

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6198659号

(P6198659)

(45) 発行日 平成29年9月20日(2017.9.20)

(24) 登録日 平成29年9月1日(2017.9.1)

(51) Int.Cl.

F 1

G 0 6 F 19/22 (2011.01)

G 0 6 F 19/22

請求項の数 8 (全 17 頁)

|           |                               |           |                     |
|-----------|-------------------------------|-----------|---------------------|
| (21) 出願番号 | 特願2014-77278 (P2014-77278)    | (73) 特許権者 | 501387839           |
| (22) 出願日  | 平成26年4月3日(2014.4.3)           |           | 株式会社日立ハイテクノロジーズ     |
| (65) 公開番号 | 特開2015-197899 (P2015-197899A) |           | 東京都港区西新橋一丁目24番14号   |
| (43) 公開日  | 平成27年11月9日(2015.11.9)         | (74) 代理人  | 110001807           |
| 審査請求日     | 平成28年9月27日(2016.9.27)         |           | 特許業務法人磯野国際特許商標事務所   |
|           |                               | (72) 発明者  | 木村 宏一               |
|           |                               |           | 東京都千代田区丸の内一丁目6番6号 株 |
|           |                               |           | 式会社日立製作所内           |
|           |                               | 審査官       | 松野 広一               |

最終頁に続く

(54) 【発明の名称】 配列データ解析装置、DNA解析システムおよび配列データ解析方法

(57) 【特許請求の範囲】

【請求項1】

サンプルDNA断片の両端からそれぞれシーケンシングされたペアである左方配列と右方配列との間を結合文字で連結した結合文字列を前記サンプルDNA断片ごとに作成し、各前記サンプルDNA断片の前記結合文字列を終端文字で結合した文字列をもとにリード配列辞書を作成するリード辞書作成部と、

ゲノムDNAの塩基文字列から生成されるクエリ配列が出現する前記リード配列辞書内の塩基文字座標であるヒット位置を検索するクエリ検索部と、

前記リード配列辞書内の前記ヒット位置を起点として、その周囲に位置する前記終端文字が出現するまでの文字列をサンプル配列として抽出し、

前記サンプル配列内の前記ヒット位置を起点として、その周囲に位置する前記結合文字を検査し、検査した前記結合文字から前記ヒット位置が存在しない側の前記終端文字が出現するまでの前記左方配列または前記右方配列をメイト配列として抽出するサンプル復元部と、

前記メイト配列が出現する前記ゲノムDNA内の塩基文字列の塩基文字座標を検索するマッピング部と、を有することを特徴とする

配列データ解析装置。

【請求項2】

前記配列データ解析装置は、さらに、クエリ作成部と、サンプル判定部とを備えており、

10

20

前記クエリ作成部は、前記ゲノムDNAの塩基文字列のうちの所定位置の塩基文字を別の塩基文字に変異させ、その変異させた塩基文字とその周囲に位置する前記ゲノムDNAの塩基文字列とを含めた前記クエリ配列を作成し、

前記サンプル判定部は、前記マッピング部により前記メイト配列の塩基文字座標が特定できたときに、前記ヒット位置と、前記メイト配列の塩基文字座標との文字間隔から特定される前記サンプルDNA断片の長さをもとに、前記サンプルDNA断片内のSNP (Single Nucleotide Polymorphism)を検出することを特徴とする

請求項1に記載の配列データ解析装置。

【請求項3】

前記配列データ解析装置は、さらに、クエリ作成部と、サンプル判定部とを備えており、

前記クエリ作成部は、前記ゲノムDNAの塩基文字列のうちの所定位置周囲に位置する前記ゲノムDNAの塩基文字列から前記クエリ配列を作成し、

前記サンプル判定部は、前記マッピング部により前記メイト配列の塩基文字座標が特定できたときに、前記ヒット位置と、前記メイト配列の塩基文字座標との文字間隔から特定される前記サンプルDNA断片の長さをもとに、前記サンプルDNA断片内の構造変異を検出することを特徴とする

請求項1に記載の配列データ解析装置。

【請求項4】

前記リード辞書作成部は、各前記サンプルDNA断片の前記結合文字列を前記終端文字で結合した文字列に対してBW (Burrows-Wheeler)変換したBW文字列を作成し、その前記BW文字列をWavelet Tree形式に変換することで、前記リード配列辞書を作成することを特徴とする

請求項1ないし請求項3のいずれか1項に記載の配列データ解析装置。

【請求項5】

前記リード辞書作成部は、各前記サンプルDNA断片の長さに対応した複数種類の前記結合文字を用いて、前記結合文字列を作成することを特徴とする

請求項1ないし請求項3のいずれか1項に記載の配列データ解析装置。

【請求項6】

前記サンプル復元部は、前記ヒット位置の周囲に前記結合文字が存在しない場合、その前記ヒット位置について前記マッピング部の処理対象から除外することを特徴とする

請求項1ないし請求項3のいずれか1項に記載の配列データ解析装置。

【請求項7】

請求項1ないし請求項6のいずれか1項に記載の配列データ解析装置と、前記サンプルDNA断片を両端からそれぞれシーケンシングし、その結果である前記左方配列と前記右方配列とのペアのリード配列を、前記配列データ解析装置に送信するシーケンサとを含めて構成される

DNA解析システム。

【請求項8】

配列データ解析装置は、リード辞書作成部と、クエリ検索部と、サンプル復元部と、マッピング部とを有しており、

前記リード辞書作成部は、サンプルDNA断片の両端からそれぞれシーケンシングされたペアである左方配列と右方配列との間を結合文字で連結した結合文字列を前記サンプルDNA断片ごとに作成し、各前記サンプルDNA断片の前記結合文字列を終端文字で結合した文字列をもとにリード配列辞書を作成し、

前記クエリ検索部は、ゲノムDNAの塩基文字列から生成されるクエリ配列が出現する前記リード配列辞書内の塩基文字座標であるヒット位置を検索し、

前記サンプル復元部は、

前記リード配列辞書内の前記ヒット位置を起点として、その周囲に位置する前記終端文字が出現するまでの文字列をサンプル配列として抽出し、

10

20

30

40

50

前記サンプル配列内の前記ヒット位置を起点として、その周囲に位置する前記結合文字を検査し、検査した前記結合文字から前記ヒット位置が存在しない側の前記終端文字が出現するまでの前記左方配列または前記右方配列をメイト配列として抽出し、

前記マッピング部は、前記メイト配列が出現する前記ゲノムDNA内の塩基文字列の塩基文字座標を検索することを特徴とする

配列データ解析方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、配列データ解析装置、DNA解析システムおよび配列データ解析方法に関する。

10

【背景技術】

【0002】

ゲノムDNAは、その塩基配列がDNA全体ですでにシーケンシングされ（読み取られ）、その塩基文字列はインターネット上のサーバなどに公開されている。研究者は、そのゲノムDNAをお手本（参照データ）として、シーケンサ装置が読み取った被験者のサンプルDNA断片との塩基位置の照合（ゲノムマッピング）を行うことにより、サンプルDNA内の変異を検出する。変異とは、例えば、一塩基多型（SNP：Single Nucleotide Polymorphism）や、SV（構造変異）などのゲノムDNAの塩基文字列とサンプルDNA断片の塩基文字列との相違箇所である。なお、サンプルDNA断片は、1本のサンプルDNAがシーケンサ装置内のフラグメント処理により多数に断片化されたものである。

