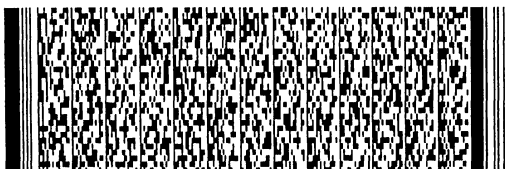


申請日期： 92. 7. 28	IPC分類
申請案號： 92120509	G06F19/00 // G06F15/00

(以上各欄由本局填註)

發明專利說明書

一、 發明名稱	中 文	推衍個體之基因體之方法及裝置
	英 文	METHOD AND APPARATUS FOR DERIVING THE GENOME OF AN INDIVIDUAL
二、 發明人 (共2人)	姓 名 (中文)	1. 羅伯森，貝瑞 2. 穆許林，理查
	姓 名 (英文)	1. ROBSON, Barry 2. MUSHLIN, Richard
	國 籍 (中英文)	1. 英國 GB 2. 美國 US
	住居所 (中 文)	1. 美國紐約州10708布朗克斯市查特菲路148號 2. 美國康乃迪克州06877里吉菲德市雷根道72號
	住居所 (英 文)	1. 148 Chatfileld Road, Bronxville, NY 10708 USA 2. 72 Regan Dr., Ridgefield, Conn 06877 USA
三、 申請人 (共1人)	名稱或 姓 名 (中文)	1. 國際商業機器股份有限公司
	名稱或 姓 名 (英文)	1. INTERNATIONAL BUSINESS MACHINES CORPORATION
	國 籍 (中英文)	1. 美國 US
	住居所 (營業所) (中 文)	1. 美國紐約州10504亞芒克市新奧爾察德路 (本地址與前向貴局申請者相同)
	住居所 (營業所) (英 文)	1. New Orchard Road, Armonk, NY 10504, USA
	代表人 (中文)	1. 傑羅 羅森梭
	代表人 (英文)	1. Gerald Rosenthal



一、本案已向

國家(地區)申請專利	申請日期	案號	主張專利法第二十四條第一項優先權
美國 US	2002/10/11	10/269,150	有

二、☐主張專利法第二十五條之一第一項優先權：

申請案號：

無

日期：

三、主張本案係符合專利法第二十條第一項☐第一款但書或☐第二款但書規定之期間

日期：

四、☐有關微生物已寄存於國外：

寄存國家：

無

寄存機構：

寄存日期：

寄存號碼：

☐有關微生物已寄存於國內(本局所指定之寄存機構)：

寄存機構：

寄存日期：

無

寄存號碼：

☐熟習該項技術者易於獲得，不須寄存。

五、發明說明 (1)

一、【發明所屬之技術領域】

本發明係關於一種資料的電子傳輸，特別係關於以電腦為基礎來表示個體的基因組。

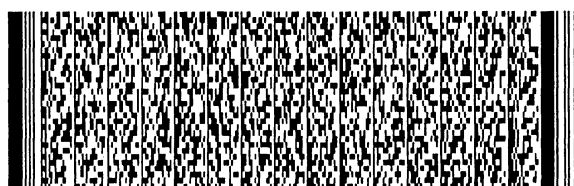
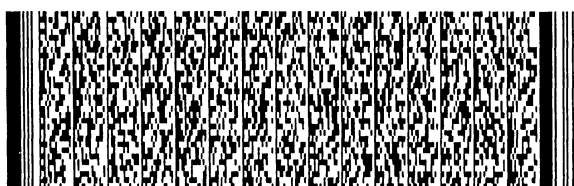
二、【先前技術】

排列人體基因組及最近其他在生物資訊領域的進展，顯示未來的醫學將可發揮利用基因組資料之優點。例如，研究員及醫護服務機構可基於預期藥物結合一病患基因序列之蛋白質碼的能力以設計藥物或篩選各式藥物。除此之外，網際網路已被廣泛使用來取得醫藥資訊，醫藥資料是在網際網路上檢索最多的資訊之一，推測到二00五年將有十億人上網，屆時如何有效地傳輸如此大量的基因組資料，將是一項新的挑戰。電腦及網際網路的使用也會因基因組序列的資料探勘而越來越頻繁。包含基因組資料之增加的傳送量需要更多有效方法來發送基因組資訊及其相關資訊。

個體基因組資料的傳輸是有困難的，因為目前資料龐大。習知的電子傳送基因組之方法太慢且很容易發生錯誤及非法存取。在傳送個體基因組資料時發生錯誤，尤其在藥治療時，會造成很嚴重的結果。因此，一個有效且精確地傳輸基因組的方法是很有必要的。

三、【發明內容】

本發明藉由提供改良的個體基因組表示法來解決上述



五、發明說明 (2)

及其他的問題。

在此揭示一種推衍個體基因組的方法。此法包含對個體存取選擇器及對群基因組存取參考模板兩步驟，選擇器包含一個所在位置值和一個鹼基質；處理選擇器及參考模板來推衍個體基因組序列代表。

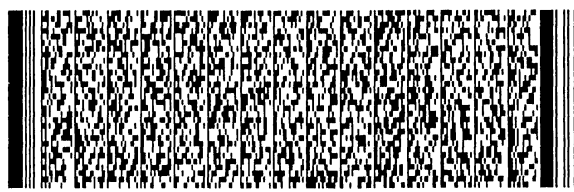
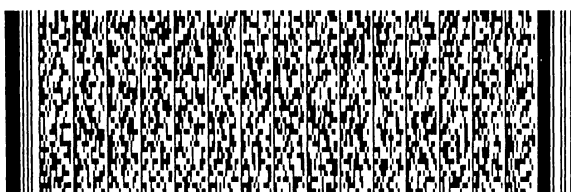
參考模板最好由表示鹼基值出現機率的資料元件所構成。此出現機率是基於群基因組中在相對的所在位置值的鹼基值發生頻率。本發明之方法進一步包含從在參考模板的資料元件計算鹼基值的步驟，非在選擇器的鹼基值。

對於本發明更完整的認識及更多的特性與優點，可參考下列細部描述及圖示。

四、【實施方式】

本發明將藉由下面基因訊息系統(GMS)加以圖示說明。在圖示實施例中，此發明是關於DNA序列資料的表示。然而要知道的是，本發明並非侷限在此特殊應用，而是可應用到其他關於基因組的資料，例如RNA的排列。

GMS和在新興臨床生物資訊領域的軟體有關，亦即臨床基因組資訊科技(IT)專注於病患的特定基因構造與其和健康狀況及疾病之關係。臨床生物資訊與習知的生物資訊是截然不同的，臨床生物資訊在意的是基因組和個別病患及集體病患群的臨床紀錄。因此，本發明不僅可在藥物研究應用上帶來益處，也可在保健資訊科技應用上獲得助益，如在電子健康目錄上的應用。



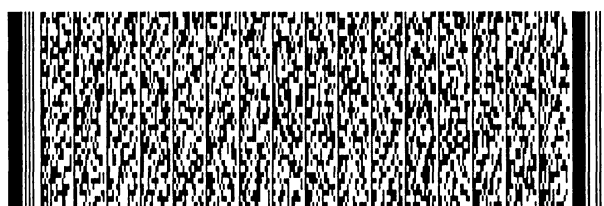
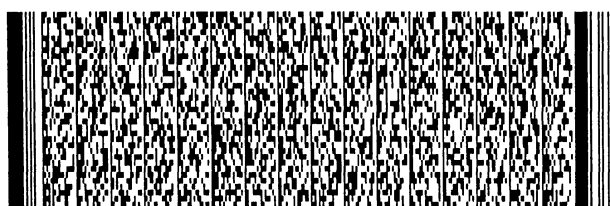
五、發明說明 (3)

