序引擎包括多個彼此串接的排序單元，每一排序單元具有一用於接收來自外部的待處理資料的第一資料輸入端、一用於接收來自前一級的排序單元的輸出資料的第二資料輸入端、一用於接收來自前一級的排序單元的一第一控制信號的第一控制輸入端、一用於接收來自外部的一第二控制信號的第二控制輸入端、一用於輸出資料給下一級的排序單元的第一輸出端、一用於輸出提供給下一級的排序單元的第一控制信號的第二輸出端、一第三輸出端和一第四輸出端，並包含：

一暫存器，具有一輸入端、及一耦接該排序單元的該

第一輸出端的輸出端；

一比較器，具有一耦接該排序單元的該第一資料輸入端的第一輸入端、一耦接該暫存器的該輸出端的第二輸入端、及一耦接該排序單元的該第二輸出端和該第三輸出端的輸出端，當該第二輸入端接收的信號邏輯值大於或等於該第一輸入端接收的信號的邏輯值時，該比較器在該輸出端輸出邏輯-1的信號；

一第一2×1多工器，具有一耦接該排序單元的該第一資料輸入端的第一輸入端、一耦接該排序單元的該第二資料輸入端的第二輸入端、一耦接該排序單元的該第一控制輸入端的控制端、及一輸出端；

一3×1多工器，具有一耦接該前一級的排序單元的第一輸出端的第一輸入端、一耦接後一級的排序單元的第一輸出端的第二輸入端、一耦接該第一2×1多工器的該輸出端的第三輸入端、一作為該排序單元的該第二控制輸入端的控制端、及一輸出端；

一第二2×1多工器，具有一耦接該暫存器的該輸出端的第一輸入端、一耦接該3×1多工器的該輸出端的第二輸入端、一耦接該比較器的輸出端的控制端、及一耦接該暫存器的該輸入端的輸出端；

一反閘，具有一耦接該排序單元的該第一控制輸入端的輸入端、及一輸出端；及

一及閘，具有一耦接該反閘的該輸出端的第一輸入端、一耦接該比較器的該輸出端的第二輸入端、及一作為

該排序單元的該第四輸出端的輸出端。

【第8項】如請求項7所述的資料處理系統，其中：

該多工排序引擎還包含一加法器，該加法器具有多個分別耦接該等排序單元的該等第三輸出端的輸入端、及一輸出端；及

當該資料處理系統操作在該預處理模式時，該多工排序引擎在執行分群處理前，使該等排序單元其中的第一至第P個排序單元的暫存器分別儲存該第一至第P分離參考字串，然後在進行分群處理時，使該第一至第P個排序單元的暫存器分別持續地儲存該第一至第P分離參考字串，以及在該第一至第P個排序單元其中每一者的該第一資料輸入端依序接收該N個編碼字串，並根據該加法器每一次在其輸出端的輸出來決定該次輸入的編碼字串所被分到的一群。

【第9項】如請求項7所述的資料處理系統，其中：

當該資料處理系統操作在該預處理模式時，該多工排序引擎在進行排序處理時，從該第一群到第(P+1)群的逐群的方式，在該等排序單元其中每一者的該第一資料輸入端依序接收待排序的每一群的編碼字串後，依照編碼值從小到大的順序逐個輸出該群的編碼字串，以獲得該N個編碼字串排序結果。

【第10項】如請求項7所述的資料處理系統，其中，當該資料處理系統操作在該序列重組模式時，該多工排序引擎進行以下操作：

使每一排序單元的該暫存器儲存一與一具有(k+1)個相同字符的片段對應且具有相對最大編碼值的參考子編碼序列；

使每一排序單元的該第一資料輸入端依序接收對應於該參考編碼序列和每一短片段的待測編碼字串的所有與連續(k+1)個字符有關的子編碼序列，以便將每一子編碼序列紀錄在該等排序單元其中一個對應的排序單元的該暫存器中，以完成與該短片段有關的德布魯因建表；

在每一排序單元的該第一資料輸入端首先接收與該等短片段其中一個具有最小回貼位置的短片段的前k個字符對應的子編碼字串，根據在該等排序單元的第四輸出端的輸出結果來決定要被輸出的子編碼序列並將其作為與該待測DNA序列有關的一編碼序列，然後在每一排序單元的該第一資料輸入端再一次接收前一次輸出的子編碼序列中與其對應的(k+1)個字符中的後k個字符所對應的子編碼字串，以便據以決定本次要輸出的子編碼序列，並根據本次輸出的子編碼序列擴展該編碼序列，並重複執行上述操作直到獲得有關於該待測DNA序列的該(等)編碼序列組合；

該多工排序引擎還將有關於該待測DNA序列的該(等)編碼序列組合儲存於該儲存模組。

【第11項】如請求項2所述的資料處理系統，其中：

該動態編程處理引擎包含多個大致呈矩陣排列的運算單元，每一運算單元是一Smith-Waterman運算單元並

包含三個信號輸入端、及一個輸出端，其中該等輸入端分別耦接在相對於該運算單元的上方、左方及左上方之運算單元的輸出端；

當該資料處理系統操作在該度片段回貼模式時，該動態編程處理引擎根據每一短片段和該參考DNA序列中在與分割自該短片段的每一小片段對應的每一候選位置所擷取的對應參考片段的字符比對結果執行作為該相似度演算的Smith-Waterman演算，以獲得該短片段對應於該候選位置的相似度分數矩陣表，該相似度分數矩陣表中的最高相似度分數作為對應於該候選位置的相似度分數；及

當該資料處理系統操作在該變體識別模式時，該動態編程處理引擎中一部分的運算單元根據該參考DNA序列和每一半倍體序列中的字符比對結果執行作為該相似度演算的Smith-Waterman演算，以獲得對應於該半倍體序列的該相似度分數矩陣表，並且在Smith-Waterman演算過程中紀錄該相似度分數矩陣表中每一分數的分數來源方向以獲得對應於該半倍體序列的該方向矩陣表。

【發明圖式】

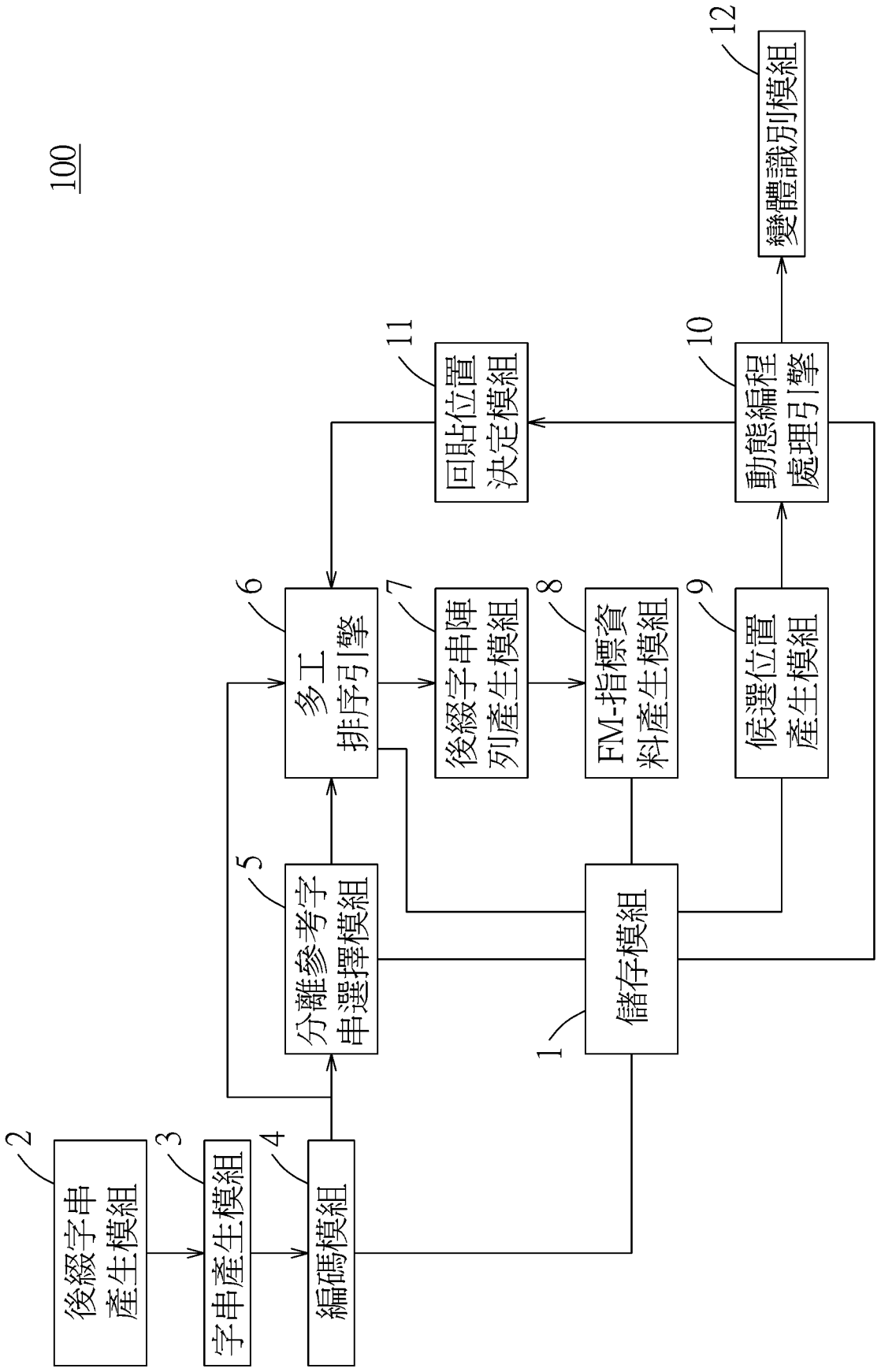
100
後綴字串產生模組
字串產生模組
編碼模組
儲存模組
分離參考字串選擇模組
多工排序引擎
後綴字串陣列產生模組
FM-指標資料產生模組
候選位置產生模組
回貼位置決定模組
動態編程處理引擎
變體識別模組
1
2
3
4
5
6
7
8
9
10
11
12

圖 1

參考序列

C	A	T	G	A	A	A	G	G	A	$

指標	後綴字串										
0	C	A	T	G	A	A	A	G	G	A	$
1	A	T	G	A	A	A	G	G	A	$	C
2	T	G	A	A	A	G	G	A	$	C	A
3	G	A	A	A	G	G	A	$	C	A	T
4	A	A	A	G	G	A	$	C	A	T	G
5	A	A	G	G	A	$	C	A	T	G	A
6	A	G	G	A	$	C	A	T	G	A	A
7	G	G	A	$	C	A	T	G	A	A	A
8	G	A	$	C	A	T	G	A	A	A	G
9	A	$	C	A	T	G	A	A	A	G	G
10	$	C	A	T	G	A	A	A	G	G	A

