(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7157141号
(P7157141)

(45)発行日　令和4年10月19日(2022.10.19)

(24)登録日　令和4年10月11日(2022.10.11)

(51)国際特許分類
Ｇ１６Ｂ　50/50　(2019.01)

ＦＩ
Ｇ１６Ｂ　50/50

請求項の数　10　（全32頁）

(21)出願番号　特願2020-509515(P2020-509515)
(86)(22)出願日　平成30年8月9日(2018.8.9)
(65)公表番号　特表2020-533666(P2020-533666A)
(43)公表日　令和2年11月19日(2020.11.19)
(86)国際出願番号　PCT/IB2018/056009
(87)国際公開番号　WO2019/043481
(87)国際公開日　平成31年3月7日(2019.3.7)
審査請求日　令和3年1月22日(2021.1.22)
(31)優先権主張番号　15/693,019
(32)優先日　平成29年8月31日(2017.8.31)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
Ｎｅｗ　Ｏｒｃｈａｒｄ　Ｒｏａｄ，　Ａｒｍｏｎｋ，　Ｎｅｗ　Ｙｏｒｋ　１０５０４，　Ｕｎｉｔｅｄ　Ｓｔａｔｅｓ　ｏｆ　Ａｍｅｒｉｃａ
(74)代理人　100112690
弁理士　太佐 種一

最終頁に続く

(54)【発明の名称】　ゲノム・ファイルのためのコンテキスト・アウェア差分アルゴリズム

(57)【特許請求の範囲】

【請求項１】

複数のゲノム・データ・ファイルに対する少なくとも１つの差分ファイルを圧縮する方法であって、前記方法は、

インプットとして第1のファイル及び第2のファイルを受信するステップであって、該第1のファイル及び該第2のファイルは、前記複数のゲノム・データ・ファイルに属し、該第1のファイルの各行が第1のファイルのフィールドからなり、該第2のファイルの各行が第2のファイルのフィールドからなるように、該第1のファイル及び該第2のファイルの両方がタブ区切りであり、ここで、該第1のファイルが、第1の標本に関連付けられた参照ゲノムからの第1のデータを含み、該第2のファイルが、第1の標本と異なる、第2の標本に関連付けられた別のゲノムからの第2のデータを含む、ステップと、

前記受信された前記第1のファイル及び受信された前記第2のファイルを精査することにより、受信された第1のファイルのソートされた行及び受信された第2のファイルの対応するソートされた行を判断し、該ソートされた行及び該対応するソートされた行は、ゲノムにおけるマッピング位置の昇順に配置されるステップと、