基因組及生物資訊的臨床應用需要特別考慮病患的隱私問題(參考2001年Jones and Bartlett出版社出版, P. R. Lee及C. L. Estes執筆, "國家的健康"第六版一書中, George J. Annas之範例"國家病患權利的條例草案"), 病人的安全及病人與內科醫師所做任何決定的結果。聯合醫療保險可攜性和責任法(HIPPA)最近實施線上醫藥資料的保密, HIPPA說明傳送、儲存及操作病患基因組資料。

既然此發明的系統包含各式醫療照護(含緊急醫療照護), 如此設計就是要和其他系統有最小的相依性。訊息網路可包含筆記型電腦或其他可攜式設備之間不使用伺服器的直接通訊, 甚至可用磁片來當成資料傳送的方法。萬一其它所有介面都故障, 可以建立及使用讀取原始傳送文件代表的基本工具。

此發明的另一項優點是符合了健康等級七組織(HL7)所推薦的臨床資訊技術標準。HL7是非營利的ANSI授權標準發展組織, 提供支援臨床病人照護及保健服務的資料之交換、管理及融合標準。例如, HL7提出了一種在醫藥應用上明確的XML實施例之臨床文件架構(CDA)。雖然HL7是這些突出標準的本體, 但這些標準仍然處於波動未定的狀態。例如, HL7對基因組資訊的建議是非常少的。

圖1是以GMS100為典範的方塊圖。圖示系統100包含一個基因組訊息模組110、一個接收模組120、一個基因體序列資訊庫130及視情況存在的臨床資訊資料庫140。基因組

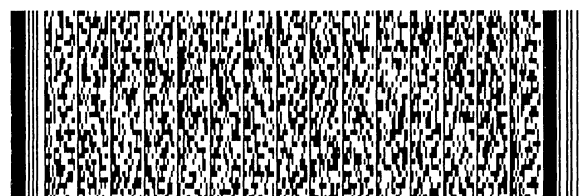
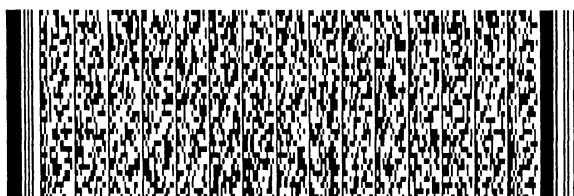


五、發明說明 (4)

訊息模組110接收由基因組序列資訊庫130而來的輸入序列和視情況來自臨床資訊資料庫140的臨床資料。基因組訊息模組110帶著輸入資料組成傳送到接收模組120的輸出資料流150。

圖二為根據本發明的一實施例之推衍個體基因組系統200的方塊圖。系統200包含一個與媒體250互動的電腦系統210。電腦系統210包含一個處理器220、一個網路介面225、一個記憶體230、一個媒體介面235及一個視情況的顯示器240。網路介面225讓電腦系統210可連接網路，而媒體介面235讓電腦系統210可和媒體250互動，如數位多用途磁碟(DVD)或硬碟。

如同已知的技術，在此所說的方法及裝置可以視為一個含有電腦可讀碼法的電腦可讀媒體的產品。電腦可讀程式碼法可以結合如電腦系統200的電腦系統來操作，以實行全部或部分的步驟來履行這些方法或創造出在此所述的儀器。配置電腦可讀碼是用來對個體存取選擇器及群基因組存取參考模板。選擇器包含位子值與鹼基值；處理選擇器及參考模板來推衍序列的個體基因組代表。電腦可讀媒體可以是可記錄媒體(如磁碟機、硬碟、如DVD的光磁碟機、或記憶卡)，或是傳輸設備(如一包含光纖之網路、全球資訊網、電纜線或使用分時多工存取、分碼多工存取或其他無線電頻率通道的無線通道)。任何已知或發展中的媒體，可以適當地用電腦系統來儲存資訊者都可使用。任何可以讓電腦讀取指令及資料的機體結構都叫電腦可讀碼法，如



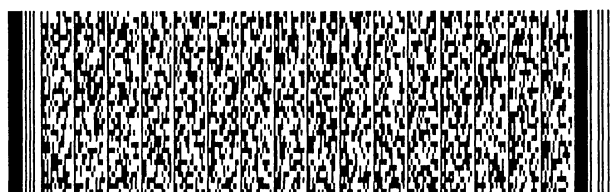
五、發明說明 (5)

在磁性媒體的磁變化或在光碟表面的高度變化。

記憶體230配置處理器220來實作揭示於此的方法、步驟及功能。記憶體230可以是分散式的或區域性的，而處理器220可以是分散式的或是單一的。記憶體230可以用電的、磁的、光學的記憶體或這些的其他組合或其他型態的儲存裝置來實作。此外，“記憶體”可作較廣的解釋，包含任何能被處理器220由可定址空間中的位址讀出及寫入的資訊。有了這個定義，可由網路介面225存取的網路資訊仍然是在記憶體230中，因為處理器220可以由網路檢索資訊。要注意的是組成處理器220的每個分散之處理器通常都有它們獨自的可定址記憶空間。也須注意的是，部分或全部的電腦系統210可以合併成專用或通用的積體電路。

視情況的顯示器240可以是任何適合於和系統200使用人互動的型式。通常，顯示器240是電腦螢幕或其他相似的顯示器。

值得注意的是，此發明之另一個可替代的實施例是在網路上實作，如網際網路。此網路另一方面可以是私人網路及/或區域網路。要了解到的是，伺服器包含一個以上的電腦系統。亦即，一組或多組圖1的元件，可位於各自的電腦系統並由各自的電腦系統執行，例如使用他們自己的處理器和記憶體。在另一個結構中，此發明的方法可在個人電腦上完成並藉由無伺服器介入的網路直接將輸出資料傳送到接收模組(如另一台個人電腦)。輸出資料也可以不用網路來進行轉存。例如，輸出資料可以簡單地下載至如



五、發明說明 (6)

軟碟的設備上，再將資料上傳到接收模組。

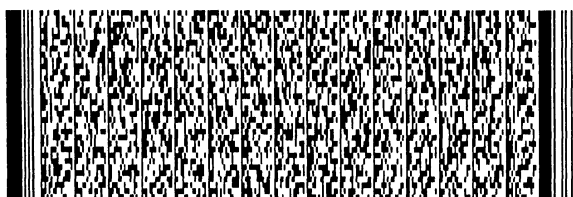
GMS 語言(GMSL)是一項用來表示臨床及基因組資料潛在廣泛什錦組合之新穎開放式語言 (lingua franca)，用來做安全及簡潔的傳輸。資料可能有許多不同來源，有著不同格式，且為了可以在下游應用中的廣大範圍內使用。GMSL 為基因組資料的註解做了最佳化。

GMSL 初步功能有：

- 在有必要下保留原始臨床文件的內容，並結合病患DNA序列或片段。
- 允許專家在儲存或傳送DNA及臨床資料前加上註解。
- 提供額外的加密或檔案保護。
- 提供病患身分等的不同程度可回復與不可回復"刷洗"式(匿名性)之工具。
- 避免錯的DNA及其它實驗室資料加到錯的病人紀錄上。
- 在不同階層提供各種的壓縮及加密型式，這可以應用到最終檔案的標準方法來予以補充。
- 由接收者來選擇最終資訊描述方法，包含選擇能看到什麼，以及
- 允許一個符合XML的特殊型式"錯開"括弧來對可交疊的DNA及蛋白質特性(和有效的XML標籤不同)編碼。