20

【0003】

ここで、シーケンサ装置は限られた長さのDNA配列しか読取ることができないため、ほぼ一定の長さに揃えられたサンプルDNA断片それぞれの両端から限られた長さ（サンプルDNA断片の一部だけ）を読み取った2本の（ペアの）リード配列を扱うペアドエンド法が知られている。つまり、ペアドエンド法では、サンプルDNA断片の中央部に、どちらのペアのリード配列にも属さない、シーケンシングされない区間が存在することになる。

よって、ペアドエンド法で読み取られたペアのリード配列を、ゲノムDNA内の塩基位置にゲノムマッピングするときには、ペアのうちの片方のリード配列だけを位置照合するよりも、もう片方のリード配列も併せて位置照合することにより、1本のサンプルDNA断片を高精度にゲノムDNA内にマッピングできる。

30

【0004】

ゲノムDNAやサンプルDNA断片の塩基文字列のデータ量が膨大であるため、以上説明したような、ペアドエンド法による2本の（ペアの）リード配列をゲノムDNA内にマッピングする際には、処理の効率化やデータの圧縮化が求められる。そこで、以下に示すように、様々なゲノムマッピングの効率化技術が提案されている。

【0005】

ゲノムワイド・トランスクリプトーム・プロファイリングでは、全長cDNA上での5'末端から3'末端に向かう順序が維持されたペアエンド・ダイタグのシーケンシングが行われ、指定されたスパーサ配列に隣接するダイタグ配列を効率的に検出する方法が知られている（特許文献1）。

40

【0006】

新型DNAシーケンサから得られる大量のショート・リードを参照ゲノム配列に高速かつ高精度にアラインメントするために、ゲノム配列をBW（Burrows-Wheeler）変換したデータを作成し、これを利用して、ショート・リードの先頭の数十塩基のシード配列と一致する配列が参照ゲノム内に出現する位置を高速に同定する技術は、一般に広く用いられている（例えば、非特許文献1）。

【0007】

また、このとき、アラインメントできる候補位置が複数あるという曖昧性を減らすため

50

に、ペアのショート・リードのマッピングも広く行われている（例えば、非特許文献1）。通常、2つのショート・リードがペアにあることを示すため、それらには同一の名前（例えば、非特許文献1では、QNAME）が与えられる。

#### 【0008】

変異解析を行うために、大量のショート・リードを参照ゲノムにマッピングして得られる大量のデータを利用する方法が、一般に広く利用されている（例えば、非特許文献3）。また、大きな計算コストがかかるマッピングの計算を行わずに変異解析を行うために、参照ゲノム配列のBW変換とショート・リード配列のBW変換を利用するアプローチが考えられる（特許文献2、非特許文献4）。

#### 【先行技術文献】

#### 【特許文献】

#### 【0009】

【特許文献1】特表2008-547080号公報

【特許文献2】特願2013-038919号公報

#### 【非特許文献】

#### 【0010】

【非特許文献1】Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup: The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9 (2009).

【非特許文献2】Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60 (2009).

【非特許文献3】M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43(5): 491-498 (2011).

【非特許文献4】K. Kimura and A. Koike: A new approach to DNA sequence variation analysis using Burrows-Wheeler transform of massive short-read data, in Proceedings of Advances in Genome Biology & Technology Conference (AGBT) 2013, Marco Island, p.202.

#### 【発明の概要】

#### 【発明が解決しようとする課題】

#### 【0011】

ペアエンド法でシーケンスされた大量のリード配列データに対して、ペア関係の情報を計算機で扱うためには、ペアの片方のリード配列にもう片方のリード配列を関連付けるための識別子（またはポインタ情報）が必要となる。これらの識別子は、ペアの相手を一意に特定できなければならないため、多数のビット数を必要とする。そのため、識別子のデータサイズが大きくなり、計算効率の低下を招く要因となる。例えば、長さ100塩基のショート・リードが80億本ある場合は40億通りの識別子が必要となり、識別子当たり少なくとも4バイト必要で、全リードに対する識別子データのサイズは4GB（ギガバイト）にも達する。

#### 【0012】

そこで、本発明は、サンプルDNA断片の両端からシーケンシングされるペアの片方のリード配列から、もう片方のリード配列を効率的に求めることを、主な課題とする。

#### 【課題を解決するための手段】

#### 【0013】

前記課題を解決するために、本発明の配列データ解析装置は、

サンプルDNA断片の両端からそれぞれシーケンシングされたペアである左方配列と右方配列との間を結合文字で連結した結合文字列を前記サンプルDNA断片ごとに作成し、各前記サンプルDNA断片の前記結合文字列を終端文字で結合した文字列をもとにリード配列辞書を作成するリード辞書作成部と、

ゲノムDNAの塩基文字列から生成されるクエリ配列が出現する前記リード配列辞書内の塩基文字座標であるヒット位置を検索するクエリ検索部と、

前記リード配列辞書内の前記ヒット位置を起点として、その周囲に位置する前記終端文字が出現するまでの文字列をサンプル配列として抽出し、

前記サンプル配列内の前記ヒット位置を起点として、その周囲に位置する前記結合文字を検査し、検査した前記結合文字から前記ヒット位置が存在しない側の前記終端文字が出現するまでの前記左方配列または前記右方配列をメイト配列として抽出するサンプル復元部と、

前記メイト配列が出現する前記ゲノムDNA内の塩基文字列の塩基文字座標を検索するマッピング部と、を有することを特徴とする。

10

その他の手段は、後記する。

【発明の効果】

【0014】

本発明によれば、サンプルDNA断片の両端からシーケンシングされるペアの片方のリード配列から、もう片方のリード配列を効率的に求めることができる。

【図面の簡単な説明】

【0015】

【図1】本発明の一実施形態に関するDNA解析システムを示す構成図である。

【図2】本発明の一実施形態に関する配列データ解析装置を示す構成図である。

【図3】本発明の一実施形態に関する各辞書の作成処理を示すフロー説明図である。

20

【図4】本発明の一実施形態に関する辞書作成処理におけるBW文字列の作成処理を示す説明図である。図4(a)は、リード配列から結合文字列を作成する処理を示す。図4(b)は、結合文字列からBW文字列を作成する処理を示す。

【図5】本発明の一実施形態に関する辞書作成処理におけるWavelet Tree形式への変換処理を示す説明図である。図5(a)は、Wavelet Tree形式に変換するために参照される二分木を示す。図5(b)は、BW文字列をWavelet Tree形式に変換する処理を示す。

【図6】本発明の一実施形態に関するSNPの解析処理を示すフロー説明図である。

【図7】本発明の一実施形態に関するSNPの解析処理における各データを示す説明図である。図7(a)は、SNP情報を示す。図7(b)は、SNP情報とゲノム配列辞書とから作成されるクエリ配列を示す。図7(c)は、リード配列辞書からクエリ配列を検索する様子を示す。図7(d)は、SNPの解析結果を示す。

30

【図8】本発明の一実施形態に関する構造変異の解析処理を示すフロー説明図である。

【図9】本発明の一実施形態に関する構造変異の解析処理における各データを示す説明図である。図9(a)は、構造変異情報を示す。図9(b)は、構造変異情報とゲノム配列辞書とから作成されるクエリ配列を示す。図9(c)は、リード配列辞書からクエリ配列を検索する様子を示す。図9(d)は、構造変異の解析結果を示す。