圖 2

指標	字串			
0	C	A	T	G
1	A	T	G	A
2	T	G	A	A
3	G	A	A	A
4	A	A	A	G
5	A	A	G	G
6	A	G	G	A
7	G	G	A	$
8	G	A	$	C
9	A	$	C	A
10	$	C	A	T

圖 3

指標	後綴字串										
10	$	C	A	T	G	A	A	A	G	G	A
9	A	$	C	A	T	G	A	A	A	G	G
4	A	A	A	G	G	A	$	C	A	T	G
5	A	A	G	G	A	$	C	A	T	G	A
6	A	G	G	A	$	C	A	T	G	A	A
1	A	T	G	A	A	A	G	G	A	$	C
0	C	A	T	G	A	A	A	G	G	A	$
8	G	A	$	C	A	T	G	A	A	A	G
3	G	A	A	A	G	G	A	$	C	A	T
7	G	G	A	$	C	A	T	G	A	A	A
2	T	G	A	A	A	G	G	A	$	C	A

圖 4

CNT表

字符	列位址
A	0
C	5
G	6
T	9

SA表

列位址	指標
0	10
1	9
2	4
3	5
4	6
5	1
6	0
7	8
8	3
9	7
10	2

F表

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

L表

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

OCC表

列位址	字符			
	A表	C表	G表	T表
0	1	0	0	0
1	1	0	1	0
2	1	0	2	0
3	2	0	2	0
4	3	0	2	0
5	3	1	2	0
6	3	1	2	0
7	3	1	3	0
8	3	1	3	1
9	4	1	3	1
10	5	1	3	1

圖5

CNT表

字符	列位址
A	0
C	5
G	6
T	9

SA表

列位址	指標
0	10
3	5
6	0
9	7

F表

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

L表

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

OCC表

列位址	字符			
	A表	C表	G表	T表
0	1	0	0	0
3	2	0	2	0
6	3	1	2	0
9	4	1	3	1

圖 6

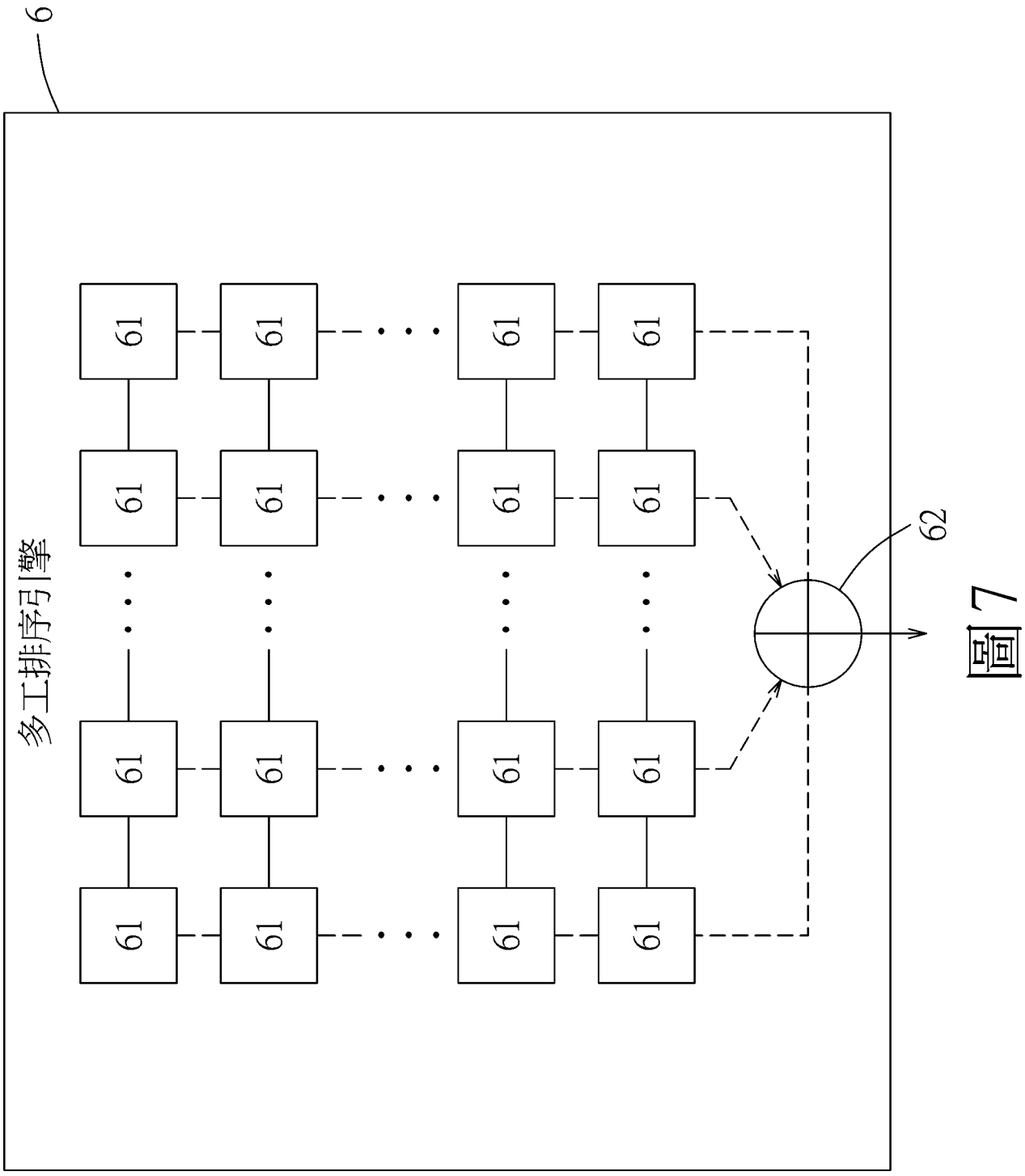

圖7

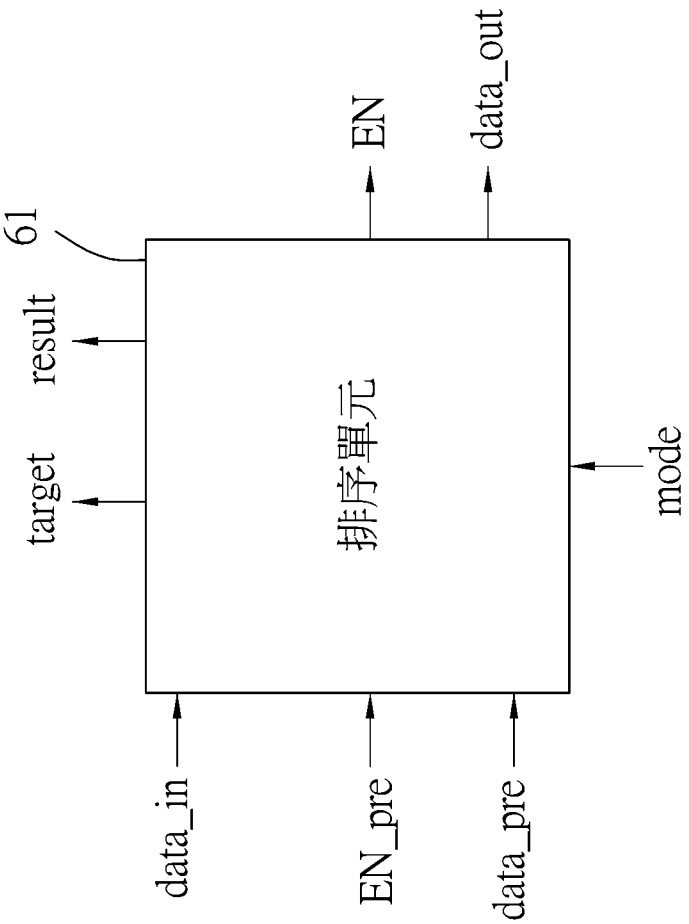

圖 8

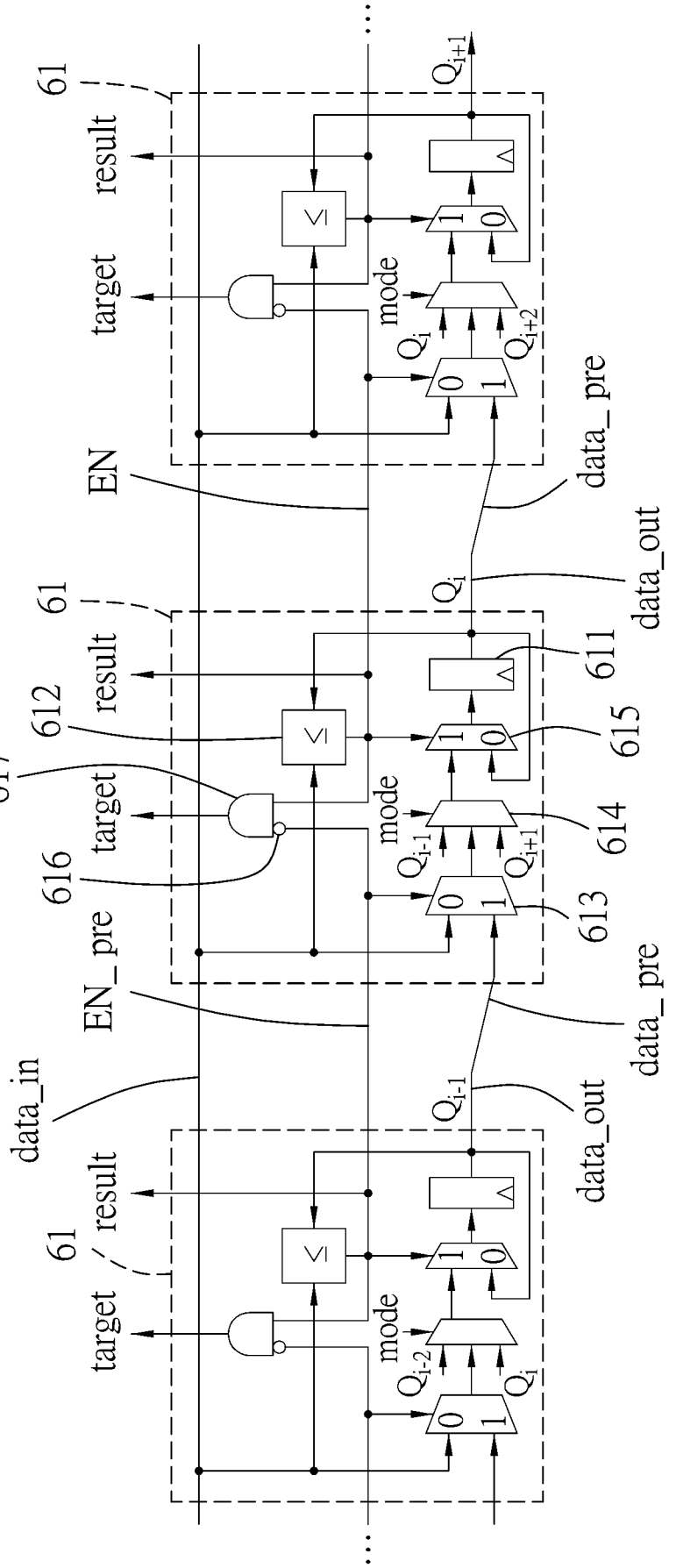
61
result
target
mode
EN
data_pre
data_out
612
617
616
EN_pre
data_in
611
615
614
613
Δ
≤
0
1