就像其他許多的電腦語言一樣，GMSL 認得兩種基本元件：指令(命令)及資料。既然GMS 為了操控可能極大的DNA或RNA序列而做了最佳化，這些元件架構也設計成簡潔的。

有一類關於位元組對映法則的指令允許四個鹼基封裝



五、發明說明 (7)

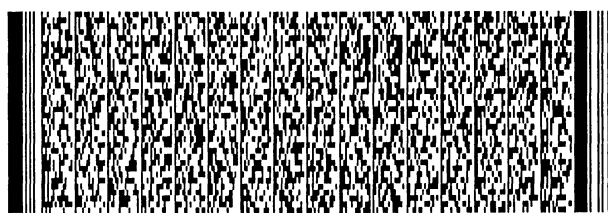
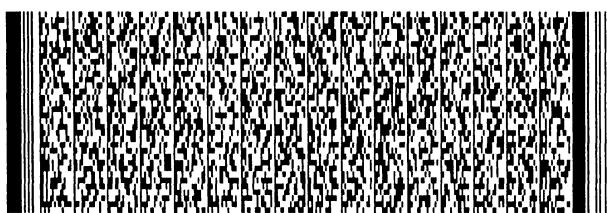
成一位元組來提供最佳壓縮串流。此特性對於處理不被註解中斷的長DNA序列是有用的。此緊密的封包會持續到遇上非DNA特性的特殊終止序列才結束。此壓縮資料可以在主要串流上傳送，或是在解碼期間從分割的檔案中讀取。另一類指令利用打開或闔上"括號"(如圓括號)來將資料分類。這種指令可用於勾勒出有待處理之基因序列的特種輪廓。不同於只能說是蜂巢式(如{a[b(c)d]e})的圓括號及標記標籤，GMS的括號可以交叉使用，如{a[b(c}d)e]。這種特性對基因組註解是重要的，因為我們有興趣的部分通常會互相交疊。它也能在同一時間用複數個方法處理一序列之相同部分或各序列之交疊部分，如有註解的或有限制的。

除了這些"混合"指令外，有些指令不和任何基因組序列的特殊部分結合，而有些指令和大量的基因組資料位元組結合。指令碼可以是初步資訊的。例如，一個特殊的指令可以指出在此點插入或刪除之一基因組鹼基，或是在此處發生之一連串此等鹼基。

當基因組序列中某些地方出現序列是實驗性地不可靠或實驗性地不清楚是否為一特殊核苷酸鹼基時(如A或G)，此序列可由指示一可靠片段於此結束或之後片段有不確定性的指令來加以中斷。因此，GMS中包括持續追蹤複數個片段的能力。GMS有能力知道片段的數量，也可以視情況需要在其上分割或加上註解，例如在XML的輸出上。

以下為一個命令片語或數個命令組成的群體之範例：

```
password ; [&7aDfx/b{by shaman protect data] ;
```



```
xml ; [ </gms : {patient}_dna>\ ] ; index ; and protein ;
```

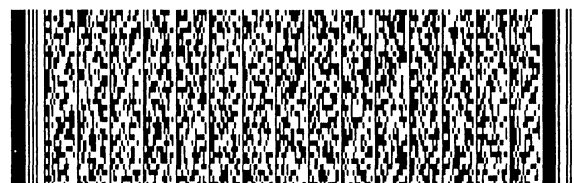
一個有用的DNA註解命令格式範例如下：

(43

這個強迫標籤到最後XML輸出檔案上，如`<open feature="whatever" type="43" level=8/>`是由括號等級決定。此命令是用來註解交疊的特性，例如DNA及蛋白質的特性不允許到XML上去（XML`<A><B></B></A>`是XML允許的，而`<A><B></A></B>`則不行）。

一般的資料敘述編碼或特定或普通的資料類別，例如包含：如

```
data ; [ ... .. / ] ;
```



五、發明說明 (9)

```

password ; [ ... .. / ] ;
filename ; [ ... .. / ] ;
number ; [ ... .. / ] ;
xml ; [ ... .. / ] ; (XML)
perl ; [ ... .. {end of data} ] ; (Perl 在清單上執行應用程序)
hl7 ; [ ... .. {end of data} ] ; (HL7 信息)
dicom ; [ ... .. {end of data} ] ; (影像)
protein ; [ ... .. / ] ;
squeeze dna ; [ ... .. / ] ; (將DNA壓縮成每位元組有四字元)

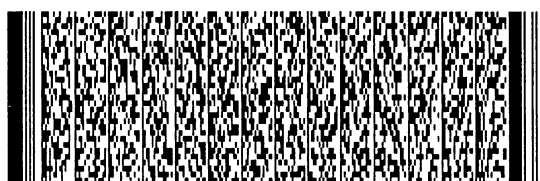
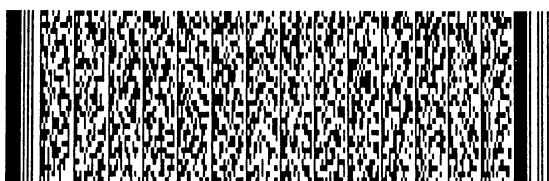
```

用" data ; / / " 形式來替換是可以的。終止括號"] " 是視情況存在的，事實上也是個檢查清單上資料敘述的同位檢查指令。在" [... .. " 裡面可以插入" 型式 " 允許的本文。目前型態的限制是非常少的，但是反斜線在某些資料型態上將會有所限制，避免在內容上被認為是符號。

在大括號內的各式指令都可以在這些" 資料 " 範圍出現，如 {xml 符號} {定義資料} {喚回資料} {用密碼開啟資料}，或是提取不同的名稱，如 {位置} 是僅在清單上用來計算數值和極可取代到資料上去。

基本語言可以組合出無數的片語，相對地形成有較少複合指令。例如：

```
filedata;[{by shaman unlock data}]
```



五、發明說明 (10)

```
number;[15 base pairs\]
```

```
squeeze dna
```

```
*
```

```
AGCTTCAGAGCTGCT\
```

這些指令在下列資料上有一保護鎖，需要一密碼(此例為 shaman)來存取。在可能範圍內這些指令亦將15鹼基對壓縮成每位元組4對。另一個範例：

```
name;[mary\];xml;[elizabeth {define data}]
```

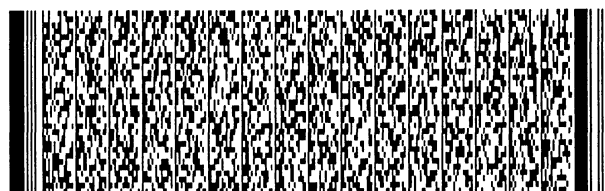
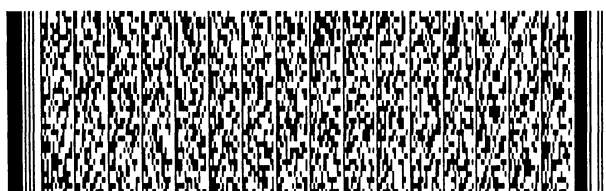
```
xml;[<test>patient{identifier} has informal code  
name{mary}</test>\];index
```

用具體指定的XML描述使用者自訂變數 "mary" 及系統變數 "identifier" (目前的病人身份) 的使用。

基因組資料輸入檔(.gmd)包含DNA序列及視情況存在之手動註解。這些DNA序列是許多鹼基的串列。泛空白字元 (white space) 是不重要的。使用字首為 "gms" 的XML型式標籤當註解，但此檔案並非是 Xml 文件。