前記ソートされた行の第1のファイルのフィールドと、前記対応するソートされた行の対応する第2のファイルのフィールドとを対比するステップと、

前記対比に基づいて複数の派生差分ファイルを生成するステップと、

汎用ファイル・コンプレッサを用いて、前記生成された複数の派生差分ファイルを圧縮するステップであって、前記複数の派生差分ファイルは、第2のファイルを再構築するた

めに第1のファイルと共に利用されるようにフォーマットされる、圧縮するステップと、
を含む、方法。
【請求項２】
前記圧縮された複数の派生差分ファイルをユーザのデバイスに格納するステップと、
前記圧縮された複数の派生差分ファイルを前記ユーザに提示するステップと、
をさらに含む、請求項１に記載の方法。
【請求項３】
前記第1のファイルは、少なくとも1つのソース・ファイルから構成され、前記第2のファイルは、少なくとも1つの目的ファイルから構成される、請求項１に記載の方法。
【請求項４】
前記受信された第1のファイルおよび前記受信された第2のファイルがソートされていないことを判断するステップと、
前記受信された第1および第2のファイル内の各行を、ソーティング・ツールを用いてソートするステップと、
をさらに含む、請求項３に記載の方法。
【請求項５】
前記受信された第1および第2のファイルが、対応可能なフォーマットであることを判断するステップ、
をさらに含む、請求項１に記載の方法。
【請求項６】
前記受信された第1のファイルのソートされた行の第1のファイルのフィールドと受信された第2のファイルのソートされた行の対応する第2のファイルのフィールドとを対比する準備として、受信された第2のファイルの対応するソートされた行の少なくとも2つの行をバッファに格納すること、
をさらに含む、請求項１に記載の方法。
【請求項７】
前記第1のファイルのフィールドがそれぞれの第1のコラムを含み、前記第2のファイルのフィールドがそれぞれの第2のコラムを含み、前記ソートされた行の第1のファイルのフィールドと、前記対応するソートされた行の対応する第2のファイルのフィールドとを対比するステップが、
前記第1のコラムと前記第2のコラムの対応するものを対比するステップと、
前記第1のコラムと前記第2のコラムの対応するものとの対比に基づいて、派生差分ファイルを生成するステップと、
をさらに含む、請求項１に記載の方法。
【請求項８】
前記第1のコラムと前記第2のコラムの対応するものを対比するステップが、
前記第1のコラムの優先コラムが前記第2のコラムの特定のコラムと一致するかどうかを判断するステップと、
前記第1のコラムの優先コラムが前記第2のコラムの特定のコラムに一致するという判断に応答して、前記第1のコラムの追加のコラムを前記第2のコラムの対応する更なるコラムと対比するステップと、
をさらに含む、請求項７に記載の方法。
【請求項９】
請求項１～８のいずれか１項に記載の前記方法の各ステップをコンピュータ・ハードウェアによる手段として構成したシステム。
【請求項１０】
請求項１～８のいずれか１項に記載の前記方法の各ステップをコンピュータに実行させる、コンピュータ・プログラム。
【発明の詳細な説明】
【技術分野】

【０００１】
本発明は、一般にコンピューティングの分野に関し、さらに具体的には計算生物学に関する。
【背景技術】
【０００２】
ゲノム解析パイプライン（パイプライン）は、生の配列リードから生物学的に意味のあるアウトプットを抽出するために、前処理、変異の発見、およびコール・セットのリファインメントなど多くのステップを伴う。パイプラインは、かかるステップ毎にアウトプット・ファイルを生成し、それらは、インプット配列リードのサイズの如何によって、メガバイト～テラバイトの間の範囲のサイズとなる。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
これらのファイルの見直しによって、あらゆるステップでこれらアウトプット・ファイルに保存された全ての情報が新規に生成されたものでないことが判明している。データのかなり大きなパーセントが、インプット・ファイルからアウトプットに単に移されているだけで、パイプラインの実行の過程で、さらにはその後でも、ストレージに対し不必要な負担を生じさせている。パイプラインの各段階は多くの時間または日数を要し得るので、これら中間的なファイルが、パイプラインにおける今後の調査、変更、または分岐のために保存される。
【課題を解決するための手段】
【０００４】
本発明の諸実施形態は、複数のゲノム・データ・ファイルに対する少なくとも１つの差分ファイルを圧縮するための方法、コンピュータ・システム、およびコンピュータ・プログラム製品を開示する。本発明は、インプットとして複数のゲノム・データ・ファイルを受信するステップを含むことができる。また、本発明は、受信された複数のゲノム・データ・ファイルを精査することによって、複数の行を判断するステップを含むことが可能である。本発明は、次いで、これら精査された複数のゲノム・データ・ファイルに関連付けられた複数の行を対比するステップを含むことができる。本発明は、対比された複数の行に基づいて、複数の派生差分ファイルを生成するステップをさらに含むことが可能である。また、本発明は、汎用ファイル・コンプレッサを用いて、生成された複数の派生差分ファイルを圧縮するステップを含むことができる。
【０００５】
本発明のこれらのおよび他の目的、特徴、および利点は、以降の本発明の例示的な諸実施形態の詳細な説明を添付の図面と併せ読むことによって明らかとなろう。これら図面中の様々な特徴は、当業者が詳細な説明と併せ本発明を理解するのを容易にする上での明瞭化のための例示である。
【図面の簡単な説明】
【０００６】
【図１】少なくとも１つの実施形態による、ネットワーク化コンピュータ環境を示す。
【図２】少なくとも１つの実施形態による、ゲノム・データ・ファイルに対する差分ファイルを圧縮するための或るプロセスを示すオペレーション・フローチャートである。
【図３】少なくとも１つの実施形態による、ゲノム・データ・ファイルに対する差分ファイルを圧縮するための典型的なプロセスを示すオペレーション・フローチャートである。
【図４】少なくとも１つの実施形態による、配列アラインメント／マップ・フォーマット（ＳＡＭ：Ｓｅｑｕｅｎｃｅ　Ａｌｉｇｎｍｅｎｔ／Ｍａｐ　ｆｏｒｍａｔ）のゲノム・データ・ファイルに対する差分ファイルを圧縮するための典型的なプロセスを示すオペレーション・フローチャートである。
【図５】少なくとも１つの実施形態による、変異コール・フォーマット（ＶＣＦ：Ｖａｒｉａｎｔ　Ｃａｌｌ　Ｆｏｒｍａｔ）のゲノム・データ・ファイルに対する差分ファイルを