圖 9

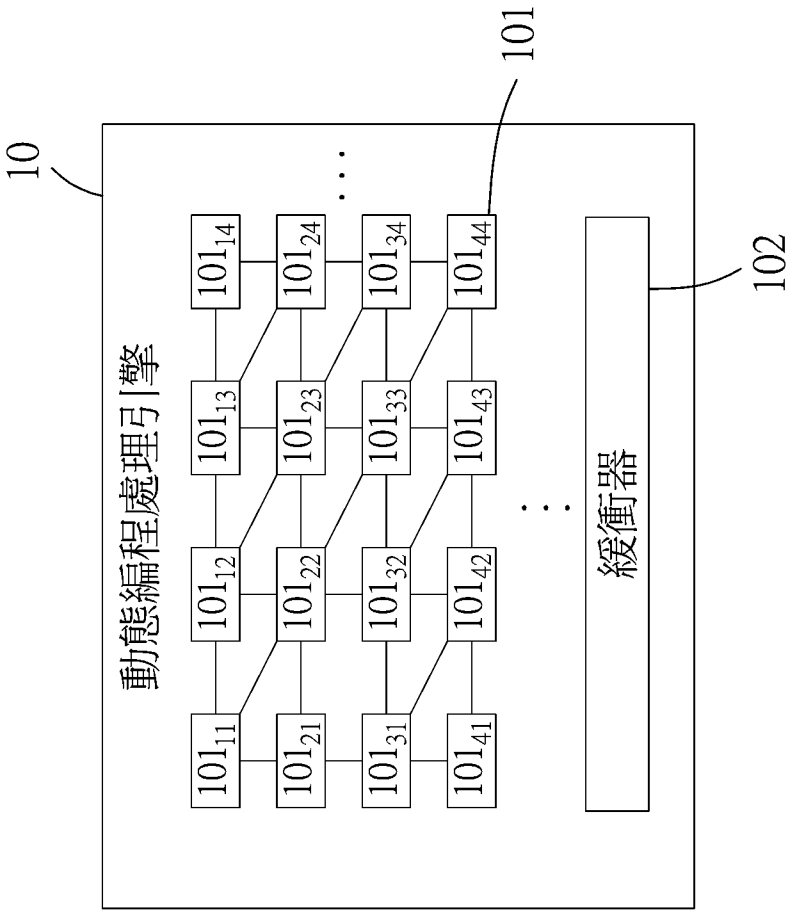
10
101
102
動態編程處理引擎
緩衝器
101_{11}
101_{12}
101_{13}
101_{14}
101_{21}
101_{22}
101_{23}
101_{24}
101_{31}
101_{32}
101_{33}
101_{34}
101_{41}
101_{42}
101_{43}
101_{44}

圖 10

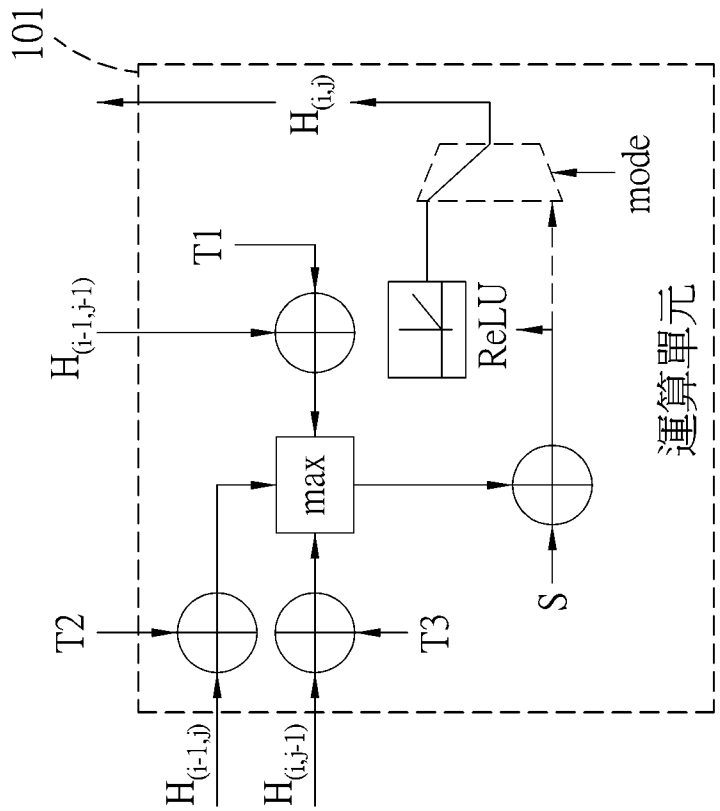

圖 11

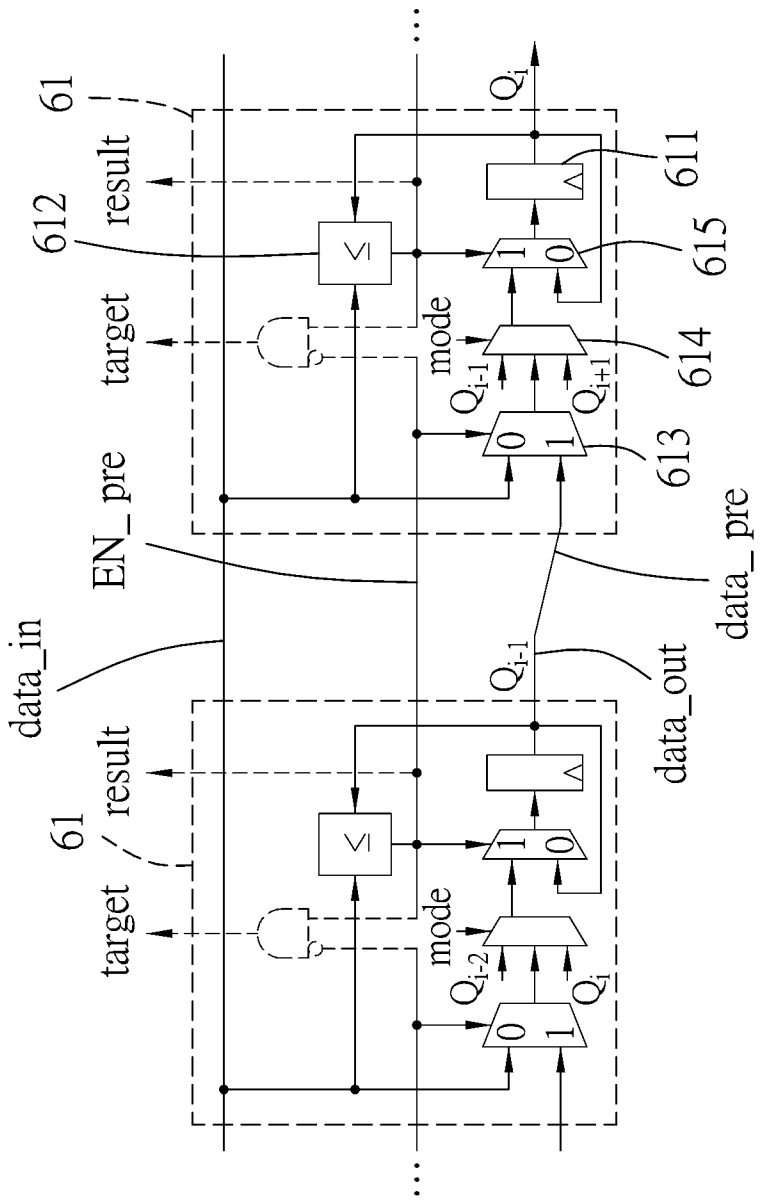
61
result
612
target
mode
EN_pre
data_in
data_pre
data_out
611
615
614
613
Q_i
Q_{i-1}
Q_{i+1}
Q_{i-2}

圖 12

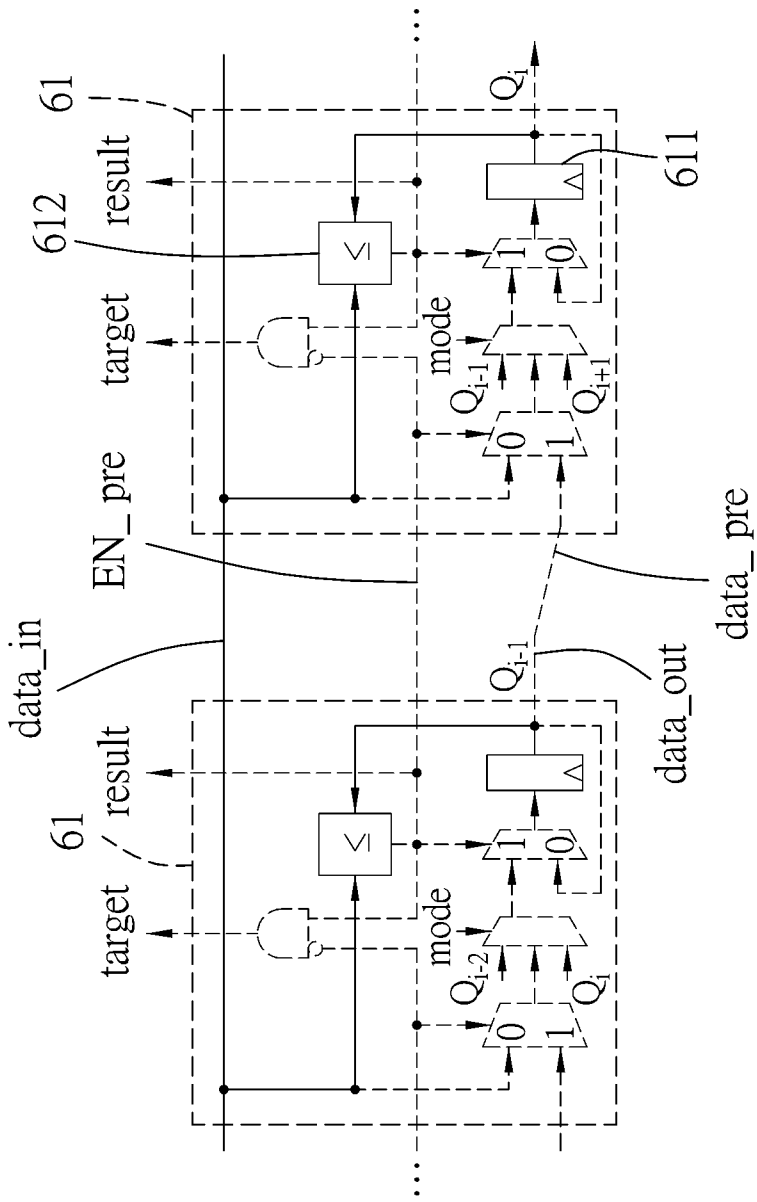
61
result
612
target
mode
EN_pre
data_in
611
data_pre
data_out
61
result
target
mode