"轉型盒"(cartridges)在此是用多種方法來改變輸入輸出的可替換程式模組。可將之想像成迷你專家系統，用來撰寫技術、需求和喜好。所有輸入轉型盒最後都產生.gms檔來當成最後且主要的輸入步驟。這個檔被轉成二進制.gmb檔並且儲存起來或傳送出去。輸入轉型盒包含如既有轉換轉型盒(Legacy Conversion Cartridges)，用來將既有臨床及基因組資料轉換成GMS語言。

當.gmi檔是一個CDA文件時，如同在一個最新臨床寶庫

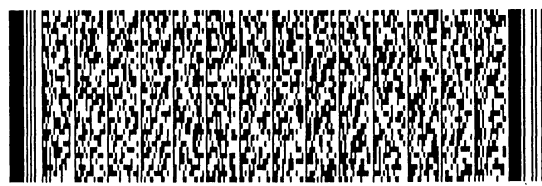
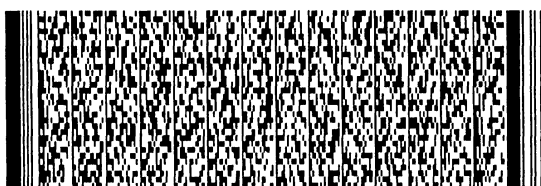


五、發明說明 (11)

檢索資料時所預期的，GMS需要知道如何將標記CDA標籤的內容轉變成標準.gms型式。使用GMS轉型盒可以完成此任務。在描述支援自動化的第一個GMS轉型盒應用中，專家選擇地修改來自CDA格式的檔案來包含額外的注解及結構。另外，前述的模板型式對指引本程序是有用的，為了讓任何修改過的文件剩下符合CDA的。此附加基因組特性的最後CDA文件表示一個“CDA基因組文件”。這樣的CDA文件現在可以自動轉成GMSL了。除了上述既有紀錄轉換轉型盒，為了從原始CDA無基因組檔案自動產生CDA基因組文件，本發明也考慮到了自動加入基因組資料。

譬如，基因組資料可使用一gms來併入：在CDA<body>結尾的名稱集字首，在他自己的CDA<section>，如下CDA結構所示：

```
<cda:clinical_document_header>
.
.<!-- header structures per CDA -->
.
</cda:clinical_document_header>
<cda:body>
.<!-- clinical sections per CDA -->
.
<cda:section>
    <cda:caption>
        IBM Genomic Messaging System
```



五、發明說明 (12)

Data

</cda:caption>

<cda:paragraph>

<cda:content>

<cda:local_markup

ignore="markup">

<!--gms: tags go

here-->

</cda:local_markup>

</cda:content>

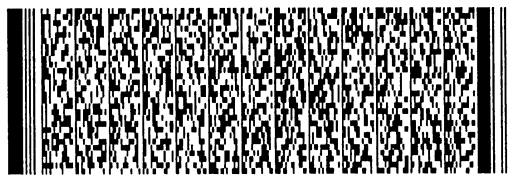
</cda:paragraph>

<cda:section>

<cda:body>

更精確地說，轉型盒首先看看這些標籤是否已存在文件中，若有，轉型盒將保留這些標籤。若這些標籤不存在，轉型盒將找一個<gms:body或<body tag(不分大小寫)。然而，如果沒有body標籤，轉型盒將在文件最後標籤前插入<gms:body或<body tag(不分大小寫)。關於GMS及包含基因組序列資料過程的更多資訊，在二00二年七月二十八日編檔，標題為“基因組訊息系統”美國專利應用號碼10/185,657，有提出討論，在此列為參考。

圖3是推衍個體基因組典範方法300的流程圖。如圖3所示，方法300包含步驟320來處理選擇器及步驟330來處理參



五、發明說明 (13)

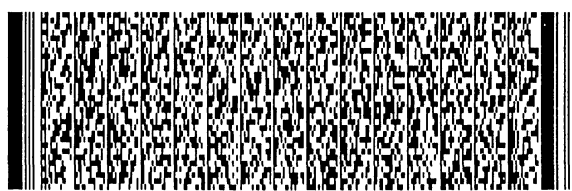
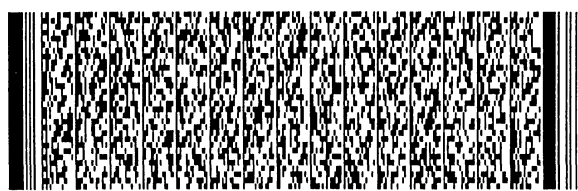
考模板。每個步驟將分別結合圖4及圖5，在下面做更細部的探討。

圖4是步驟320(圖3)處理選擇器的更細部流程圖。如圖4所示，處理選擇器包含一個404的步驟來取得選擇器。一旦取得了選擇器，接著就是步驟406決定一位置值及步驟410決定一鹼基值。位置值表示在核苷酸序列的一個位置。鹼基值表示一個核苷酸鹼基。較佳的核苷酸鹼基包含嘌呤：腺嘌呤(A)與鳥糞嘌呤(G)及嘧啶：胞嘧啶(C)與胸嘧啶(T)或尿嘧啶(U)(及尿嘧啶是在RNA上)，但不以此為限。例如，選擇器含有鹼基值及位置值(A, 6)，表示存在核苷酸序列第六個位置是核苷酸鹼基腺嘌呤。

藉由位置值和鹼基值可在個體基因組序列代表放置合適的鹼基值，如步驟416所示。個體基因組序列代表是由處理選擇器與參考模板所推衍出的核苷酸序列。(將結合圖5於下面描述更細部的細節)，在前述例子中，選擇器包含鹼基值及位置值(A, 6)，腺嘌呤將置於個體基因組序列代表的第六個位置。

如步驟414所示，選擇器的處理會持續到沒有任何選擇器為止，將在步驟408中偵測到。

在一個較佳實施例中，鹼基值與位置值或多個鹼基值與多個位置值代表著多態性。多態性可以定義為某群體中穩定基因組的變化區域(亦即，在群體中至少通常存在1%的個體，相對於個體化的隨機突變)。另外，鹼基值與位置值可以表示特別感興趣的基因組區域。典範的感興趣區域包



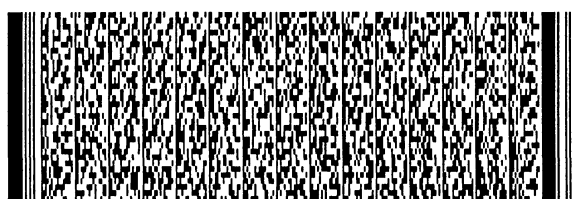
五、發明說明 (14)

含對一個或一群蛋白質編碼的基因組範圍。

以選擇器來表示個體基因組包含鹼基值與位置值，亦即多態性、感興趣範圍或兩者皆是，允許僅傳送個體基本基因組的精華資料。然後此傳送資料會在如GMS的接收端與參考模板調和。因此，可以將基因組資料傳送做得更有效率且更精準。

接著開始處理參考模板。參考模板是一個群組基因組的核苷酸序列代表。“群組”一字可以是描述個體的任意群，子群或小集團。但較佳者是子群。適合本發明的子群可以由幾個參數來定義，可包含種族、群族、部族、宗族、家族及手足，但不以此為限。本發明方法可以替視為一群組的每個子群決定代表的核苷酸序列。藉由將個體組成子群可辨識出更多共通的基因組特性(如一胜肽的前導區與基因的插入子的區)和更多型態蛋白質特性(如糖化作用)。

圖5是步驟330(圖3)處理參考模板的流程圖。如圖5所示，處理參考模板包含步驟504來取得資料元件。資料元件包含一個位置值與一個鹼基值或是複數個鹼基值，將於下面描述其更細部的細節。一旦取得資料元件，接著就進行步驟508來決定位置值。位置值是決定在個體基因組序列代表的位置，並不含於選擇器中。因此，在上述經典範例中，選擇器有鹼基值及位置值(A, 6)在個體基因組序列代表的第六個位置已放置了一個腺嘌呤。因此，不需從參考模板決定位置值給第六個核苷酸位置。