【発明を実施するための形態】

【0016】

以下、本発明の一実施形態を、図面を参照して詳細に説明する。

【0017】

40

図1は、DNA解析システムを示す構成図である。

配列データ解析装置1は、通常の計算機の構成を有するサーバ等のコンピュータで実現される。

配列データ解析装置1は、中央処理部(CPU: Central Processing Unit)201、プログラムなどが記憶される記憶部であるメモリ202、操作のためのGUI(Graphical User Interface)や、解析結果などを表示する表示部203、配列辞書(図2のリード配列辞書14、ゲノム配列辞書15)などを記憶する記憶部として機能するハードディスクドライブ(HDD)204、SNPなどの変異情報やパラメータ入力などを行うキーボード等の入力部205、インターネット等に接続するためのネットワークインタフェース(NIF)206がバス207に接続された構成を備えている。

50

H D D 2 0 4 に記憶される配列辞書は、配列データ解析装置 1 の外部に設置された記憶装置に記憶してもよいし、ネットワークを介してデータセンタなどに記憶してもよい。

以下で説明される各種のフローチャートは、C P U 2 0 1 のプログラム実行などで実現される。

#### 【0018】

配列データ解析装置 1 の N I F 2 0 6 には、ゲノムサーバ 8 とシーケンサ 9 とがネットワークを介して接続されている。

シーケンサ 9 は、各サンプル D N A 断片に対して、その両端（5' 末側のリード配列と、3' 末側のリード配列）のペアをシーケンシングし（読み取り）、その結果を配列データ解析装置 1 に提供する。

なお、リード配列（塩基配列）の表記法として、5' 末側の塩基文字を左側に記載し、3' 末側の塩基文字を右側に記載する方法が一般的であるため、以下では、5' 末側を「左方」とし、3' 末側を「右方」とする。

シーケンサ 9 は、超並列型（いわゆる次世代型）D N A シーケンサとして構成され、多数の（例えば、1 億本の）サンプル D N A 断片を並列にシーケンシングすることができる。

同様に、ゲノムサーバ 8 は、ゲノム D N A をシーケンシングした結果であるゲノム配列を、配列データ解析装置 1 に提供する。

#### 【0019】

なお、ゲノム D N A をシーケンシングしたゲノム配列を参照配列として、サンプル D N A 断片を解析対象とする代わりに、m R N A 配列を参照配列として、c D N A サンプルのペアド・エンド・シーケンシング・データを解析対象とすることにより、スプライス・バリエーションの検出を行うことができる。スプライス・バリエーションによりエクソンが失われることは、構造変異の「欠失」に相当し、新たなエクソンが取りこまれることは構造変異の「挿入」に相当するからである。

#### 【0020】

図 2 は、配列データ解析装置 1 を示す構成図である。この図 2 の説明では、各構成要素の概要を説明し、各構成要素の詳細については、後記する説明で明らかにする。

配列データ解析装置 1 は、シーケンサ 9 からリード配列集合 1 1 （左方配列 1 1 a と右方配列 1 1 b ）の入力を受け付ける。

リード配列集合 1 1 は、シーケンサ 9 にシーケンシングされたサンプル D N A 断片ごとのリード配列（左方配列 1 1 a と右方配列 1 1 b ）の集合である。

左方配列 1 1 a は、サンプル D N A 断片のうちの 5' 末側の端点を基点として、3' 末側に向かってシーケンシングされたリード配列である。

右方配列 1 1 b は、サンプル D N A 断片のうちの 3' 末側の端点を基点として、5' 末側に向かってシーケンシングされたリード配列である。

ここで、左方配列 1 1 a および右方配列 1 1 b の長さは、例えば 1 0 0 塩基程度であり、サンプル D N A 断片が 3 0 0 塩基程度であるとする、中央の 1 0 0 塩基程度は左方配列 1 1 a にも右方配列 1 1 b にも含まれないシーケンシング対象外の箇所である。または、2 万塩基程度の長さのサンプル D N A 断片の場合は、シーケンシング対象外の箇所は 1 9 8 0 0 塩基程度となる。

#### 【0021】

D N A 解析処理の前準備として、リード辞書作成部 2 1 は、リード配列集合 1 1 からリード配列辞書 1 4 を作成する。ゲノム辞書作成部 2 2 は、参照ゲノム配列 1 2 からゲノム配列辞書 1 5 を作成する。

なお、参照ゲノム配列 1 2 は、解析の対象とする生物種ごとに定められており、各染色体の全長の配列の集合である。

#### 【0022】

配列データ解析装置 1 は、今回の解析対象となる変異を示す解析用情報 1 3 として、S N P 情報 1 3 a または構造変異情報 1 3 b の入力を受け付ける。S N P とは、特定箇所に

10

20

30

40

50

におけるサンプルDNA断片の塩基内容がゲノムDNAの同じ箇所の塩基内容と異なることである。構造変異とは、連続した複数の塩基の並びの挿入や欠失である。

#### 【0023】

クエリ作成部23は、ゲノム配列辞書15を参照して、解析用情報13が示す変異を含むクエリ配列16を作成する。

クエリ検索部24は、リード配列辞書14内から、クエリ配列16にマッピングする（クエリ配列16が出現する）塩基位置座標であるヒット位置17aを検索する。

サンプル復元部25は、ヒット位置17aを含むサンプルDNA断片のリード配列（サンプル配列17）を復元する。ここで、ヒット位置17aは、左方配列11aか、右方配列11bかのいずれかに含まれるので、ヒット位置17aが含まれない片方のリード配列を、メイト配列17bとする。

10

#### 【0024】

マッピング部26は、ゲノム配列辞書15を参照して、サンプルDNA断片のメイト配列17bが由来するゲノムDNA内の位置を特定（ゲノム・マッピング）する。

サンプル判定部27は、メイト配列17bのマッピングの成否により、メイト配列17aのサンプルDNA断片が解析用情報13が示す変異を含むか否かを判定する。そして、サンプル判定部27は、解析用情報13の判定結果を出力する。

#### 【0025】

図3は、各辞書の作成処理を示すフロー説明図である。まず、リード辞書作成部21が、リード配列集合11からリード配列辞書14を作成するリード辞書作成処理（S101～S105）を説明する。

20

S101において、配列データ解析装置1は、シーケンサ9からリード配列集合11（左方配列11aと右方配列11b）の入力を受け付ける。

図4（a）では、説明をわかりやすくするために、サンプルDNA断片301、305の2本を例示し、サンプルDNA断片301の左方配列11a「GA」2塩基分、右方配列11b「T」1塩基分とし、サンプルDNA断片305の左方配列11a「C」2塩基分、右方配列11b「TA」2塩基分とする。

#### 【0026】

シーケンサ9は、読み取ったリード配列集合11を、左方配列11aを示す符号361と、右方配列11bを示す符号362のように、FASTQ形式で配列データ解析装置1に通知する。符号361、符号362では、2つのサンプルDNA断片301、305のリード配列を列挙しており、1つのリード配列に対して4行の情報が記載される。

30

FASTQ形式の1行目「@seq1, @seq2」は、サンプルDNA断片の識別子（ID）であり、2行目「GA, T, C, TA」は、リード配列である。例えば、符号361の1行目「@seq1」と、符号362の1行目「@seq1」とが一致するので、同じサンプルDNA断片301から読み取られたペアであることがわかる。