圖 13

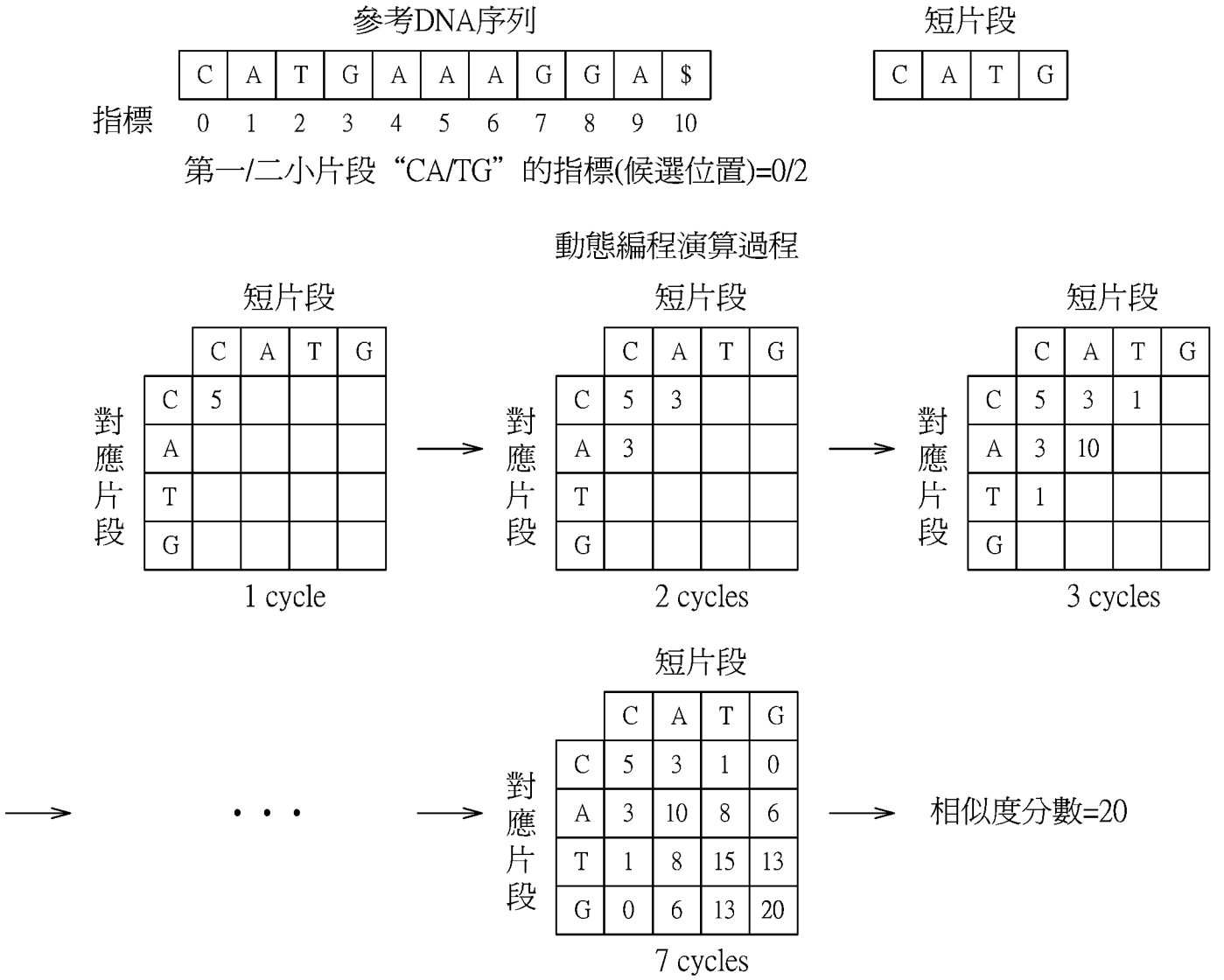
參考DNA序列
C A T G A A A G G A $
指標 0 1 2 3 4 5 6 7 8 9 10
短片段
C A T G
第一/二小片段“CA/TG”的指標(候選位置)=0/2
動態編程演算過程
短片段
對應片段
C A T G
C 5
A
T
G
1 cycle
短片段
對應片段
C A T G
C 5 3
A 3
T
G
2 cycles
短片段
對應片段
C A T G
C 5 3 1
A 3 10
T 1
G
3 cycles
短片段
對應片段
C A T G
C 5 3 1 0
A 3 10 8 6
T 1 8 15 13
G 0 6 13 20
7 cycles
相似度分數=20

圖 14

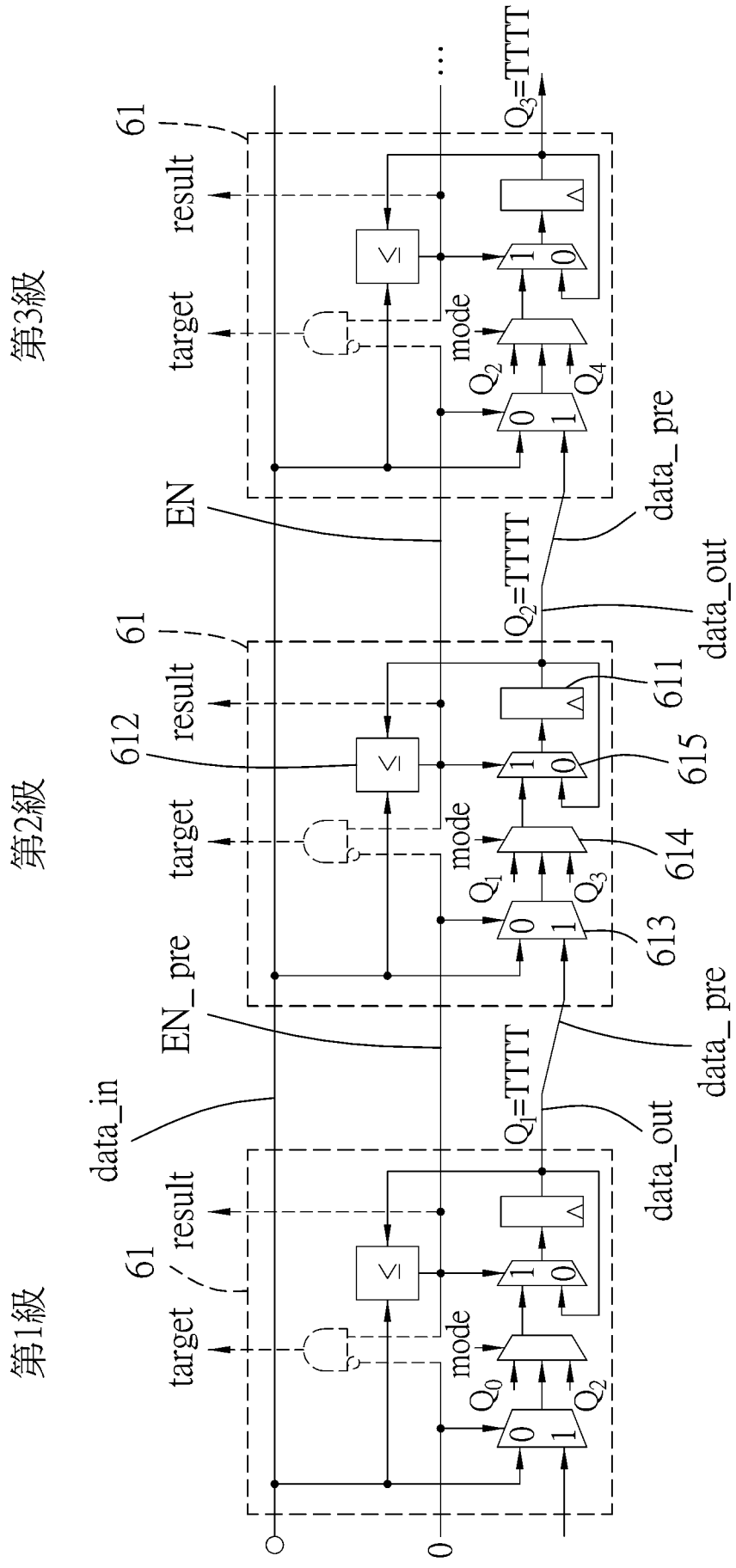
第3級
第2級
第1級
61
result
target
mode
EN
EN_pre
data_in
data_pre
data_out
data_ pre
Q_0
Q_1
Q_2
Q_3
Q_4
Q_1=TTTT
Q_2=TTTT
Q_3=TTTT
611
612
613
614
615
0