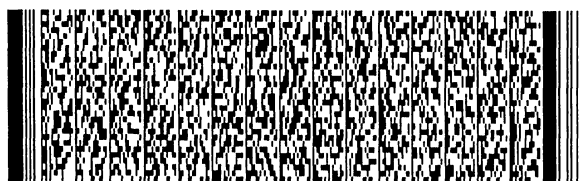


五、發明說明 (15)

如步驟508，一旦從參考模板決定出了位置值，如步驟520所示，接著將算出鹼基值。此步驟將結合圖6於下面作更細部的討論。藉由決定出的位置值及算出的鹼基值，可以在個體基因組序列代表上放置適當鹼基值，如步驟518所示。如步驟516，參考模板程序持續中。參考模板程序會持續到沒有任何資料元件，將在步驟408中偵測到。

圖6是步驟520(圖5)計算鹼基值的流程圖。在參考模板的資料元件表示在群基因組的位置值及鹼基值。資料元件可表示步驟604單一鹼基值或步驟618之複數個鹼基值。當資料元件是單一鹼基值(如步驟608)，將提出算出的鹼基值(如步驟610)且放置於個體基因組序列代表的位置值所示位置。如步驟618所示，當資料元件是複數個鹼基值時，則需決定是否有最大的資料元件，如步驟619。最大資料元件可以定義為數值最高的資料元件。若沒有最大資料元件存在，將提出如步驟620複數個鹼基值，且放置於個體基因組序列代表的位置值所示位置上。在此所說沒最大資料元件存在的情況，將於下面作更細部的細部討論。如步驟622，若最大資料元件存在，則需將它找出來。如步驟616，若資料元件不是單一鹼基值，也不是複數個鹼基值，則資料元件是空的，此時要重複此位置的程序。

表示複數個鹼基值的資料元件會發生在，例如，當在群基因組的特殊位置值上表示有複數個鹼基值時。在此例中，資料元件表示在此位置值的特殊鹼基值發生的機率，亦即腺嘌呤、鳥糞嘌呤、胞嘧啶及胸嘧啶之一發生的機



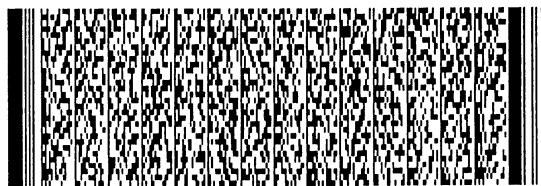
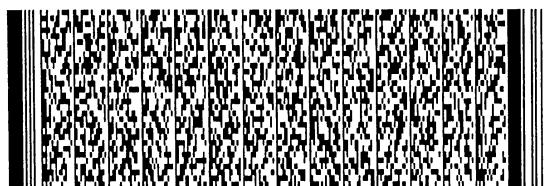
五、發明說明 (16)

率，基於在群基因組相對位置腺嘌呤、鳥糞嘌呤、胞嘧啶及胸嘧啶的發生頻率。在群基因組的相對位置表示在組成群基因組的複數個序列存在單一位置。例如下列參考模板：

... (40, 30, 10, 20) (20, 20, 60) (50, 10, 40) (33, 33, 34)
(90, 5, 5) ...

每個括號的數值集合表示在群基因組某特殊位置的特別鹼基值之出現機率。上面例子中，出現機率表示在相對位置上具有特殊鹼基值的群基因組之百分比。因此，若第一個括號內的數值集合分別表示腺嘌呤、鳥糞嘌呤、胞嘧啶及胸嘧啶的出現機率，則在這個位置此群的40%有腺嘌呤，30%有鳥糞嘌呤，10%有胞嘧啶，20%有胸嘧啶。此外，剩下的四個括號數值集合表示四種DNA鹼基值中有一種不會在此位置出現（亦即三個出現機率值總和是100%）。關於包含發生值機率的參考模板細節，在此同時期，標題為“推衍代表核苷酸來表示群基因組的方法與設備”（代理人檔案號碼YOR920010649US1）美國專利應用號碼_____有提出討論，在此列為參考。

如步驟622，為了決定最大資料元件，將決定由資料元件表示的最大出現機率，如步驟624。接著將最大出現機率的鹼基值放置於個體基因組序列代表的決定位置值之位置上。



五、發明說明 (17)

使用查驗表來決定對應到的最高出現機率之鹼基值，如步驟628及626所示。查驗表藉著指示出現機率的位置來指出何鹼基值對應至何出現機率。亦即在括號內的數值集合。一個典範的查驗表如下：

位置 鹼基值

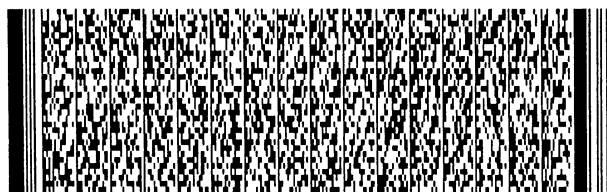
1	A
2	C
3	G
4	T

因此由上表可知，第一個發生值機率表示腺嘌呤，第二個發生值機率表示鳥糞嘌呤，第三個發生值機率表示胞嘧啶，第四個發生值機率表示胸嘧啶。嚴格來說，對於上示的第一括號數值集合， $\dots(40, 30, 10, 20)\dots$ ，查驗表可以顯示成：

位置 範例 鹼基值

1	40	A
2	30	C
3	10	G
4	20	T

除此之外，發生值機率可和參考模板從完全一致地。



五、發明說明 (18)

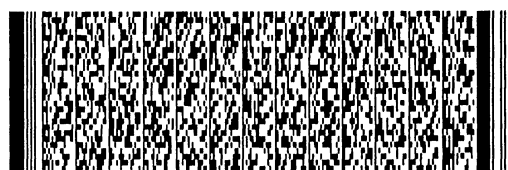
例如，第一個值都對應到腺嘌呤的出現機率、第二個值都對應到鳥糞嘌呤的出現機率、第三個值都對應到胞嘧啶的出現機率及第四個值都對應到胸嘧啶的出現機率。

較佳者，提出了四個可能的鹼基值中的三個出現機率，而第四個的出現機率推衍自總和為100%而少於另三個的總合之差值。

沒有最大資料元件的情況會發生在當個體基因組序列代表不含在選擇器有位置時，及參考模板包含表示複數個鹼基值之出現機率的資料元件但沒有最大資料元件時(例如二個或二個以上的鹼基值有相同的出現機率時)。這裡有個例子，當參考模板包含資料元件(40, 40, 10, 10)。在此例中，我們較喜歡將複數個資料值的資料元件代表放入序列中。因此，可在序列中的此位置表示多個鹼基值。

實例

下面是例示的選擇器和例示參考模板。參考模板包含位置值及資料元件。有些資料元件表示單一鹼基值，而有一些則表示複數個鹼基值。選擇器包含鹼基值和位置值。



五、發明說明 (19)

位置 →

1	2	3	4	5	6	7	8	9
A	G	50, 30, 10	C	T	0, 20, 80	A	40, 0, 0	G

10	11	12	13	14	15
C	0, 40, 60	C	40, 0, 60	G	G

個體選擇器表示成：(C, 6,)(A, 8,)