圧縮するための典型的なプロセスを示すオペレーション・フローチャートである。
【図６】少なくとも１つの実施形態による、２つのゲノム・データ・ファイルを対比するための典型的なプロセスを示すブロック図である。
【図７】少なくとも１つの実施形態による、配列アラインメント／マップ・フォーマット（ＳＡＭ）のゲノム・データ・ファイルの階層構造を識別するための典型的なプロセスを示すオペレーション・フローチャートである。
【図８】少なくとも１つの実施形態による、変異コール・フォーマット（ＶＣＦ）のゲノム・データ・ファイルの階層構造を識別するための典型的なプロセスを示すオペレーション・フローチャートである。
【図９】少なくとも１つの実施形態による、図１に描かれたコンピュータおよびサーバの内部および外部のコンポーネントのブロック図である。
【図１０】本開示の或る実施形態による、図１に描かれたコンピュータ・システムを含む例示のクラウド・コンピューティング環境のブロック図である。
【図１１】本開示の或る実施形態による、図８の例示のクラウド・コンピューティング環境の機能層のブロック図である。
【発明を実施するための形態】
【０００７】
請求対象の構造体および方法の詳細な実施形態が、本明細書で開示されるが、但し、当然のことながら、これら開示される実施形態は、様々な形態で具現化が可能な請求対象の構造体および方法の単なる例示である。しかしながら、本発明は多くの異なる形態で具現化することが可能で、本明細書中で述べる例示的な実施形態に限定されると解釈されるべきではない。むしろ、これらの例示的な実施形態は、本開示が、当業者によって徹底的且つ完全に理解され、本発明の範囲が十分に伝達されるようにするため提供される。本説明中で、提示された実施形態を、不必要に曖昧にするのを避けるために、周知の機能および技法の詳細は省略されることがある。
【０００８】
本発明は、任意の可能な技術的詳細の集約度で、システム、方法、もしくはコンピュータ・プログラム製品またはこれらの組合せとすることができる。このコンピュータ・プログラム製品は、プロセッサに本発明の態様を実行させるためのコンピュータ可読プログラムを有するコンピュータ可読ストレージ媒体（または媒体群）を含むことが可能である。
【０００９】
このコンピュータ可読ストレージ媒体は、命令実行デバイスが使用するための命令を保持し格納できる有形のデバイスとすることができる。該コンピュータ可読ストレージ媒体は、例えば、以下に限らないが、電子ストレージ・デバイス、磁気ストレージ・デバイス、光ストレージ・デバイス、電磁気ストレージ・デバイス、半導体ストレージ・デバイス、または前述のデバイスの任意の適切な組合せであってよい。コンピュータ可読ストレージ媒体のさらに具体的な例の非包括的リストには、携帯型コンピュータ・ディスケット、ハード・ディスク、ランダム・アクセス・メモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、消去およびプログラム可能読み取り専用メモリ（ＥＰＲＯＭ：ｅｒａｓａｂｌｅ ｐｒｏｇｒａｍｍａｂｌｅ ｒｅａｄ－ｏｎｌｙ ｍｅｍｏｒｙまたはフラッシュ・メモリ）、静的ランダム・アクセス・メモリ（ＳＲＡＭ：ｓｔａｔｉｃ ｒａｎｄｏｍ ａｃｃｅｓｓ ｍｅｍｏｒｙ）、携帯型コンパクト・ディスク読み取り専用メモリ（ＣＤ－ＲＯＭ：ｃｏｍｐａｃｔ ｄｉｓｃ ｒｅａｄ－ｏｎｌｙ ｍｅｍｏｒｙ）、デジタル多用途ディスク（ＤＶＤ：ｄｉｇｉｔａｌ ｖｅｒｓａｔｉｌｅ ｄｉｓｋ）、メモリ・スティック、フレキシブル・ディスク、パンチカードまたは記録された命令を有する溝中の嵩上げ構造体などの機械的符号化デバイス、および前述の任意の適切な組合せが含まれる。本明細書で用いられるコンピュータ可読ストレージ媒体は、無線波または他の自由に伝播する電磁波、ウェーブガイドまたは他の送信媒体（例えば、光ファイバ・ケーブルを通過する光パルス）、またはワイヤを通って送信される電気信号などの、本質的に一時的な信号であると解釈されるものではない。