#### 【0027】

図3のS102において、リード辞書作成部21は、ペアをなす左方配列11aの塩基文字と、右方配列11bの塩基文字とを結合文字「&」で結合することで、1本のサンプルDNA断片を示す結合文字列を作成する。なお、塩基文字とは、4種類の塩基それぞれを示す「A、C、G、T」と、不明の塩基を示す「N」である。

40

図4（a）では、同じサンプルDNA断片301のペアである左方配列11a「GA」と右方配列11b「T」とを結合文字「&」302で結合して、末尾に終端文字「\$」303を付加することで、結合文字列304「GA&T\$」を得る。同様に、サンプルDNA断片305のペアからは、結合文字列306「C&TA\$」を得る。

なお、サンプルDNA断片の長さに応じて、複数種類の結合文字を使い分けてもよい。例えば、約300塩基のサンプルDNA断片から生成する結合文字列内の結合文字には「&」を用い、約20000塩基のサンプルDNA断片から生成する結合文字列内の結合文字には「#」を用いてもよい。これにより、結合文字からサンプルDNA断片のペア相手だけではなく、サンプルDNA断片の長さを取得することができる。

50

## 【0028】

図3のS103において、リード辞書作成部21は、結合文字列304、306をBW変換してBW文字列311を作成する。

図4(b)では、例えば、以下の手順により、BW文字列311が作成される。この計算過程で、先頭から1文字ずつ比較して文字列比較を行う際、\$どうしが比較されたら比較を終了し、&どうしが比較されたら比較を継続する。

(手順1) 結合文字列304を巡回シフト(cyclic shift)して文字列のリスト307を得るとともに、結合文字列306も巡回シフトして文字列のリスト308を得る。

(手順2) 2つのリスト307、308をマージすることで、マージ済みリスト309を得る。

(手順3) マージ済みリスト309をアルファベット順にソートすると、ソート済みリスト310を得る。そのとき、文字のソート順位は、例えば、「\$<#<&<A<C<G<T<N」とする。

(手順4) ソート済みリスト310の各行の末端の文字を連結して、BW文字列311を得る。

このようにして得られたBW文字列311は、ソート済みなので、同じ文字が連続する頻度が高い。よって、BW文字列311をランレングス圧縮することで、データ量を圧縮することができる。

## 【0029】

図3のS104において、リード辞書作成部21は、BW文字列311をWavelet Tree形式に変換することにより、検索を効率的に行うためのリード配列辞書14を作成する。

図5(a)は、Wavelet Tree形式に変換するために参照される二分木320を示す。この二分木320では、文字列に使用される全文字(\$、&、A、C、G、T、N)321が根となる。

二分木320は、文字列に使用される全文字(\$、&、A、C、G、T、N)321を、再帰的に二分類し、分類の末端で高々2種類の文字しか含まないように分類する方法を示した二分木である。

二分木320の根では、全文字321を、Aと、T(W)324と、それ以外(S)325と、に分類する。Sに分類された文字(\$、&、C、G、N)325も、同様にMとKに二分類する。以下、これを再帰的に繰り返し、分類の末端で高々2種類の文字しか含まないようにする。但し、ペアを表す符号&(302)と終端記号\$(303)は、分類の末端331と一緒に現れるように分類する。

## 【0030】

図5(b)のWavelet Tree340は、BW文字列311を示す二分木である。Wavelet Tree340の根は、二分木320のWとSの二分類に従って、BW文字列311をバイナリ文字列341に変換したものである。リード辞書作成部21は、Wに分類される文字を抜き出した部分文字列342、および、Sに分類される文字を抜き出した部分文字列343を作成する。

リード辞書作成部21は、(AとTの2種類の文字しか含まない)部分文字列342を、0と1のバイナリ文字列344に変換する。

リード辞書作成部21は、部分文字列343も同様に、Sに分類された文字のMとKへの二分類(二分木320で示される)に従ってバイナリ文字列345を作成する。

リード辞書作成部21は、Mに分類される文字を抜き出した部分文字列346、および、Kに分類される文字を抜き出した部分文字列347を作成する。

## 【0031】

以上説明した各部分文字列を作成する処理を繰り返すことにより、図示した6本のバイナリ文字列が得られる。分解の末端351では、ペアを表す符号&(302)と終端記号\$(303)からなる文字列350がバイナリ文字列351で表される。バイナリ文字列351の長さ(ビット数)はリードの総数(ペアの総数の2倍)に等しい。

よって、Wavelet Tree340は、BW文字列311を可逆変換したものであり、Wavele

10

20

30

40

50

t Tree 3 4 0 から BW 文字列 3 1 1 を復元することができる。

#### 【0032】

図3のS105において、リード辞書作成部21は、Wavelet Tree 3 4 0 から作成されたリード配列辞書14を記憶手段に出力する。ここで、リード辞書作成部21は、Wavelet Tree 3 4 0に加えて、そのWavelet Tree 3 4 0のrank関数やselect関数を計算するための、わずかな（BW文字列3 1 1に対して3.5%程度のデータ量の）補助データを、リード配列辞書14に付加してもよい。

rank(p,c)とは、配列要素0～pのうちの文字「c」の出現回数を返す関数である。

select(i,c)とは、(i+1)番目の文字「c」が出現する配列位置を返す関数である。

補助データは、例えば、参考文献「Kouichi Kimura, Yutaka Suzuki, Sumio Sugano, and Asako Koike. Journal of Computational Biology. November 2009, 16(11): 1601-1613.」に記載の「hierarchical binary string」である。この補助データは、BW文字列3 1 1から、任意に与えられた塩基配列と一致するリード配列断片を全て求める、などの検索を効率的に行うためのデータである。

#### 【0033】

BW変換された文字列上でrank関数とselect関数を効率的に計算するための補助データが与えられているため、\$と&の2種類の区切り文字を用いたBW変換の特徴を利用して、任意に与えられた文字列sに対して、sに一致する全てのリード断片配列を効率的に求めることができ、また、さらに、各リード断片配列に対して、その左方と右方にある配列を（\$が現れるまで延長して）復元する計算を効率的に行える。

#### 【0034】

以上、図3～図5を参照して、リード辞書作成処理（S101～S105）を説明した。一方のゲノム辞書作成処理（図3のS101b～S105b）もリード辞書作成処理（S101～S105）と同様に作成できる。

S101bでは、ゲノム辞書作成部22は、S101のリード配列集合11の代わりに参照ゲノム配列12の入力を受け付ける。

S102bでは、ゲノム辞書作成部22は、参照ゲノム配列12が示す複数のゲノムDNAの染色体配列（塩基文字列）を、そのまま終端文字「\$」で連結して、1本の文字列を作成する。ここで、参照ゲノム配列12はペア形式ではないので、S102のような結合文字「&」によるペア連結処理は不要である。

S105bでは、ゲノム辞書作成部22は、S105のリード配列辞書14の代わりにゲノム配列辞書15を出力する。

#### 【0035】

図6は、SNPの解析処理を示すフロー説明図である。

S121において、配列データ解析装置1は、今回の解析対象となる変異を示す解析用情報13として、SNP情報13aの入力を受け付ける。

図7(a)では、SNP情報13aを示すテーブル400を例示する。テーブル400の各行に示すように、SNPごとに、染色体名、染色体上の塩基位置座標、参照ゲノム配列内での塩基の種類（標準塩基）、SNPとして現れる塩基の種類（変異塩基）の情報を含む。