圖 15

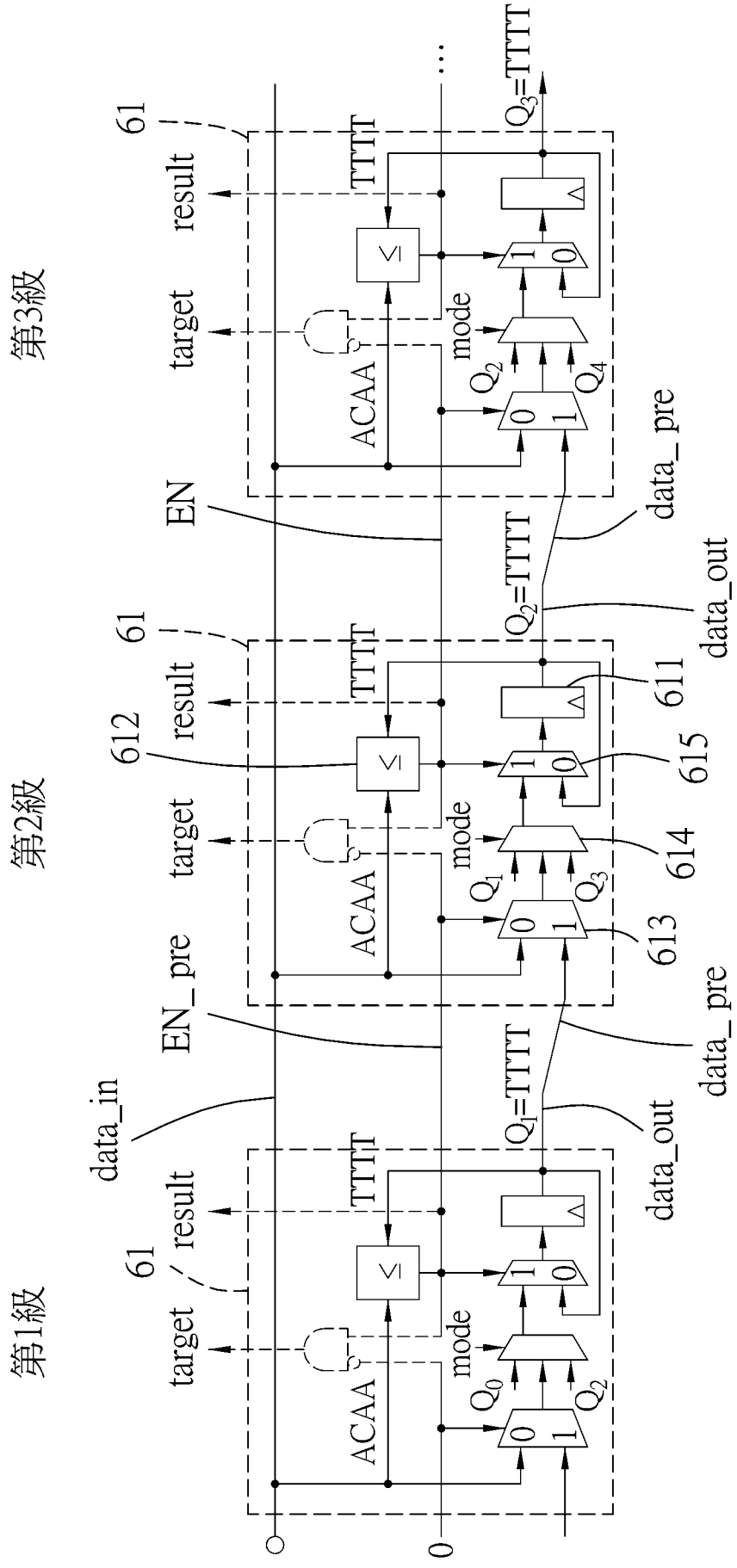
第3級
第2級
第1級
61
result
target
mode
ACAA
TTTT
EN
EN_pre
data_in
data_pre
data_out
612
611
615
614
613
Q_0
Q_1
Q_2
Q_3
Q_4
Q_1=TTTT
Q_2=TTTT
Q_3=TTTT
0

圖 16

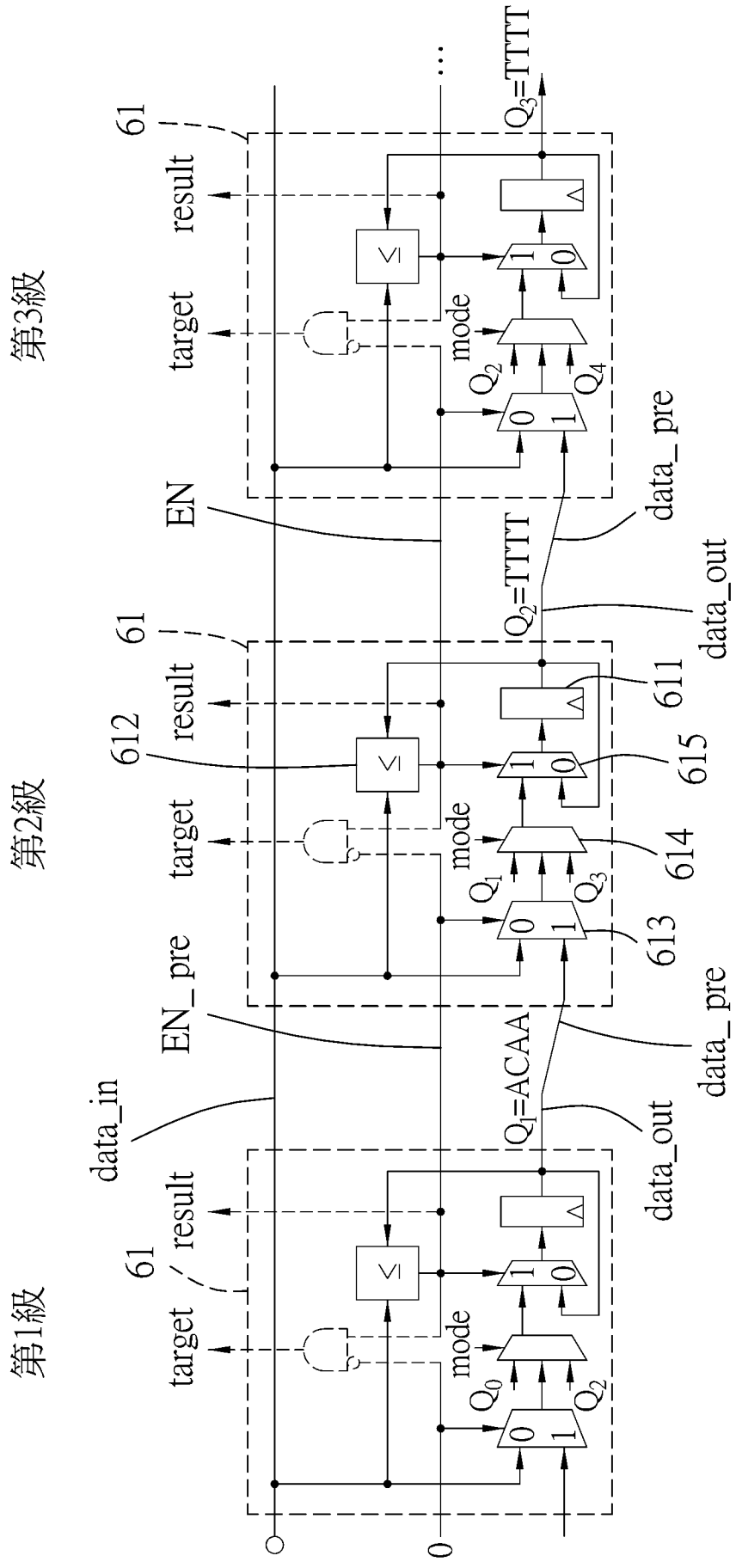
第3級
第2級
第1級
61
result
target
mode
EN
EN_pre
data_in
data_pre
data_out
data_ pre
Q_0
Q_1
Q_2
Q_3
Q_4
Q_1=ACAA
Q_2=TTTT
Q_3=TTTT
611
612
613
614
615
0
1

圖 17

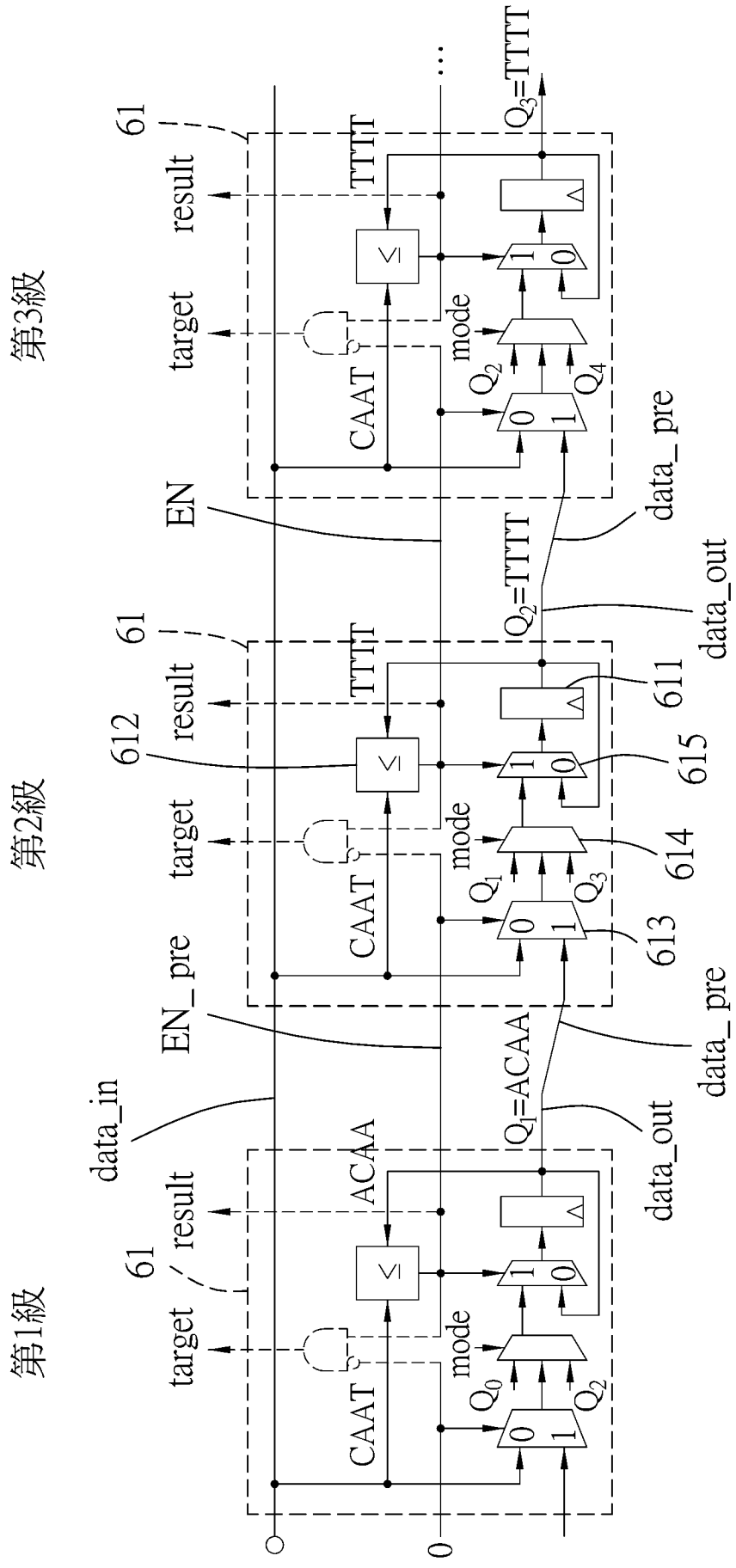
第1級
第2級
第3級
data_in
EN_pre
EN
61
result
target
mode
CAAT
ACAA
TTTT
data_pre
data_out
Q_0
Q_1
Q_2
Q_3
Q_4
Q_1=ACAA
Q_2=TTTT
Q_3=TTTT
612
611
615
614
613
0