【００１０】
本明細書に述べられるコンピュータ可読プログラム命令は、コンピュータ可読ストレージ媒体から、それぞれのコンピューティング／処理デバイスに、または、例えばインターネット、ローカル・エリア・ネットワーク、広域ネットワークもしくはワイヤレス・ネットワークまたはこれらの組合せなどのネットワークを介して、外部のコンピュータもしくは外部のストレージ・デバイスにダウンロードすることが可能である。このネットワークには、銅送信ケーブル、光送信ファイバ、ワイヤレス通信、ルータ、ファイアウォール、スイッチ、ゲートウェイ・コンピュータ、もしくはエッジ・サーバまたはこれらの組合せが含まれてよい。それぞれのコンピューティング／処理デバイス中のネットワーク・アダプタ・カードまたはネットワーク・インターフェースは、ネットワークからコンピュータ可読プログラム命令を受信し、そのコンピュータ可読プログラム命令を、ストレージのため、それぞれのコンピューティング／処理デバイス内のコンピュータ可読ストレージ媒体中に転送する。

【００１１】
本発明のオペレーションを実行するためのコンピュータ可読プログラム命令は、アセンブラ命令、命令集合アーキテクチャ（ＩＳＡ：ｉｎｓｔｒｕｃｔｉｏｎ－ｓｅｔ－ａｒｃｈｉｔｅｃｔｕｒｅ）命令、マシン命令、マシン依存命令、マイクロコード、ファームウェア命令、状態設定データ、集積回路の構成データ、または、Ｓｍａｌｌｔａｌｋ、Ｃ＋＋などのオブジェクト指向プログラミング言語、および「Ｃ」プログラミング言語もしくは類似のプログラミング言語などの従来式の手続き型プログラミング言語を含む、１つ以上のプログラミング言語の任意の組合せで記述されたソース・コードもしくはオブジェクト・コードであってよい。このコンピュータ可読プログラム命令は、スタンドアロン・ソフトウェア・パッケージとしてユーザのコンピュータで専ら実行することも、ユーザのコンピュータで部分的に実行することもでき、一部をユーザのコンピュータで一部を遠隔コンピュータで実行することもでき、あるいは遠隔のコンピュータまたはサーバで専ら実行することもできる。後者の場合は、ローカル・エリア・ネットワーク（ＬＡＮ：ｌｏｃａｌ ａｒｅａ ｎｅｔｗｏｒｋ）または広域ネットワーク（ＷＡＮ：ｗｉｄｅａｒｅａ ｎｅｔｗｏｒｋ）を含む任意の種類のネットワークを介して、遠隔コンピュータをユーザのコンピュータに接続することもでき、あるいは（例えばインターネット・サービス・プロバイダを使いインターネットを介し）外部のコンピュータへの接続を行うことも可能である。いくつかの実施形態において、例えば、プログラム可能論理回路、フィールドプログラム可能ゲート・アレイ（ＦＰＧＡ：ｆｉｅｌｄ－ｐｒｏｇｒａｍｍａｂｌｅ ｇａｔｅ ａｒｒａｙ）、またはプログラム可能論理アレイ（ＰＬＡ：ｐｒｏｇｒａｍｍａｂｌｅ ｌｏｇｉｃ ａｒｒａｙ）を含む電子回路は、本発明の諸態様を実行すべく、該電子回路をカスタマイズするためコンピュータ可読プログラム命令の状態情報を利用することによって、該コンピュータ可読プログラム命令を実行することができる。