テーブル400の1行目は、SNPが染色体7番上の123456塩基目の位置にあり、参照ゲノムの塩基「A」が塩基「G」に変異することを示す。

#### 【0036】

図6のS122において、クエリ作成部23は、ゲノム配列辞書15を参照して、SNP情報13aが示すSNPを含むクエリ配列16を作成する。

図7(b)の説明欄420は、S122の例示であり、横軸421は染色体上の塩基位置座標である。まず、クエリ作成部23は、ゲノム配列辞書15を参照して、テーブル400の1行目で示されたSNPの位置424の周辺（例えば、左右に10塩基程度）の塩基配列422を求める。クエリ作成部23は、SNPの位置424で塩基配列422の塩基を変異させた配列423を作成し、これをクエリ配列16とする。または、クエリ作成

部 2 3 は、変異を含む塩基配列 4 2 3 の代わりに、変異を含まない塩基配列 4 2 2 をクエリ配列 1 6 とすることで、S N P の位置 4 2 4 に現れる標準塩基を検出することができる。

#### 【0037】

図 6 の S 1 2 3 において、クエリ検索部 2 4 は、リード配列辞書 1 4 内から、クエリ配列 1 6 にマッピングする（塩基文字列の並びが一致する）塩基位置座標であるヒット位置 1 7 a を検索する。

なお、BW 変換されたリード配列辞書 1 4 をもとに計算する過程で、2 本の文字列（リード配列辞書 1 4、クエリ配列 1 6）を先頭から 1 文字ずつ比較する際には、終端文字「\$」どうしが比較される場合は比較を終了し、結合文字列「&」どうしが比較される場合は比較を継続する。

#### 【0038】

S 1 3 1 ~ S 1 3 9 において、配列データ解析装置 1 は、S 1 2 3 で求めたヒット位置 1 7 a ごとのループ処理を実行する。

S 1 3 2 において、サンプル復元部 2 5 は、S 1 3 1 からのループで現在選択しているヒット位置 1 7 a を含むサンプル D N A 断片のリード配列（サンプル配列 1 7）を復元する。この復元処理では、BW 変換により作成されているリード配列辞書 1 4 から、rank 関数と select 関数を用いることにより、ヒット位置 1 7 a を起点として、終端文字「\$」が出現するまでリード配列辞書 1 4 内を延長して走査することにより、左右の終端文字「\$」に挟まれたサンプル配列 1 7 が取得される。

このような rank 関数と select 関数を用いる方法は、例えば、文献「Ferragina, P. and Manzini, G 著, "Opportunistic data structures with applications", In 41st IEEE Symposium on Foundations of Computer Science, FOCS, pages 390-398」に記載されている。

サンプル配列 1 7 は、結合文字列 3 0 4, 3 0 6 のように、結合文字「&」および終端文字「\$」を含む。このサンプル配列 1 7 を結合文字列「&」で分離することで、サンプル D N A 断片ごとの識別子を使わずに、ペアをなす 2 本のリード配列（左方配列 1 1 a、右方配列 1 1 b）が得られる。

#### 【0039】

S 1 3 3 において、サンプル復元部 2 5 は、以下に示すように、サンプル配列 1 7 から結合文字「&」が見つかるまで走査（検査）することで、メイト配列 1 7 b を取得する（S 1 3 4）。

図 7（c）の [1] の場合のように、ヒット位置 1 7 a の配列 4 2 3 から右方に延長した先に結合文字「&」が現れるので、取得されるメイト配列 1 7 b は、結合文字「&」より右側の右方配列 1 1 b である。

図 7（c）の [2] の場合のように、ヒット位置 1 7 a の配列 4 2 3 から左方に延長した先に結合文字「&」が現れるので、取得されるメイト配列 1 7 b は、結合文字「&」より左側の左方配列 1 1 a である。

図 7（c）の [3] の場合のように、ヒット位置 1 7 a の配列 4 2 3 から右方に延長しても、左方に延長しても結合文字「&」が現れない場合は、サンプル配列 1 7 はペアを構成する相手がいない単独のリード配列である。このような単独のリード配列は信頼性が低いとみなして無視してよい。

あるいは、変異を導入する前の配列 4 2 3 をゲノム配列辞書 1 5 に問い合せて、ゲノム内で配列 4 2 3 が出現する箇所が一つしかないと確認できた場合に限り、（信頼性は低い）S N P を検出できたと判定してもよい。

#### 【0040】

S 1 3 5 において、マッピング部 2 6 は、ゲノム配列辞書 1 5 を参照して、S 1 3 4 で取得したメイト配列 1 7 b が由来するゲノム D N A 内の位置を特定（ゲノム・マッピング）する。マッピング部 2 6 は、例えば、メイト配列 1 7 bの中から短い（例えば、20 塩基程度の）部分配列を切り出して、その部分配列がゲノム内に 1 か所だけ現れるか否かをゲノム配列辞書 1 5 に問い合せる。

10

20

30

40

50

もし、1か所も現れない場合は、マッピング部26は、その部分配列中にシーケンシング・エラーや多型が含まれていると考えられるため、別の部分配列で再度、問い合わせを行う。

また、もし、複数箇所に現れる場合は、マッピング部26は、部分配列の長さを増やすか、または、別の部分配列を用いて、再度、問い合わせを行う。こうして、短い部分配列のゲノム内の位置を特定できた場合、それを含むメイト配列17bのゲノム上に位置も特定できる（換言すると、マッピングに成功する）。一方、位置を特定できない場合は、マッピングに失敗する。

S135でYesならS136に進み、Noなら今回のループを終了して（S139）、次のヒット位置を選択するためにS131に戻る。

10

#### 【0041】

S136において、サンプル判定部27は、S132の配列17aのサンプルDNA断片が解析用情報13で示されるSNPの変異を含むか否かを、以下で説明するように距離が正常（整合）かによって判定する。例えば、サンプル判定部27は、マッピングが成功したメイト配列17bのマッピング位置と、SNP情報13aが示すSNPの位置（ヒット位置17a）との距離が、サンプルDNA断片の長さにほぼ等しければ、サンプル配列17を構成する左方配列11aと右方配列11bとが整合したとして、「SNP検出」（つまり、SNP情報13aに整合）と判定する。

S136でYesならS137に進み、NoならS138に進む。

#### 【0042】

20

S137において、サンプル判定部27は、「SNP検出」と判定されたSNP情報13aの検出数カウンタ値を1つ増やす。

S138において、サンプル判定部27は、「SNP非検出」と判定されたことにより、検出数カウンタ値の増加を行わない。

#### 【0043】

S131～S139のループを実行した後、S141において、サンプル判定部27は、SNP情報13aと、その検出数カウンタ値（SNPが検出されたサンプルDNA断片の個数）とを対応づけた解析用情報13の判定結果を出力（ユーザに報告）する。

図7（d）のテーブル460は、S141で出力される情報の一例である。テーブル460には、図7（a）のテーブル400で示される各SNPごとに、SNP位置424で変異塩基（SNP）が検出されたリード断片の数（変異塩基検出数）、および、標準塩基が検出されたリード断片の数（標準塩基検出数）が、検出数カウンタ値から読み取られて、書き込まれている。

30

#### 【0044】

図8は、構造変異の解析処理を示すフロー説明図である。図6のSNPの解析処理との差分に着目して、以下で図8を説明する。

S121bでは、配列データ解析装置1は、今回の解析対象となる変異を示す解析用情報13として、SNP情報13aの代わりに、構造変異情報13bの入力を受け付ける。

図9（a）の構造変異情報13bは、テーブル600の各行に示すように、各変異ごとに、染色体名、染色体上の塩基位置座標、変異のタイプ（挿入か欠失か）、変異長の情報を含む。