圖 18

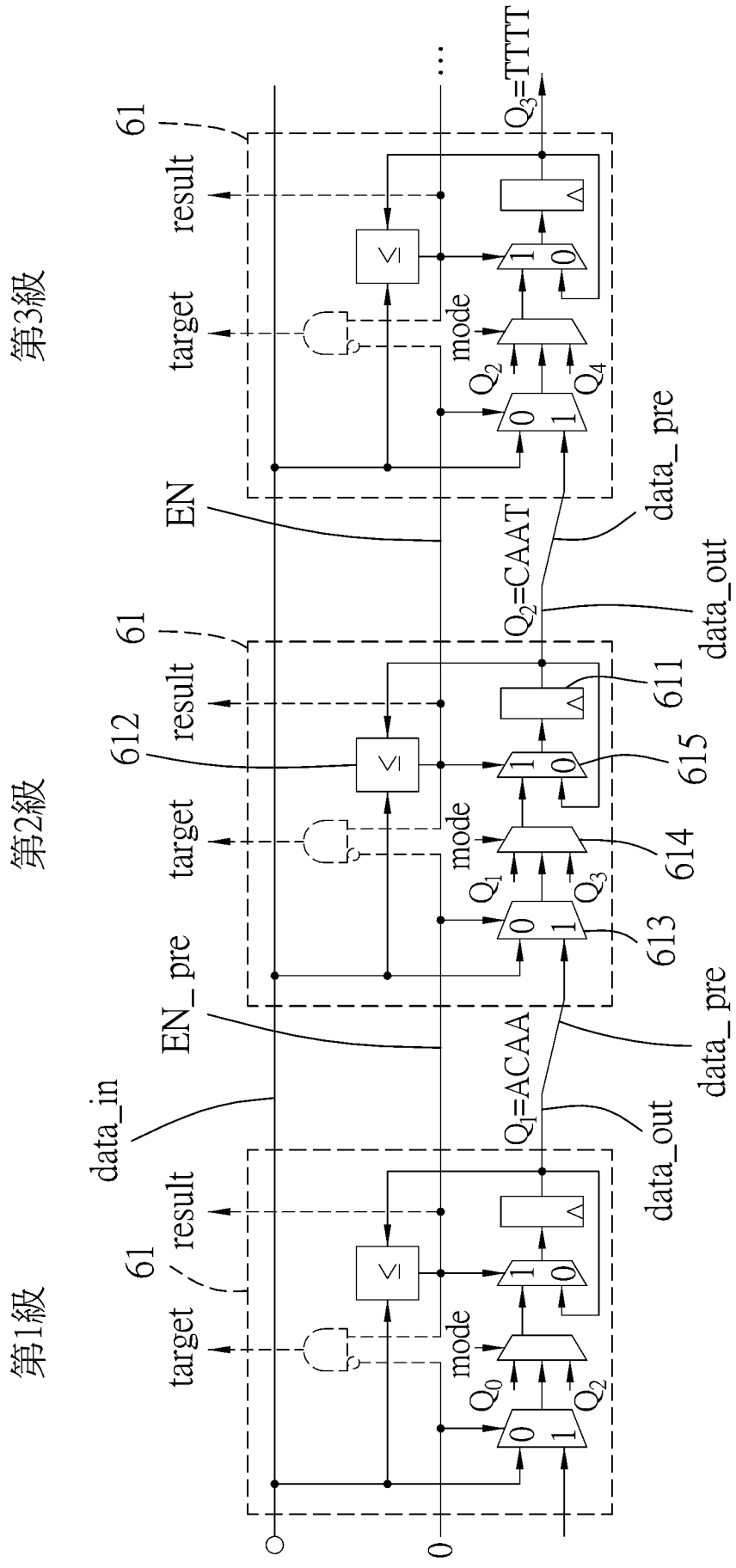
第3級
第2級
第1級
61
result
target
mode
EN
EN_pre
data_in
data_out
data_ pre
Q_0
Q_1
Q_2
Q_3
Q_4
Q_1=ACAA
Q_2=CAAT
Q_3=TTTT
611
612
613
614
615
0

圖 19

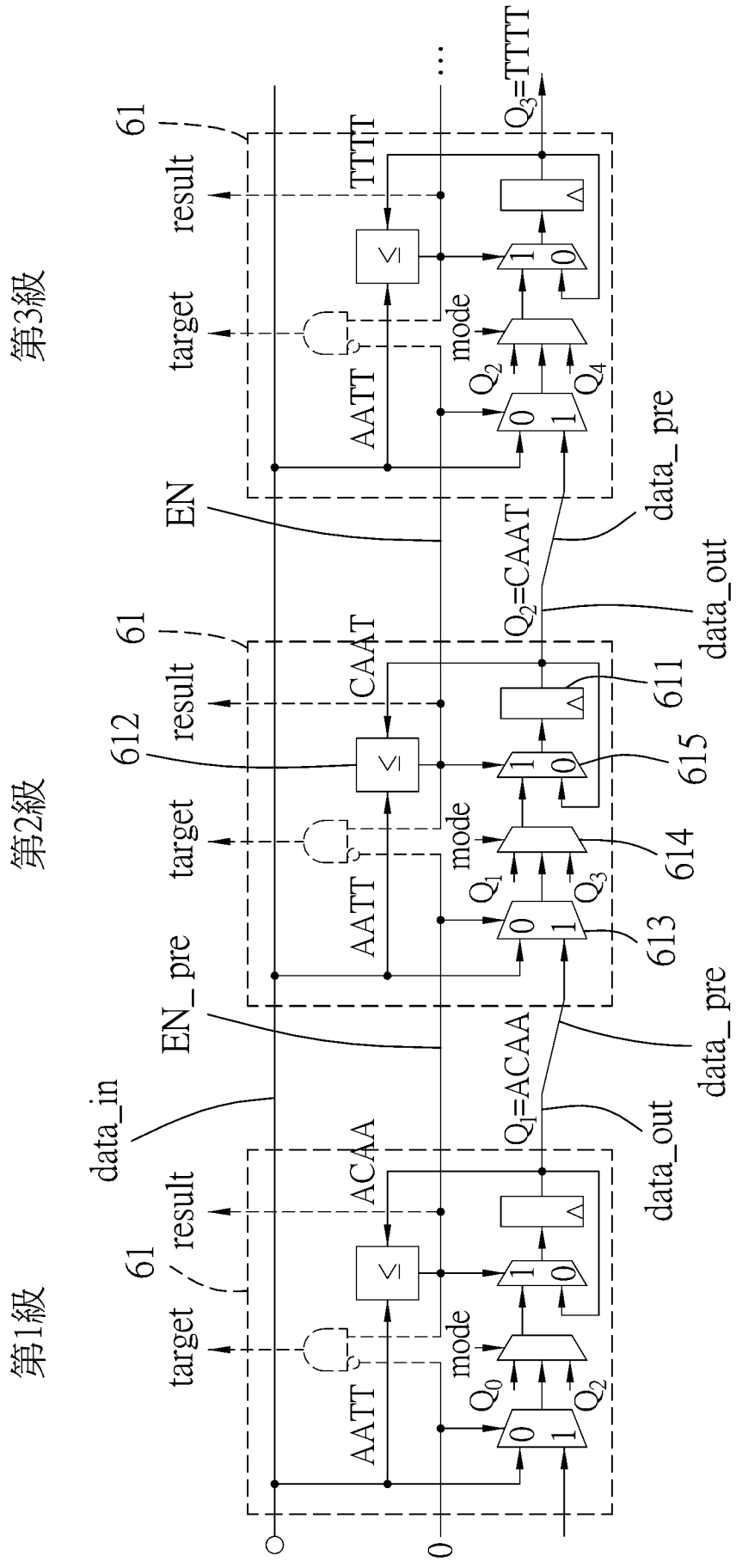
第3級
第2級
第1級
61
result
target
mode
AATT
TTTT
CAAT
ACAA
EN
EN_pre
data_in
data_pre
data_out
data_ pre
Q_3=TTTT
Q_2=CAAT
Q_1=ACAA
Q_0
Q_2
Q_1
Q_3
Q_4
612
611
615
614
613
0

圖 20

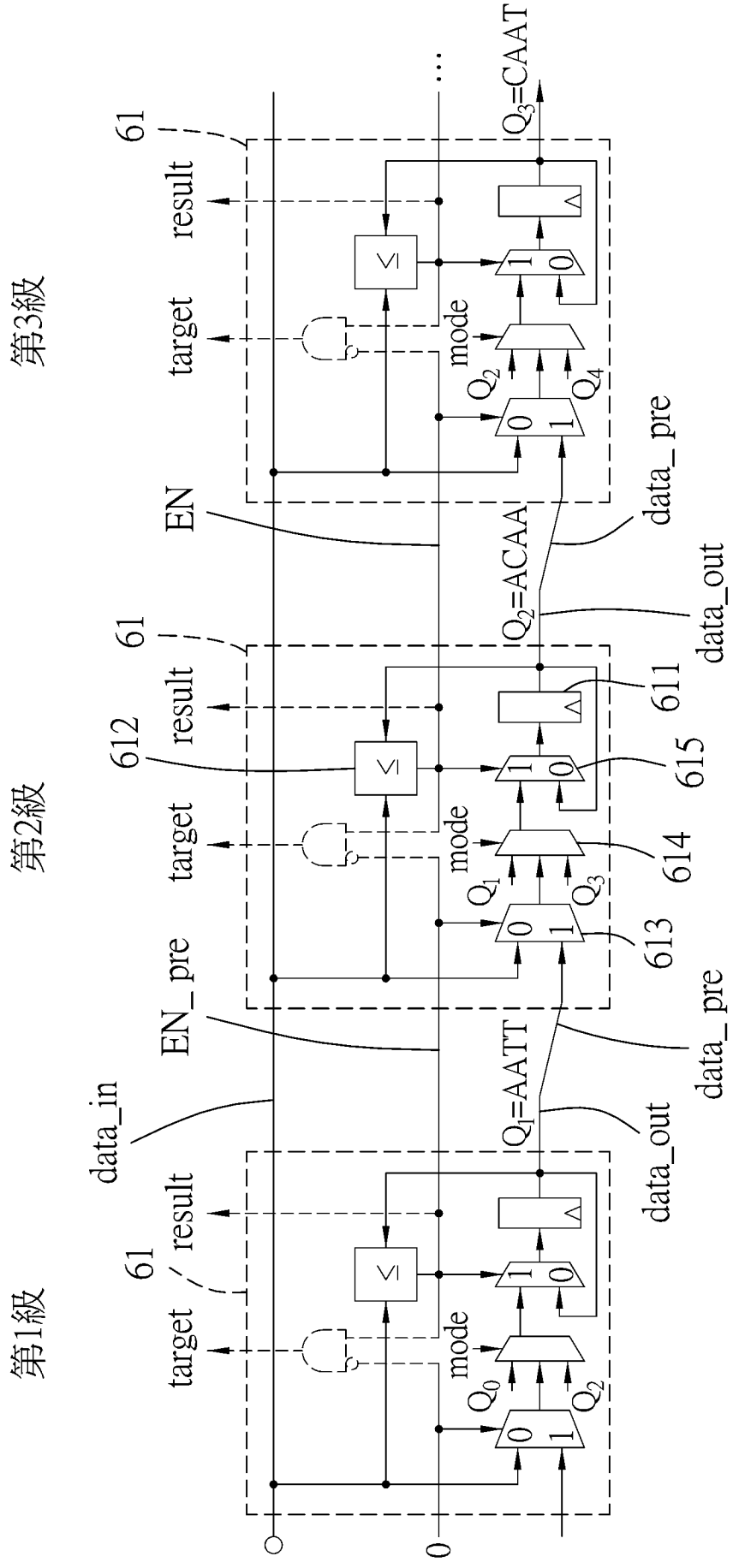
第3級
第2級
第1級
61
result
target
mode
EN
EN_pre
data_in
data_pre
data_out
data_ pre
Q_0
Q_1
Q_2
Q_3
Q_4
Q_1=AATT
Q_2=ACAA
Q_3=CAAT
611
612
613
614
615
0