【００１２】
本発明の諸態様は、本発明の諸実施形態による方法、装置（システム）、およびコンピュータ・プログラム製品のフローチャート図もしくはブロック図またはその両方を参照しながら本明細書で説明される。当然のことながら、フローチャート図もしくはブロック図またはその両方の各ブロック、およびフローチャート図もしくはブロック図またはその両方のブロックの組合せは、コンピュータ可読プログラム命令によって実装することが可能である。

【００１３】
これらのコンピュータ可読プログラム命令を、汎用コンピュータ、特殊用途コンピュータ、またはマシンを形成する他のプログラム可能データ処理装置のプロセッサに提供し、そのコンピュータまたは他のプログラム可能データ処理装置のプロセッサを介して実行されるこれらの命令が、フローチャートもしくはブロック図またはその両方のブロックもしくはブロック群中に特定されている機能群／動作群を実装するための手段を生成するようにすることができる。また、コンピュータ、プログラム可能データ処理装置、もしくは他

のデバイスまたはこれらの組合せに対し特定の仕方で機能するよう命令することが可能なこれらのコンピュータ可読プログラム命令を、コンピュータ可読ストレージ媒体に格納し、格納された命令を有するコンピュータ可読ストレージ媒体が、フローチャートもしくはブロック図またはその両方のブロックまたはブロック群中に特定されている機能／動作の諸態様を実装する命令群を包含する製造品を構成するようにすることができる。

【００１４】

さらに、これらコンピュータ可読プログラム命令を、コンピュータ、他のプログラム可能データ処理装置、または他のデバイスにロードし、そのコンピュータ上で、他のプログラム可能装置上で、または他のデバイス上で一連のオペレーション・ステップを実施させて、コンピュータ実装のプロセスを作り出し、当該コンピュータ上で、他のプログラム可能装置上でもしくは他のデバイス上で実行される命令が、フローチャートもしくはブロック図またはその両方のブロックもしくはブロック群中に特定されている機能群／動作群を実装するようにすることも可能である。

【００１５】

諸図面中のフローチャートおよびブロック図は、本発明の様々な実施形態による、システム、方法、およびコンピュータ・プログラム製品から可能となる実装のアーキテクチャ、機能性、およびオペレーションを表している。この点に関し、フローチャートまたはブロック図中の各ブロックは、特定の論理機能（群）を実装するための一つ以上の実行可能命令を含む、モジュール、セグメント、または命令の部分を表し得る。一部の別の実装においては、ブロック中に記載された機能が、図面に記載された順序から外れて行われてよい。例えば、連続して示された２つのブロックが、関与する機能性に応じ、実際にはほぼ同時に実行されることがあり、時にはこれらのブロックが逆の順序で実行されることもあり得る。さらに、ブロック図もしくはフローチャート図またはその両方の各ブロック、およびブロック図もしくはフローチャート図またはその両方中のブロック群の組合せは、特定の機能または動作を実施する特殊用途ハードウェア・ベースのシステムによって実装でき、または特殊用途ハードウェアとコンピュータ命令との組合せによって実行できることにも留意すべきである。