個體基因組序列代表可以使用下面的演算法來計算：

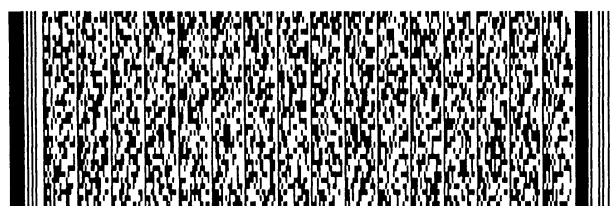
對於在模版的每個位置：

如果在這位置的值是單一鹼基，複製此值到同一位置的結果序列。

如果在這位置的值是複數個值，在選擇器尋找符合此位置之一(位置值/鹼基值)對：

如果有找到，從選擇器複製此鹼基到相同的位置上。

否則根據建立的協定(亦即查驗表)，在此混合物中找最大資料元件，且複製此鹼基值到複數個值之位置。上例的查驗表是：



五、發明說明 (20)

位 置 鹼 基 值

1	A
2	C
3	G
4	T

此個體基因組序列代表可以讀取如下：

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	G	A	C	T	C	A	A	G	C	G	C	G	G	G

雖然本發明以前述之較佳實施例揭露如上，然其並非用以限定本發明。任何熟習此技藝者，在不脫離本發明之精神和範圍內，當可作些許之更動與潤飾。下述範例將說明本發明之精神和範圍。本例只用來說明目的，所以不應用以限定本發明。



圖式簡單說明

- 圖1 為一個基因組訊息系統(GMS)範例；
- 圖2 為GMS硬體實作範例方塊圖範例；
- 圖3 為整個推衍個體基因組方法的流程圖；
- 圖4 為選擇器處理程序流程圖；
- 圖5 為參考模板處理程序流程圖；
- 圖6 為由參考模板來計算鹼基值的流程圖。

元件符號說明

- 140 臨床資訊資料庫
- 130 基因組序列資料庫
- 110 基因訊息模組
- 120 接收模組
- 220 處理器
- 230 記憶體
- 240 顯示器
- 225 網路介面
- 235 媒體介面
- 250 媒體

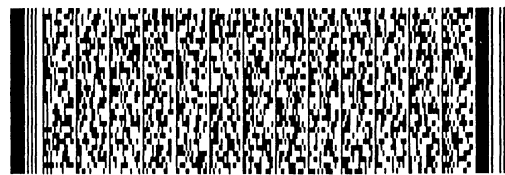
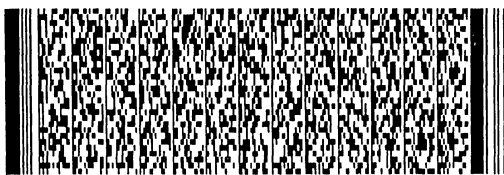


四、中文發明摘要 (發明名稱：推衍個體之基因體之方法及裝置)

提供一種以電腦為基礎來推衍個體基因組的方法。此法包含對個體存取選擇器及對群基因組存取參考模板兩步驟，選擇器包含一個所在位置值和一個鹼基值；並處理選擇器及參考模板來推衍個體基因組序列代表。參考模板較佳由表示一鹼基值出現機率的資料元件所構成。此出現機率是基於在此群基因組中在對應的位置值的鹼基值發生率。本發明之方法更進一步包含從在參考模板中的資料元件計算一鹼基值，非在選擇器的鹼基值。

五、英文發明摘要 (發明名稱：METHOD AND APPARATUS FOR DERIVING THE GENOME OF AN INDIVIDUAL)

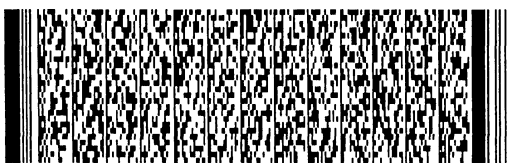
A computer-based method is provided for deriving a genome of an individual. The method comprises the steps of accessing a selector for an individual and a reference template for a group genome, the selector comprising a locus value and a base value; and processing the selector and the reference template to derive a sequence representative of the genome of the individual.



四、中文發明摘要 (發明名稱：推衍個體之基因體之方法及裝置)

五、英文發明摘要 (發明名稱：METHOD AND APPARATUS FOR DERIVING THE GENOME OF AN INDIVIDUAL)

The reference template preferably comprises data components representing a probability of occurrence of a base value. The probability of occurrence is based on base value occurrences at corresponding locus values in the group genome. The method of the present invention further comprises computing a base value from the data component in the reference template, for base values not in the selector.