圖 21

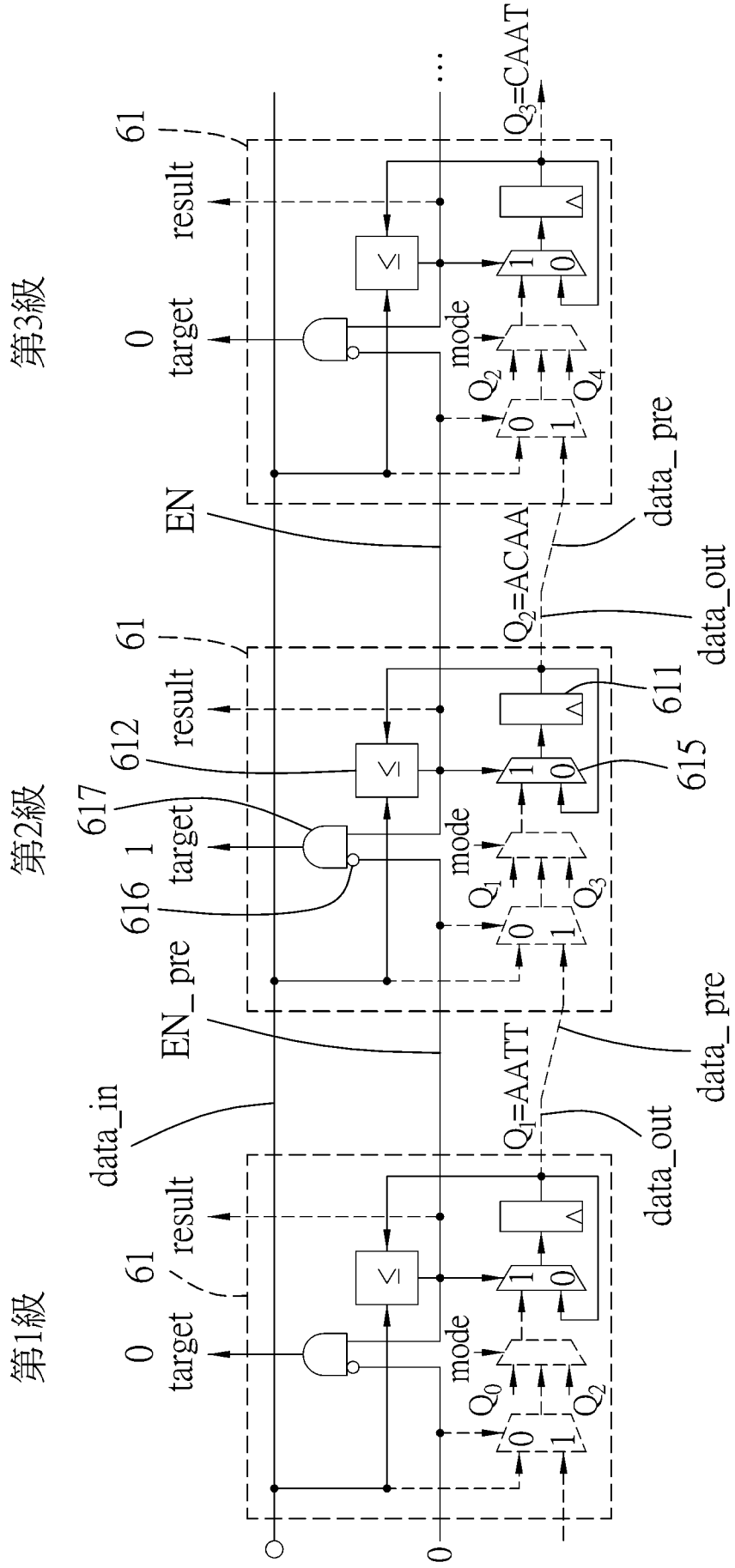
第3級
第2級
第1級
61
result
target
mode
EN
EN_pre
data_in
data_out
data_ pre
Q_3=CAAT
Q_2=ACAA
Q_1=AATT
Q_0
Q_2
Q_1
Q_3
Q_4
611
612
615
616
617

圖 22

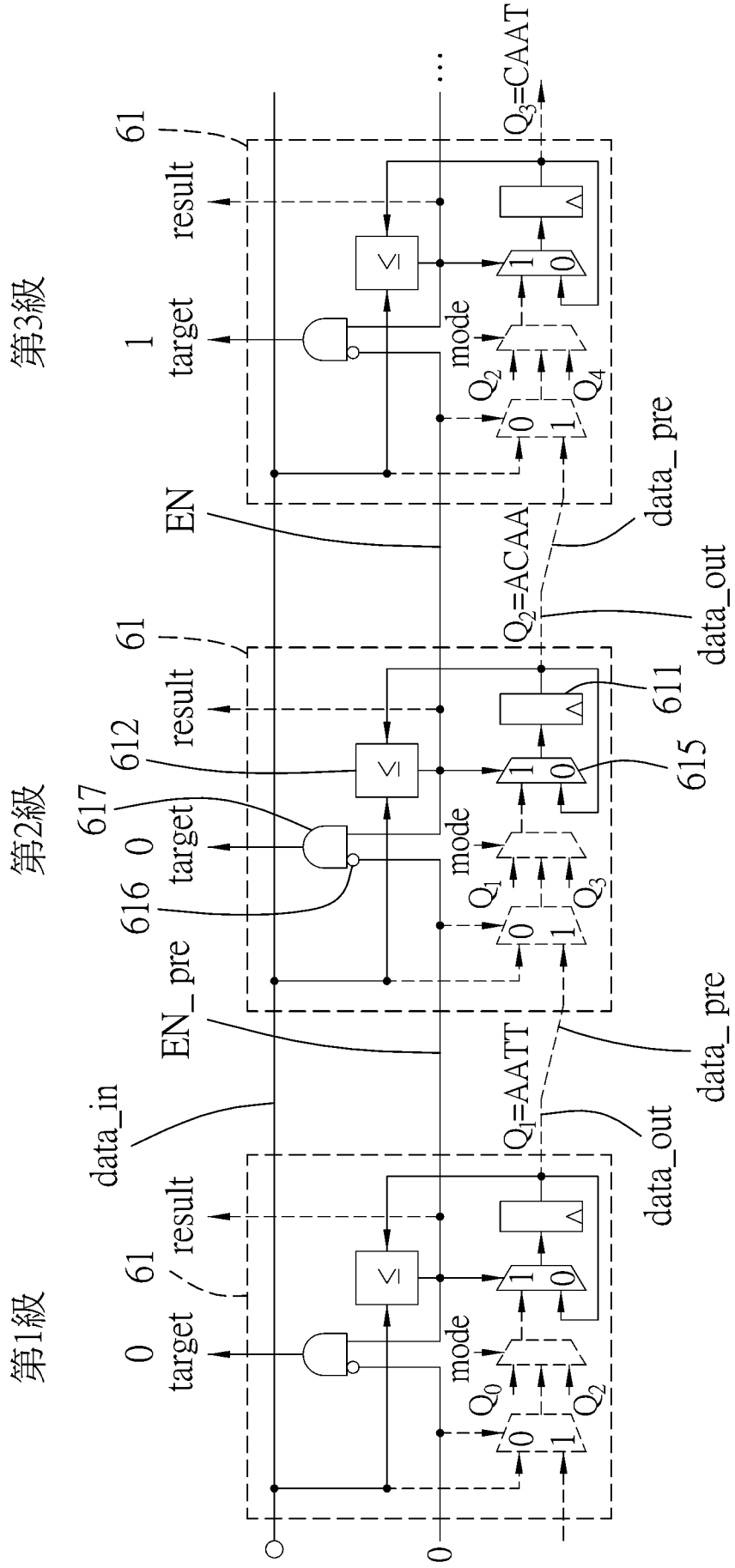
第3級
第2級
第1級
61
result
1
target
mode
EN
data_pre
data_out
Q_2=ACAA
Q_3=CAAT
612
617
0
target
616
EN_pre
611
615
Q_1=AATT
data_in
Q_0
Q_1
Q_2
Q_3
Q_4