【００１６】

以降で説明する例示的な諸実施形態は、ゲノム・データ・ファイルに対する差分ファイルを圧縮するためのシステム、方法、およびプログラム製品を提供する。しかして、本諸実施形態は、ゲノム解析パイプラインにより生成される中間的なファイルのストレージ必要容量を低減することによって、計算生物学の技術分野を改良する効果を有する。さらに具体的には、少なくとも２つのゲノム・データ・ファイル（例えば、少なくとも１つのソース・ファイルおよび１つの目的ファイル）を、インプットとしてゲノムベース差分圧縮プログラムに投入することができ、次いで、ゲノムベース差分圧縮プログラムが、ソート・アルゴリズムを含み得る既存のオープン・ソース・ツールを利用してゲノム・データ・ファイルの行をソートすることができる。ゲノムベース差分圧縮プログラムは、次いで、ソース・ファイルおよび目的ファイルの各行を逐次的に精査することが可能で、ソース・ファイルと目的ファイルと間で対応する行を対比し、派生差分ファイルを生成することが可能である。次いで、ゲノムベース差分圧縮プログラムは、既存の汎用ファイル・コンプレッサを用いて、生成された派生差分ファイルを圧縮することができる。

【００１７】

前述したように、ゲノム解析パイプライン（パイプライン）は、生の配列リードから生物学的に意味の或るアウトプットを抽出するために、前処理、変異の発見、およびコール・セットのリファインメントなど多くのステップを伴う。該パイプラインは、かかるステップ毎にアウトプット・ファイルを生成し、それらは、インプット配列リードのサイズの如何によって、メガバイト～テラバイトの間の範囲のサイズになる。

【００１８】

これらのファイルの見直しによって、あらゆるステップでこれらアウトプット・ファイルに保存された全ての情報が新規に生成されたものでないことが判明している。データの

かなり大きなパーセントが、インプット・ファイルからアウトプットに単に移されているだけで、パイプラインの実行の過程で、さらにはその後でも、ストレージに対し不必要な負担を生じさせている。パイプラインの各段階は多くの時間または日数を要し得るので、これら中間的なファイルが、パイプラインにおける今後の調査、変更、または分岐のために保存される。

【００１９】

デルタまたは差分圧縮アルゴリズムが存在し得るが（例えば、ｘｄｅｌｔａ３またはｖｃｄｉｆｆ）、但し、これらの圧縮アルゴリズムは、ソース・ファイルと目的ファイルとの間に、比較的に大きく連続した共通の部分列がある場合により効果的であり得る。配列アラインメント／マップ・フォーマット（ＳＡＭ）および変異コール・フォーマット（ＶＣＦ）などの或る種のフォーマットによるゲノム・データは、フィールドから成る比較的小さな記録を含むことがあり、一部のフィールドおよびサブフィールドが１つのバージョンと次のバージョンとで異なることがある。

【００２０】

したがって、小さな記録群に対してゲノム解析パイプラインによって生成された中間的ファイルのストレージ必要容量、または繰り返し情報の量を最小化するためのデータ低減方法を考案することが、なかんずく有益であり得る。

【００２１】

少なくとも１つの実施形態によれば、本ゲノムベース差分圧縮プログラムは、インプット・ファイルとアウトプット・ファイルの対（つい）の間の差分ファイル（すなわち、差分）を、一致、置き換え、挿入、および削除など、単純なコラム操作のセットに低減することができる。ソースおよび目的ファイルは、それぞれのコラムとともに全行を通貫して構文解析することが可能である。差分ファイルは、ソース・ファイルを目的ファイルに変換するために、ソース・ファイルに対し行うことが可能な操作のセットの連続的な記述であってよい。派生差分ファイルは、．ｚｉｐ（すなわち、ファイル圧縮および解凍に使用される或るファイル・フォーマットおよびソフトウェア・アプリケーション）など、既存の汎用ファイル・コンプレッサを用いて圧縮することができる。