六、指定代表圖

(一)、本案代表圖為：第___1___圖

(二)、本案代表圖之元件代表符號簡單說明：

140 臨床資訊資料庫

130 基因組序列資料庫

110 基因訊息模組

120 接收模組



圖 一

100

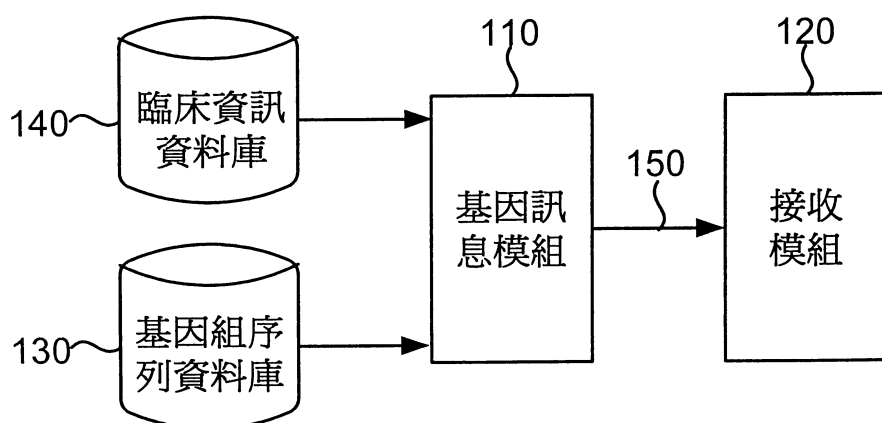


圖 二

200

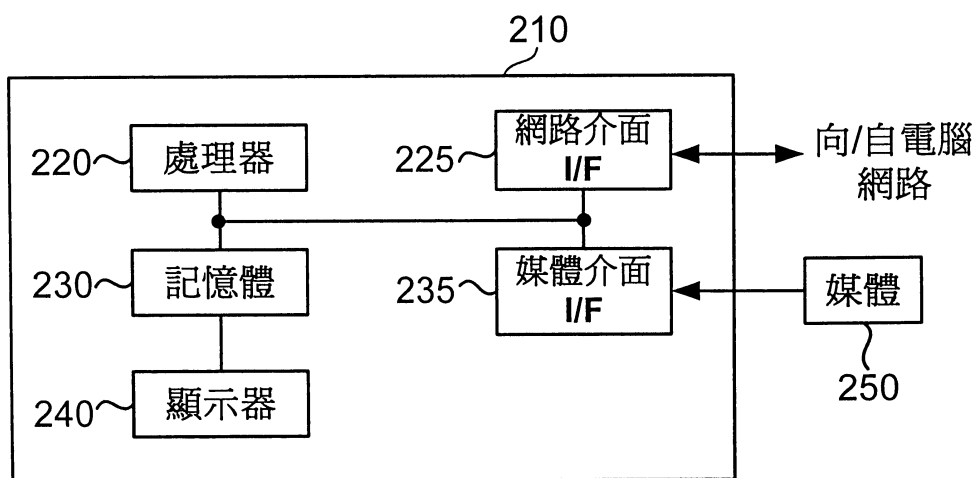


圖 三

300

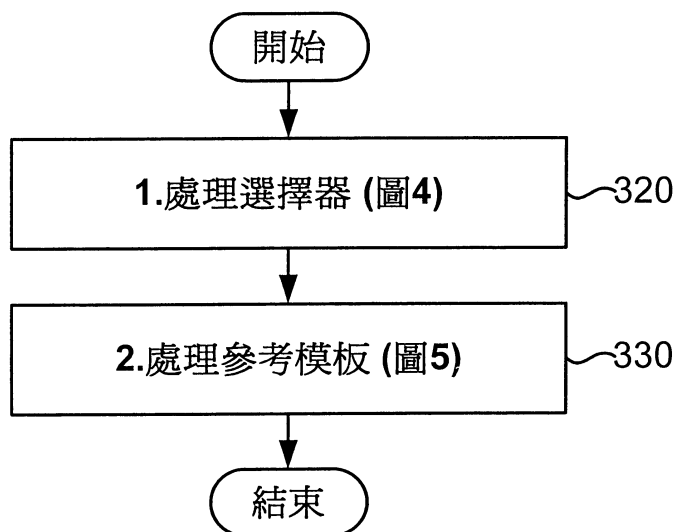


圖 四

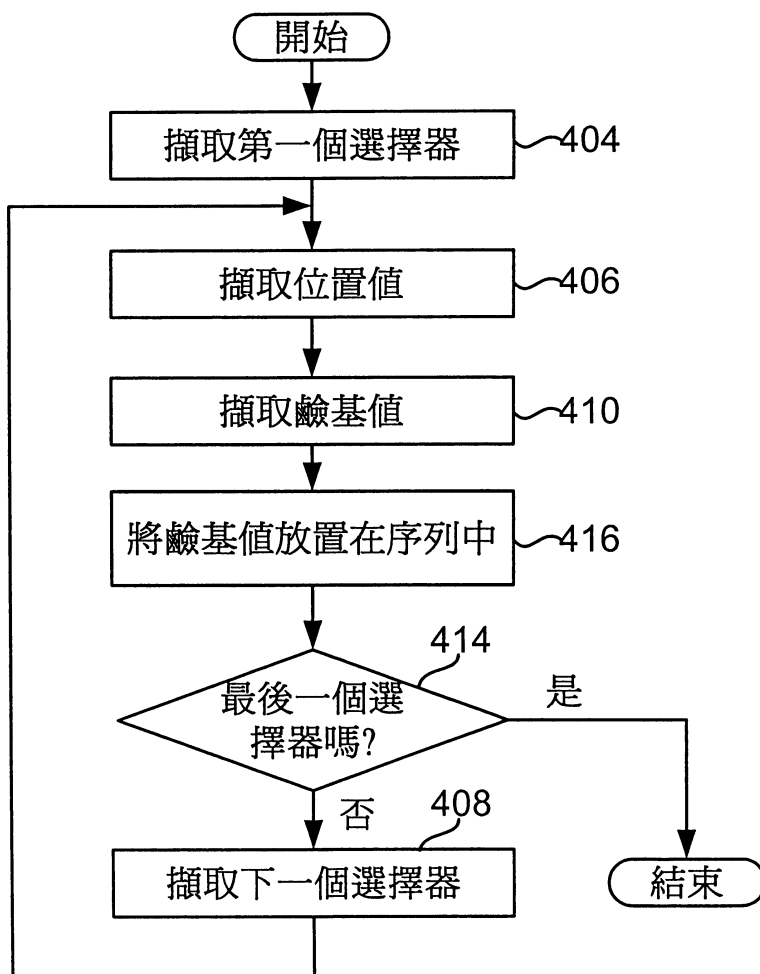
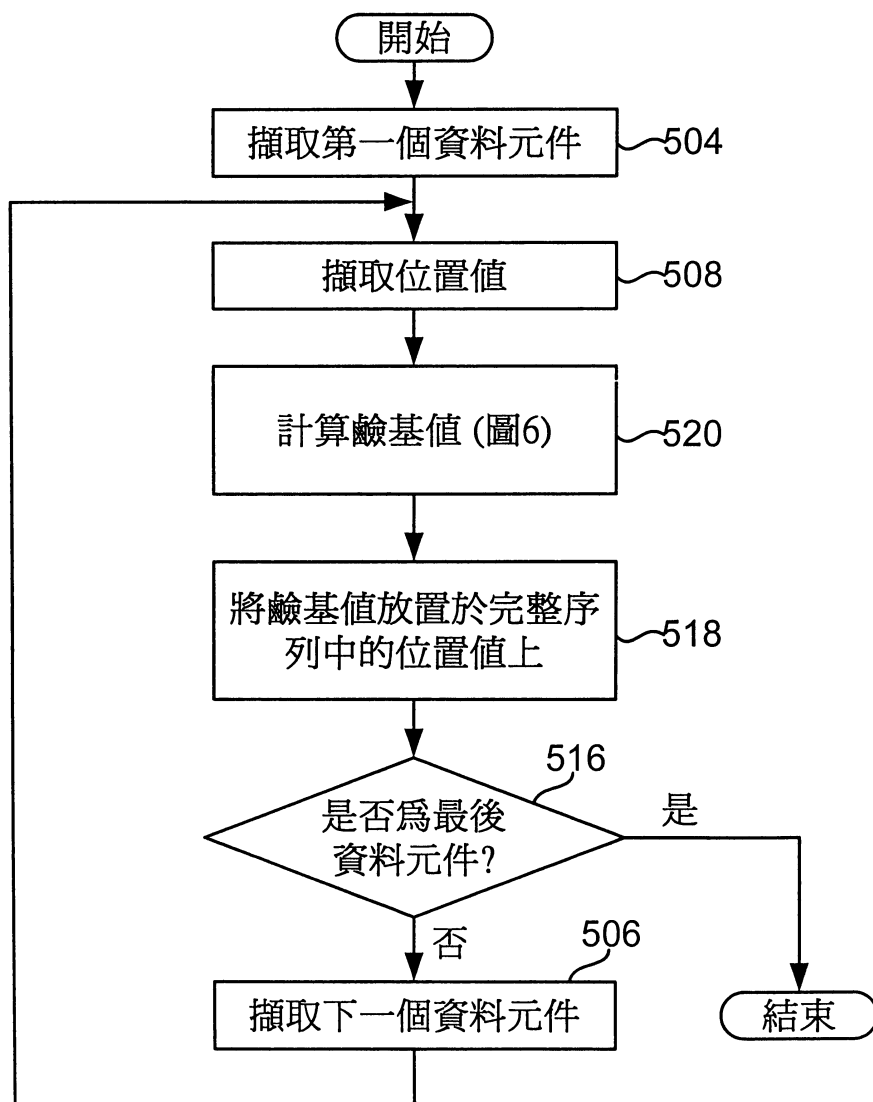
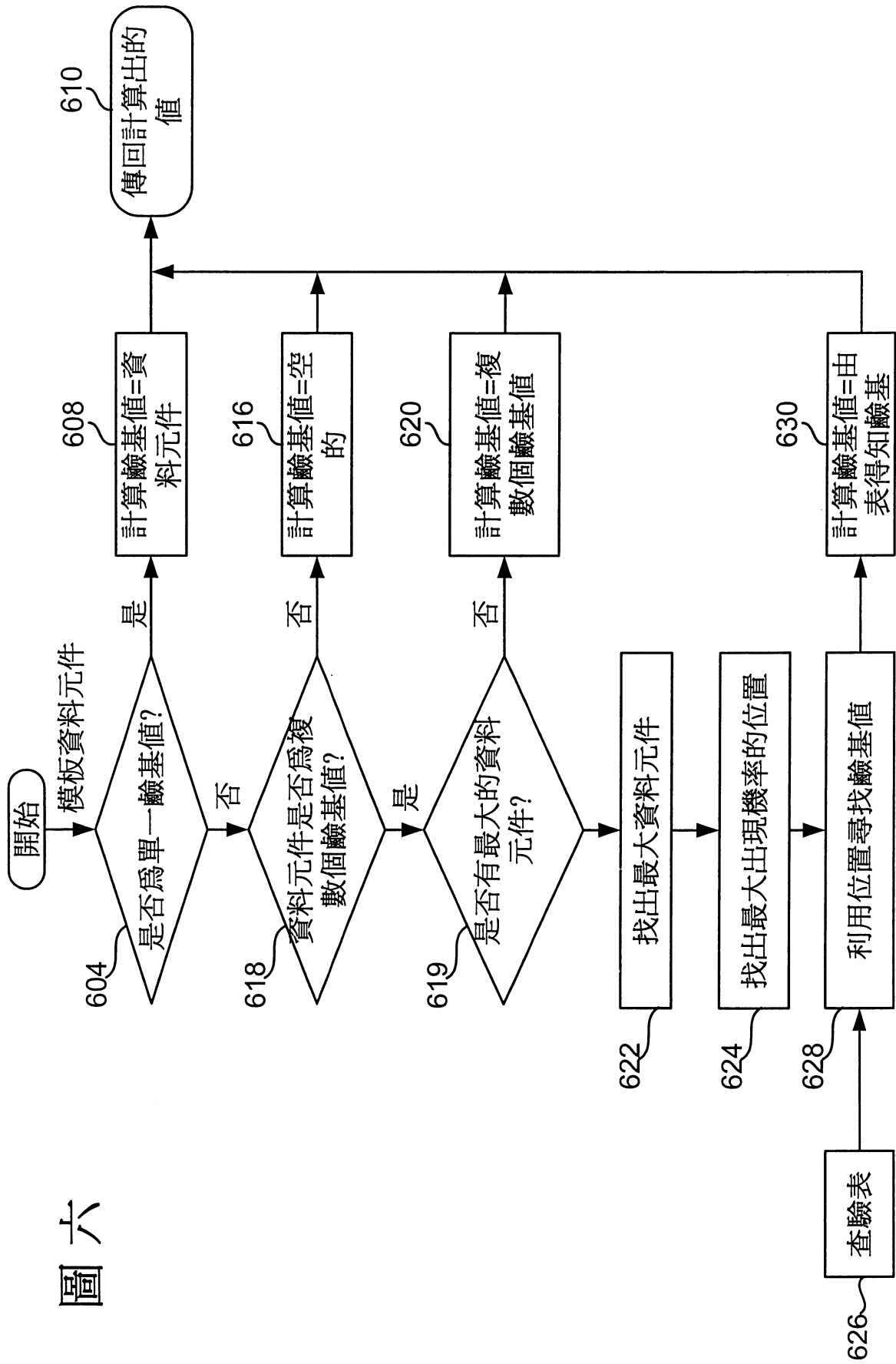