圖 23

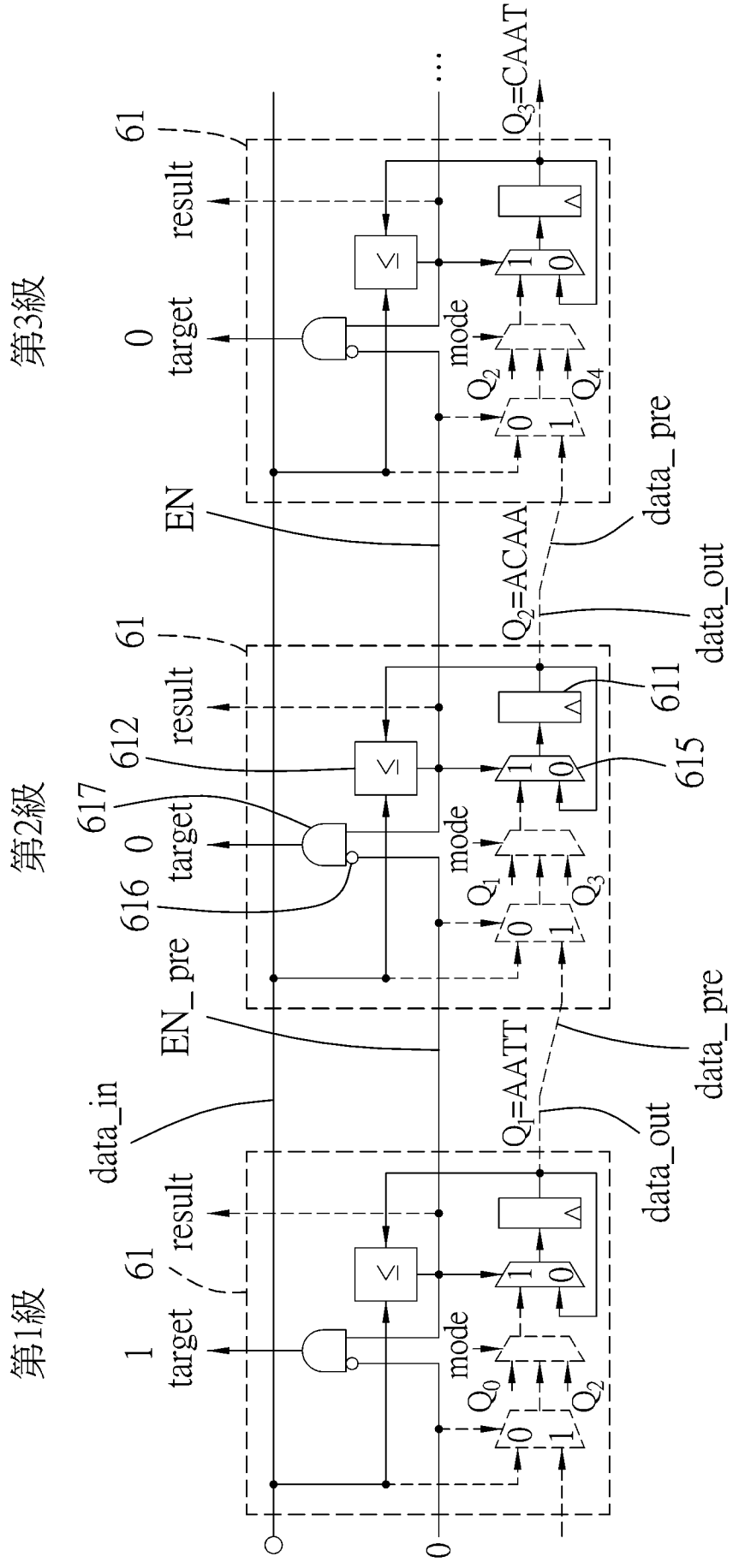
第3級
第2級
第1級
61
result
target
mode
EN
data_pre
data_out
data_in
EN_pre
Q_3=CAAT
Q_2=ACAA
Q_1=AATT
611
612
615
616
617

圖 24

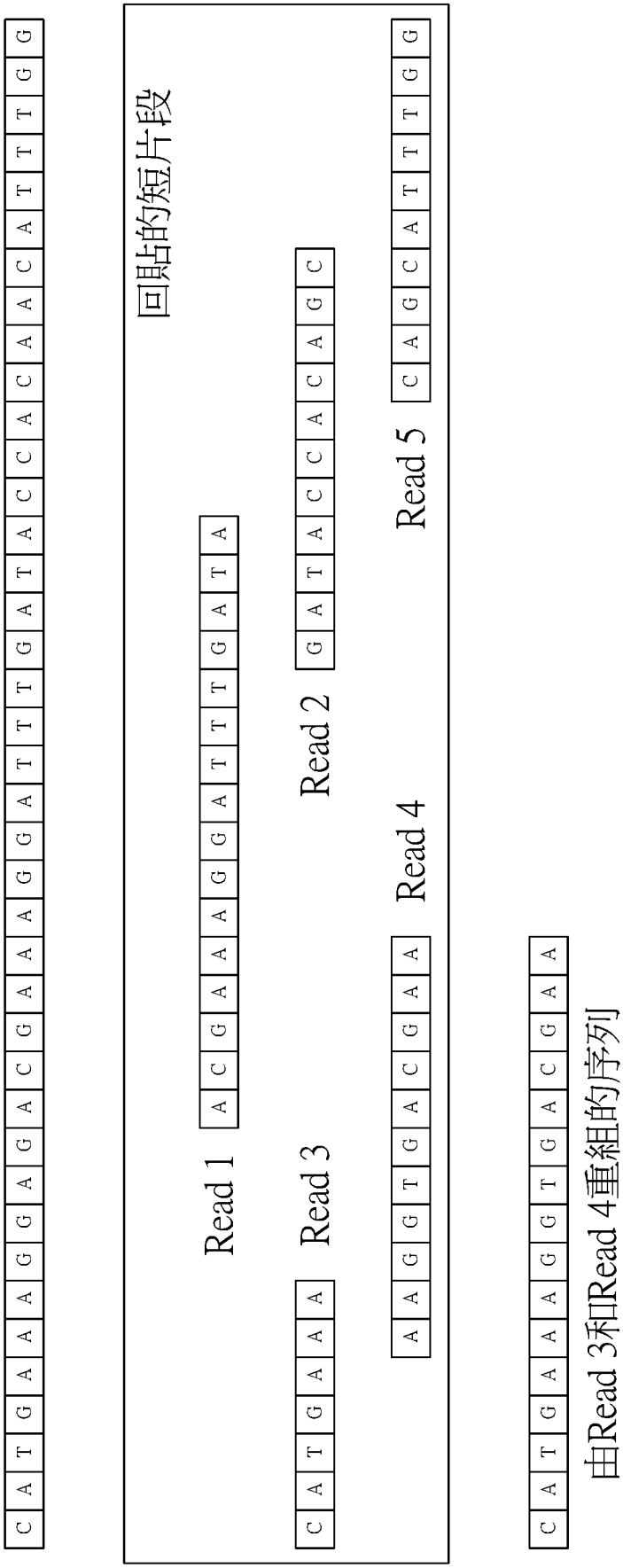
參考DNA序列
C A T G A A A G G A G A A A G G A T T T G A T T A C C C A C A A C A T T G G
回貼的短片段
Read 1
A C G A A A G G A T T T G A T A A
Read 3
C A T G A A A
Read 2
G A T A C C A C A G C
Read 4
A A G G T G A C G A A A
Read 5
C A G C A T T T G G
由Read 3和Read 4重組的序列
C A T G A A A G G T G A C G A A

圖 25

相似度分數矩陣表

b \ a		G	T	A	C	A	T
		5	3	1	0	0	0
G	5	3	10	8	6	4	5
T	3	1	8	15	13	11	9
A	1	0	6	13	13	18	16
A	0	0	5	11	11	16	23
T	0	0	3	9	16	14	21
C	0						

方向矩陣表

b \ a		G	T	A	C	A	T
G		↗	↑	↑			
T	→	↗	↗	↑	↑	↑	↗
A	→	→	→	↗	↗	↗	↑
A			→	↗	↗	↗	↑
T			↗	→	→	→	↗
C			→	→	↗	→	→

圖 26

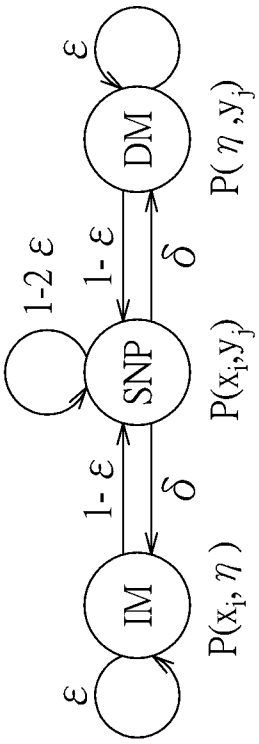

圖 27

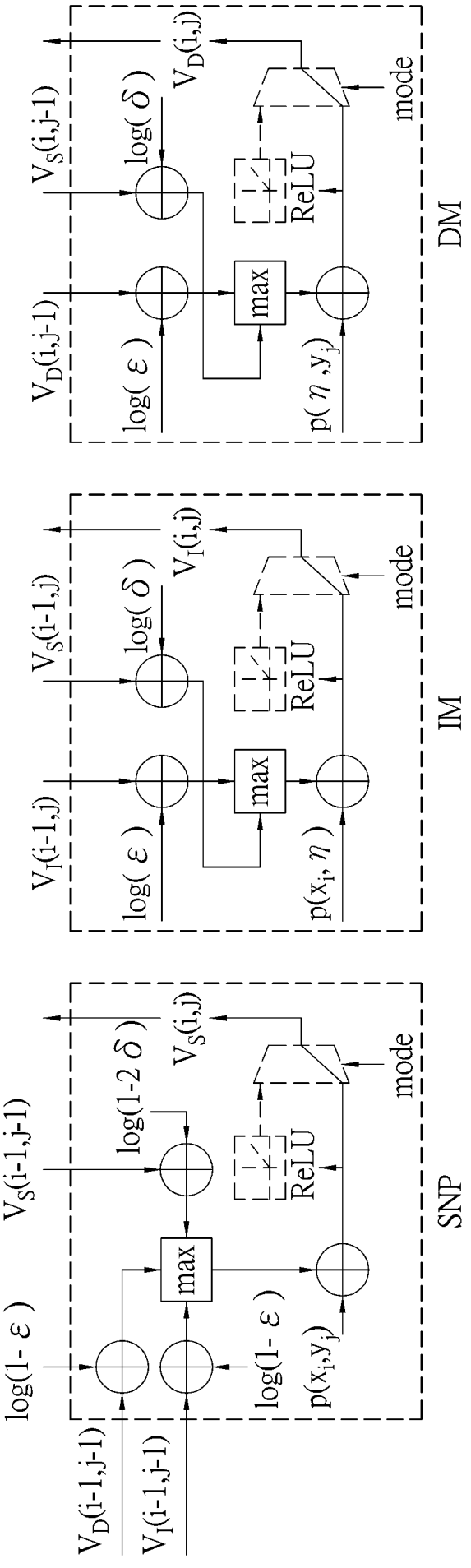
SNP
IM
DM
mode
ReLU
max
log(1- ε)
log(1-2 δ)
log(ε)
log(δ)
p(x_i,y_j)
p(x_i, η)
p(η ,y_j)
V_D(i-1,j-1)
V_I(i-1,j-1)
V_S(i-1,j-1)
V_I(i-1,j)
V_S(i-1,j)
V_D(i,j-1)
V_S(i,j-1)
V_S(i,j)
V_I(i,j)
V_D(i,j)

圖 28