40

テーブル600の1行目は、構造変異が染色体3番上の654321塩基目の位置にあり、標準ゲノムから500個の連続した塩基の並びが失われる欠失が生ずることを示す。

#### 【0045】

図8のS122bでは、クエリ作成部23は、ゲノム配列辞書15を参照して、構造変異情報13bが示す構造変異周辺のクエリ配列16を作成する。

図9（b）の説明欄620は、S122bのクエリ配列16の作成方法を示す。横軸421は染色体上の塩基位置座標である。まず、ゲノム配列辞書15を参照して、構造変異が生ずる位置624の周辺（例えば、左方と右方に数十塩基程度離れた位置）の短い（例えば20塩基程度の）塩基配列622と623を求め、これらをそれぞれクエリ配列16

50

とする。つまり、クエリ配列 1 6 は構造変異が生ずる位置 6 2 4 の（左方または右方の）近くにある短い配列である。

#### 【0046】

図 8 の S 1 3 3 b では、S 1 3 3 の処理（図 8 では図示省略）の実行後に、構造変異解析の対象外（変異情報が得られない）か否かを判定する。S 1 3 3 b で Y e s なら今回のループを終了し（S 1 3 9）、N o なら S 1 3 4 へ進む。以下、S 1 3 3 b の判定処理について、図 9（c）の説明欄 6 4 0 を参照して説明する。

#### 【0047】

図 9（c）の〔1〕の場合のように、ヒット位置 1 7 a である位置 6 2 4 の左方にあるクエリ配列 6 2 2 から右方に延長して復元されたサンプル配列 1 7 内に結合文字「&」が現れる場合は、クエリ配列 6 2 2 は左方配列 1 1 a に含まれる。

10

よって、結合文字「&」の右方の配列 6 4 1（右方配列 1 1 b）がメイト配列 1 7 b として復元される。さらに、クエリ配列 6 2 2 と配列 6 4 1 との間には構造変異が生じる位置 6 2 4 が含まれる可能性がある。

一方、図示は省略するが、クエリ配列 6 2 2 の左方に延長して復元された文字列内に結合文字「&」が現れる場合は、構造変異の可能性は無いので、対象外（変異情報が得られない）と判定する。

#### 【0048】

図 9（c）の〔2〕の場合のように、ヒット位置 1 7 a である位置 6 2 4 の右方にあるクエリ配列 6 2 3 に対しても、図 9（c）の〔1〕に対して左右を反転して同様な判定を行う。クエリ配列 6 2 3 から左方に延長して復元されたサンプル配列 1 7 内に結合文字「&」が現れる場合は、クエリ配列 6 2 3 は右方配列 1 1 b に含まれる。

20

よって、結合文字「&」の左方の配列 6 4 2（左方配列 1 1 a）がメイト配列 1 7 b として復元される。さらに、クエリ配列 6 2 3 と配列 6 4 2 との間には構造変異が生じる位置 6 2 4 が含まれる可能性がある。

一方、図示は省略するが、クエリ配列 6 2 3 の右方に延長して復元された文字列内に結合文字「&」が現れる場合は、構造変異の可能性は無いので、対象外（変異情報が得られない）と判定する。

以上、S 1 3 3 b の判定処理について、説明した。

#### 【0049】

30

S 1 3 6 の代わりに行われる S 1 3 6 b では、サンプル判定部 2 7 は、メイト配列 1 7 b と、その抽出元であるサンプル配列 1 7 におけるペアのリード配列とのペア間距離が、サンプル DNA 断片の長さとはほぼ等しい場合は、距離が正常値であると評価する（S 1 3 6 b, Y e s）。

S 1 3 6 b で N o なら、構造変異を検出したので、サンプル判定部 2 7 は、該当する構造変異情報 1 3 b の「変異有り検出数」のカウンタ値を 1 つ増やす（S 1 3 7 b）。

ここで、サンプル判定部 2 7 は、「欠失」の変異タイプと、「挿入」の変異タイプとで別々に検出数を集計してもよい。そのため、ペア間の距離がサンプル DNA 断片の長さから予想されるよりも長い場合は（差分の長さに対応する）欠失が生じていると判定し、逆に、ペア間の距離がサンプル DNA 断片の長さから予想されるよりも短い場合は（差分の長さに対応する）挿入が生じていると判定してもよい。

40

S 1 3 6 b で Y e s なら、構造変異は非検出なので、サンプル判定部 2 7 は、該当する構造変異情報 1 3 b の「変異無し検出数」のカウンタ値を 1 つ増やす（S 1 3 8 b）。

#### 【0050】

S 1 4 1 b では、配列データ解析装置 1 は、S 1 3 7 b および S 1 3 8 b のカウンタ値を構造変異の検出結果として報告するため、図 9（d）のテーブル 6 6 0 で示すような情報を出力する。

このテーブル 6 6 0 には、テーブル 6 0 0 で示される各構造変異ごとに、変異有りと判定されたリード断片の数、および、変異無し（即ち、標準ゲノムと一致する）と判定されたリード断片の数を、変異有り検出数、および、変異無し検出数として報告する。

50

## 【0051】

以上説明した本実施形態では、リード辞書作成部21は、サンプルDNA断片のペアである左方配列11aと右方配列11bとの間を結合文字で連結し、サンプルDNA断片のペア間を末端文字で結合することで、結合文字列を作成する。そして、リード辞書作成部21は、結合文字列をBW変換してからWavelet Tree化したリード配列辞書14を作成する。

サンプル復元部25は、リード配列辞書14内のクエリ配列16のヒット位置17aから、そのヒット位置17aを含むサンプルDNA断片のペア（メイト配列17b）を復元するときに、リード配列辞書14内に埋め込まれた結合文字を手がかりに、メイト配列17bを復元することができる。よって、サンプルDNA断片ごとの識別子は不要となるため、任意のリード配列に対してペアをなすメイト配列17bを効率良く計算できる。

10

## 【0052】

なお、サンプルDNA断片ごとの識別子を用いないことにより、以下に示すように、配列データ解析装置1が負担するデータ量および処理量を削減することができる。

比較例として、図4の符号361の「@seq1」のようにサンプルDNA断片ごとの識別子を用いる場合、サンプルDNA断片から読み取られるリード配列集合11が左方配列11aと右方配列11bとを合わせて10億本存在する場合、5億種類の識別子が必要となる。1つの識別子当たり4バイトのデータサイズとすると、全識別子データのサイズは4ギガバイトを要する。さらに、サンプルDNA断片ごとの識別子を用いる場合には、ヒットしたリード配列の識別子を検索キーとして、そのペアをなすメイト配列17bを検索するための負荷がかかってしまう。

20

## 【0053】

一方、本実施形態では、リード配列辞書14内に埋め込む制御用文字（結合文字、末端文字）のデータサイズは、1つのサンプルDNA断片あたり1ビットで済むため、全サンプルDNA断片の合計で10億ビット（＝0.125ギガバイト）となり、識別子を用いた場合のデータサイズの32分の1程度で済む。さらに、ヒットしたリード配列と、そのペアをなすメイト配列17bとは結合文字をはさんでリード配列辞書14内に隣接して配置されているため、メイト配列17bの検索負荷を低く抑えることができる。