【００２２】

少なくとも１つの実施形態によれば、本ゲノムベース差分圧縮プログラムは、医療産業において、ストレージおよびファイル・システム、製品、またはクラウド提供物が、より小さな必要容量によるパイプラインをサポートし作動させ、様々な通信ネットワークを介した伝送を加速化する差別化を可能にすることができる。また、本ゲノムベース差分圧縮プログラムは、様々な中間ステップから、再スタートさせることによって別の分岐の実行を可能にすることができ、仮説および変異のより迅速な確認を可能にすることができる。

【００２３】

少なくとも１つの実施形態によれば、ゲノム・データ・ファイルは、テキストの行（すなわち、記録）のセットとして表されてよく、行の各々は、（例えば、ＳＡＭまたはＶＣＦフォーマットのようにタブで区切られた）固定数のコラムまたはフィールドを含む。ソース・ファイルと目的ファイルとを対比することによって、差分を計算することができる。ソースおよび目的ファイルの行の各々は、それらのマッピング位置の昇順にソートすることが可能である。次いで、ソースおよび目的ファイルは行ごとに同調して精査することができ、ソースおよび目的ファイル中の同じ位置の対応する行が対比され、差分ファイルを得ることができる。この派生差分ファイルは、次いで、汎用コンプレッサを使って圧縮されてよい。

【００２４】

少なくとも１つの実施形態によれば、本ゲノムベース差分圧縮プログラムは、ソースおよび目的ファイルをソートし、リファレンス・ゲノム・ファイルにマップすることが可能であるという事実を考慮に入れることが可能である。しかして、本ゲノムベース差分圧縮プログラムは、ソース・ファイルと目的ファイルとの間の２つの類似の記録に対し、階層識別システムを用いてよい。正確な一致が見出された後、次いで、両方の記録を、個別の

フィールドに構文解析し、差分についてフィールド毎ベースで対比することが可能である。

【００２５】

少なくとも１つの実施形態によれば、本ゲノムベース差分圧縮プログラムは、同じ位置にマップされた目的ファイル中の行をバッファ保存し、同じ位置のソース・ファイル中の次の行に対して目的ファイルを再スキャンすることによって、行の照合を改良することができる。ゲノム・ファイル中の各行は、リファレンス・ゲノムにマップされた「リード」に対応してよい。一般に、リファレンス・ゲノム中の同じ位置にマップされた多くのリードが存在し得る。事実上、ゲノムの同じ領域にオーバーラップしているリードの平均数をカバレッジ係数とすることができる。カバレッジは、同じ位置にマップされ得る数十、さらには数百のリードと高くなり得る。

【００２６】

少なくとも１つの実施形態によれば、より小さな所定粒度を有し得るコラム中の差分を見出すために、より細かい粒度（例えば、サブコラム）を用いることができる。ゲノム・フォーマット中の各記録（すなわち、行）の最後のコラムは、キー－値ペアのセットを含んでよい。パイプラインの諸ステップで、いくつかの新規のキー－値ペアが追加され得、いくつかは削除され得る。しかして、より小さな粒度を用いることによって、より小さい差分を得ることができる。

【００２７】

少なくとも１つの実施形態によれば、差分の符号化（すなわち、圧縮方法）は、ファイルの対の間の差の形でデータを格納することができる。たとえ、配列データの間の差を調べるいくつかの既存の差分圧縮アルゴリズム（例えば、ｖｃｄｉｆｆおよびｘｄｅｌｔａ）が存在していても、これらの既存の差分アルゴリズムは、ゲノム・データ・ファイルに適用されたとき、実際の差よりも大きな派生差分ファイルを生成し得る。これら既存の差分アルゴリズムは、ゲノム・データ・ファイルのソートされた順序、および行のタブ区切り構造の利点を利用できていない可能性がある。ＳＡＭおよびＶＣＦフォーマットのファイルを含め、これらファイルは、効率的な差分圧縮のためゲノムベース差分圧縮プログラムから取り除かれてよい、ヌクレオチド配列および品質値に加え、過度の補助情報を含み得る。