圖 五

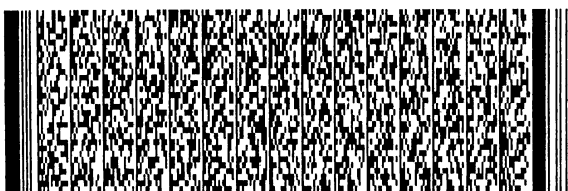


圖六



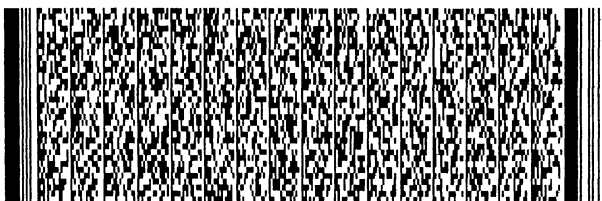
六、申請專利範圍

1. 一個推衍個體基因組的方法，包含之步驟為：
對一個體存取一選擇器及對一群基因組存取一參考模板，該選擇器包含一位置值和一鹼基值；及
處理該選擇器和該參考模板來推衍該個體基因組之一序列代表。
2. 如申請專利範圍第 1 項的方法，其中該位置值表示在一核 酸序列的一個位置。
3. 如申請專利範圍第 1 項的方法，其中該鹼基值表示一核 酸鹼基。
4. 如申請專利範圍第 1 項的方法，其中該選擇器包含複數個位置值及複數個鹼基值。
5. 如申請專利範圍第 1 項的方法，其中該參考模板包含表示一鹼基值的資料元件群。
6. 如申請專利範圍第 5 項的方法，其中該資料元件表示對該鹼基值的一出現機率。
7. 如申請專利範圍第 6 項的方法，其中該出現機率是基於在此群基因組中在對應的位置值群的鹼基值發生率。



六、申請專利範圍

- 8.如申請專利範圍第7項的方法，更進一步包含：
從在該參考模板的該資料元件計算一鹼基值，不是在該選擇器的鹼基值群。
- 9.如申請專利範圍第8項的方法，更進一步包含此步驟：
找出一最大資料元件。
- 10.如申請專利範圍第8項的方法，其中計算出來的該鹼基值包含複數個鹼基值。
- 11.如申請專利範圍第9項的方法，其中該最大資料元件表示一最大出現機率。
- 12.如申請專利範圍第9項的方法，其中找出該最大元件包含混合表的使用。
- 13.一個推衍個體基因組的系統，包含：
一個儲存電腦可讀碼的記憶體；及
一個結合到該記憶體的處理器，安裝此處理器來實作該電腦可讀碼，安裝該電腦可讀碼來：
對一群基因組存取一參考模板及對一個體存取一選擇器，該選擇器包含一位置值和一鹼基值；及
處理該參考模板及該選擇器來推衍該個體該基因組之一序列代表。



六、申請專利範圍

14.如申請專利範圍第13項的系統，其中該參考模板包含表示一鹼基值出現一機率的資料元件群。

15.如申請專利範圍第14項的系統，其中該出現機率是基於在此群基因組中在對應的位置值群的鹼基值發生率群。

16.如申請專利範圍第14項的系統，其中該電腦可讀碼進一步安裝來：

從在該參考模板的該資料元件計算一鹼基值，不是在該選擇器的鹼基值群。

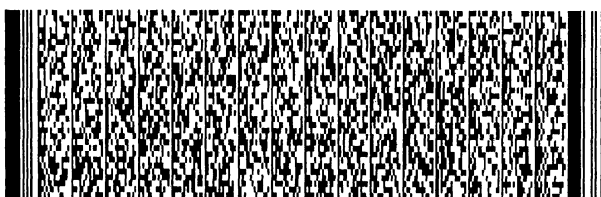
17.一種含有電腦可讀紀錄媒體製品，包含推衍個體基因組的電腦可讀碼，該電腦可讀碼包含：

一指令碼來對一群基因組存取一參考模板及對一個體存取一選擇器，該選擇器包含一位置值和一鹼基值；及

一指令碼來處理該參考模板及該選擇器來推衍該個體該基因組之一序列代表。

18.如申請專利範圍第17項的紀錄媒體製品，其中該參考模板包含表示一鹼基值出現一發生機率的資料元件群。

19.如申請專利範圍第18項的紀錄媒體製品，其中該出現機率是基於在此群基因組中在對應的位置值群的一鹼基值



六、申請專利範圍
發生率。

20.如申請專利範圍第18項的紀錄媒體製品，其中該電腦可讀碼進一步包含：

一指令碼來計算從該參考模板中之該資料元件而來的一鹼基值，不是在該選擇器的鹼基值群。