## 【0054】

なお、本発明は前記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、前記した実施例は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。

30

また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。

また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。また、上記の各構成、機能、処理部、処理手段などは、それらの一部または全部を、例えば集積回路で設計するなどによりハードウェアで実現してもよい。

また、前記の各構成、機能などは、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。

## 【0055】

40

各機能を実現するプログラム、テーブル、ファイルなどの情報は、メモリや、ハードディスク、SSD（Solid State Drive）などの記録装置、または、IC（Integrated Circuit）カード、SDカード、DVD（Digital Versatile Disc）などの記録媒体に置くことができる。

また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線や情報線を示しているとは限らない。実際にはほとんど全ての構成が相互に接続されていると考えてもよい。

## 【符号の説明】

## 【0056】

1 配列データ解析装置

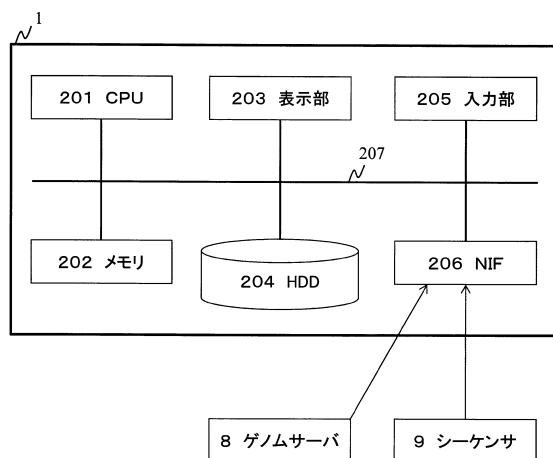
50

- 8 ゲノムサーバ
- 9 シーケンサ
- 11 リード配列集合
- 11a 左方配列
- 11b 右方配列
- 12 参照ゲノム配列
- 13 解析用情報
- 13a SNP情報
- 13b 構造変異情報
- 14 リード配列辞書
- 15 ゲノム配列辞書
- 16 クエリ配列
- 17 サンプル配列
- 17a ヒット位置
- 17b メイト配列
- 21 リード辞書作成部
- 22 ゲノム辞書作成部
- 23 クエリ作成部
- 24 クエリ検索部
- 25 サンプル復元部
- 26 マッピング部
- 27 サンプル判定部

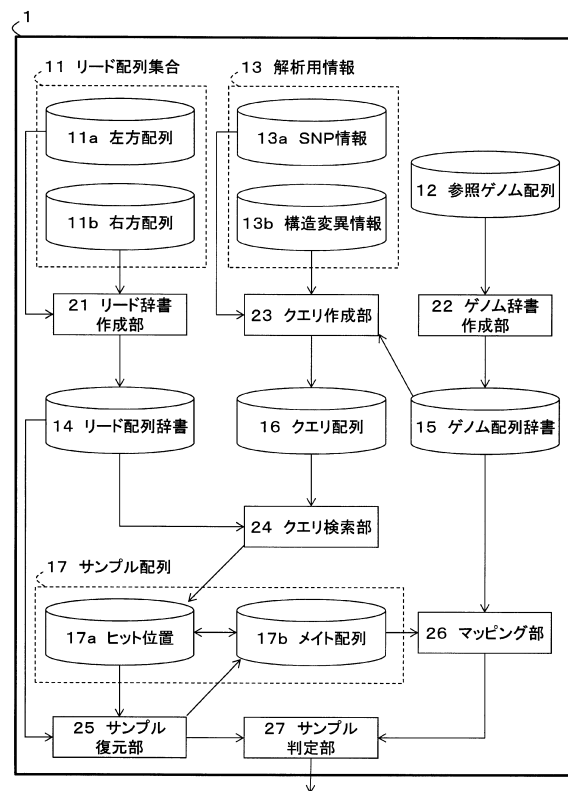
10

20

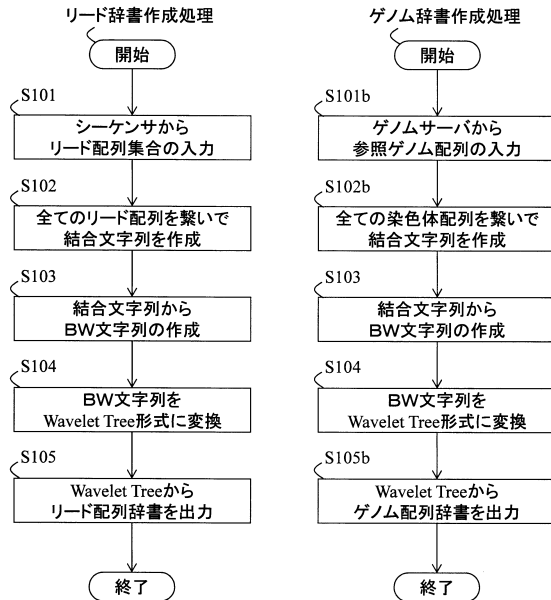
【図 1】



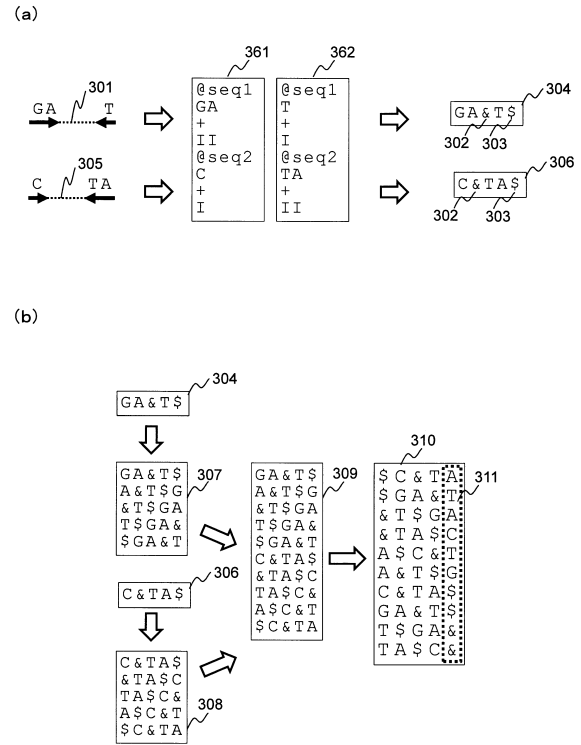
【図 2】



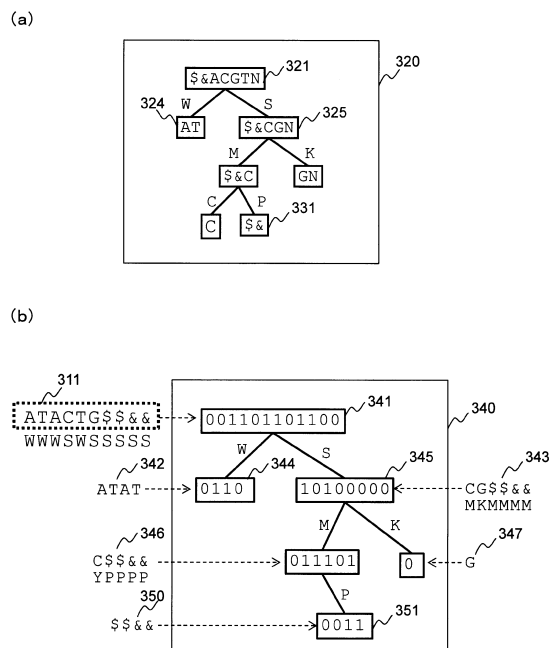
【図 3】



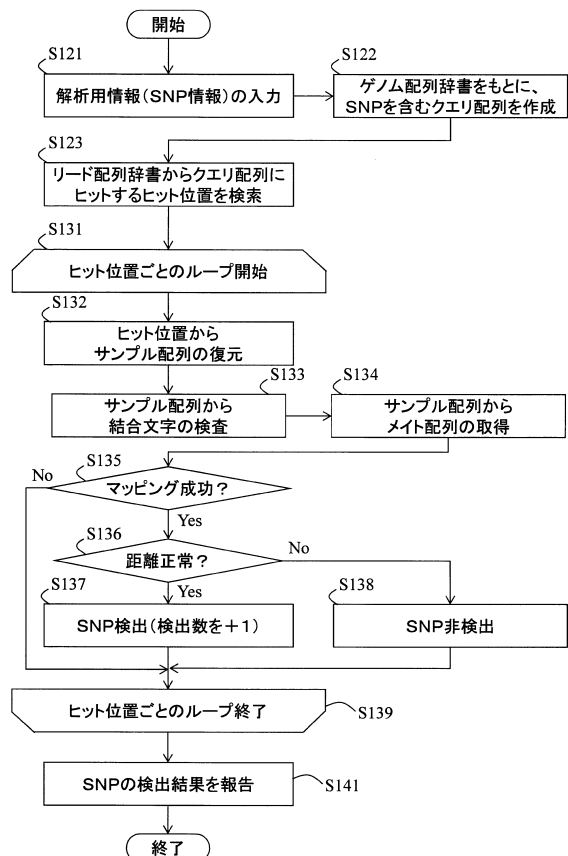
【図 4】



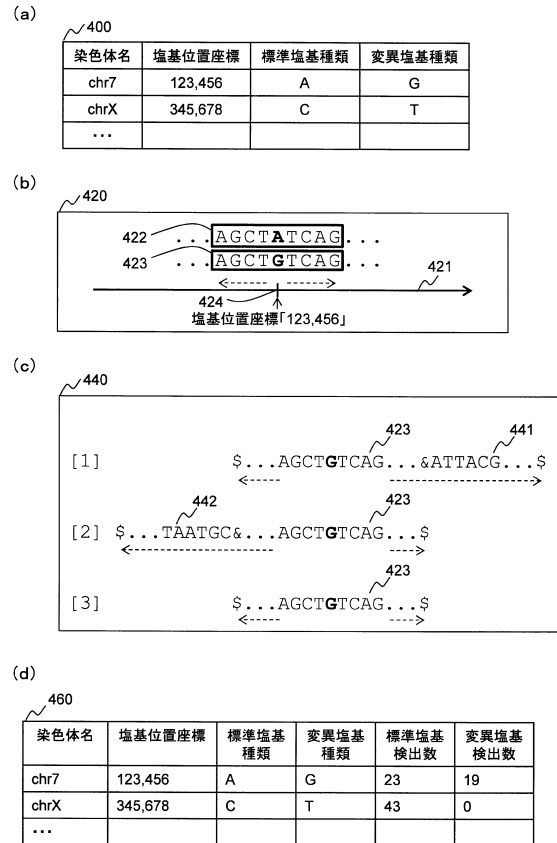
【図 5】



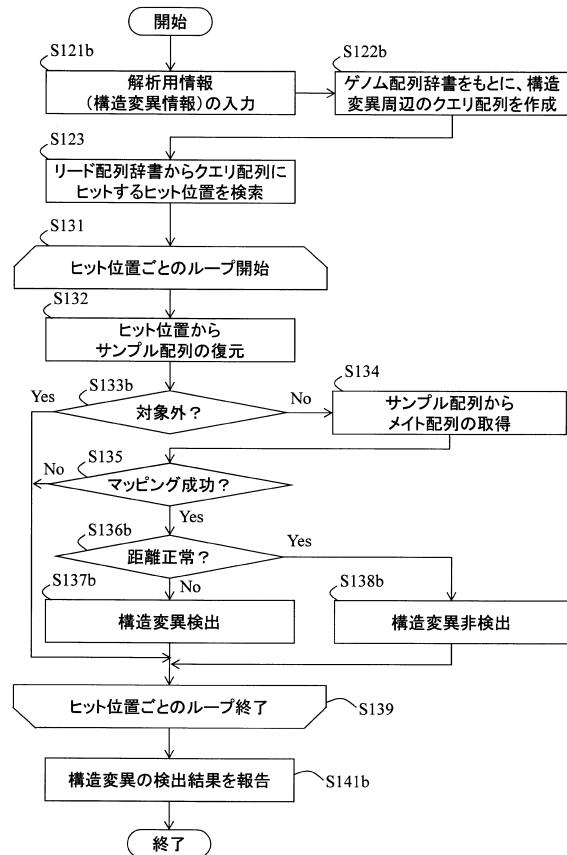
【図 6】



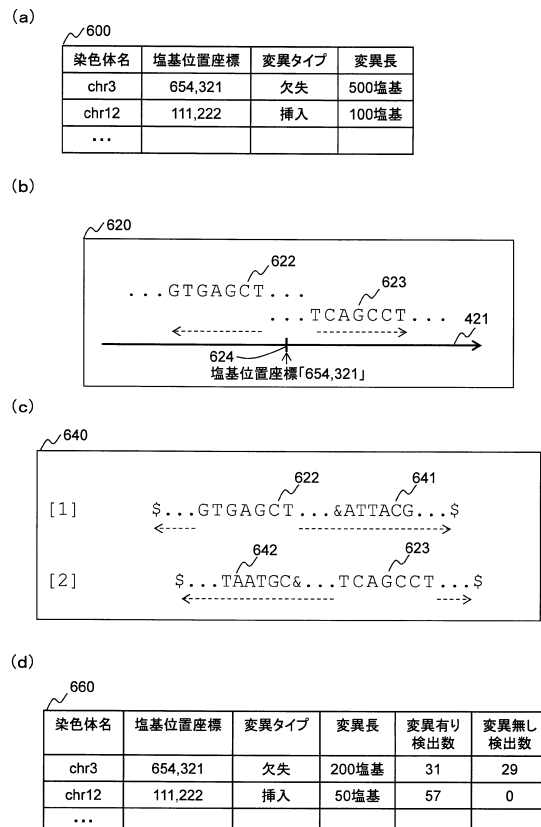
【図 7】



【図 8】



【図 9】



---

フロントページの続き

(56)参考文献 特開2009-116559 (JP, A)

特表2008-547080 (JP, A)

特表2014-502513 (JP, A)

Heng LI et al, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 2009年 7月15日, Vol.25, No.14, pp.1754-1760

Heng LI et al, The Sequence Alignment/Map format and SAMtools, Bioinformatics, 2009年 6月 8日, Vol.25, No.16, pp.2078-2079

Bouchu, Marie-Noelle; Large, Sylvie; Steng, Michaela; Langlois, Bernard; Praly, Jean-Pierre, An unprecedented access to trifluoromethylthiosugar derivatives from thiocyanate precursors upon tre, Carbohydrate Research, 1998年, 314(1-2), 37-45

(58)調査した分野(Int.Cl., DB名)

G06F 19/10-19/28

PubMed

JSTPlus/JMEDPlus/JST7580 (JDreamIII)