【００２８】

少なくとも１つの実施形態によれば、このアルゴリズムの設計は、パイプラインとデータ・ファイルとの主要特徴に基づくことができる。パイプラインの最初のステップは、配列のアラインメントであり、これはＳＡＭファイルをもたらし、このアラインメントでは、あらゆる生の配列リードが、リファレンス・ゲノム中の特定の位置にマップされるか、またはマップされずに遺棄される。アラインメントの位置は、再アラインメントの過程で変更されない限り、残りのパイプラインの間ずっと同じのままに留まり得る。ほとんどのパイプライン・ツールは、アラインメント・ステップの後でだけファイルをソートすることができる。本ゲノムベース差分圧縮プログラムは、ファイルの対の間の差を同調させて計算するため、このソートを利用することが可能である。２つのファイルの間でソート順序が大きく変化すると、差分の大きさもまた増大し得る。本ゲノムベース差分圧縮プログラムは、差分が受容可能な閾値を超えて増加し得るかどうかを検出することができる。かかる場合に、本ゲノムベース差分圧縮プログラムは、差分圧縮を中止し目的ファイルを変更しないでおくことが可能である。

【００２９】

少なくとも１つの実施形態によれば、本ゲノムベース差分圧縮プログラムは、ソースおよび目的ファイルのヘッダを迅速に検索し、第一記録に到達することができる。ソース・ファイル中のあらゆる個別記録に対し、差分ファイル中に対応する処置があり得る。あらゆるソース記録に対し、本ゲノムベース差分圧縮プログラムは、最初に、目的ファイル中に一致を見出す試みをすることができる。この一致を見出すための検索領域は、ＳＡＭの場合は、同じ値を持つリファレンス配列名（ＲＥＦまたはＲＮＡＭＥ）および１から始まる左端のマッピング位置（ＰＯＳ：ｐｏｓｉｔｉｏｎ）に、ＶＣＦの場合は、リファレン

ス・ゲノムからの染色体識別子またはアセンブリ・ファイル（＃ＣＨＲＯＭ）中のコンティグ（例えば、オーバーラップしている配列データまたはリード）と位置（ＰＯＳ）を指している角括弧内の識別文字列を有する記録に限定され得る。本ゲノムベース差分圧縮プログラムは、目的ファイルを、一致を見出すことができるまでまたは検索領域が尽きるまで、逐次的に読み取ることが可能である。目的ファイルからの不一致記録は、１つのソース記録に対する一致の検索の過程で遭遇することができ、それらは、本ゲノムベース差分圧縮プログラムが、ソース・ファイルの次の記録に進む際に不一致記録を改めて処理するために、バッファ（例えば、ファイルが一時的に格納されるメモリの小部分）にしまわれてよい。バッファにしまわれた記録の数が受容可能な閾値を超えた場合、本ゲノムベース差分圧縮プログラムは検索を中止することができる。一致が見出せない場合、当該ソース記録は、差分ファイル中に「記録を削除」と標識されてよい。本ゲノムベース差分圧縮プログラムが、目的ファイルを逐次に読取っている間に、ソース記録のものよりも小さい可能性のあるＲＥＦ、ＰＯＳまたは＃ＣＨＲＯＭ、ＲＥＦを有する記録に遭遇した場合、目的ファイルからのその特定の記録は、「記録を挿入」と表記されてよく、差分ファイル中に加えられてよい。

【００３０】

少なくとも１つの実施形態によれば、本ゲノムベース差分圧縮プログラムは、特定の請求項に基づき、ソース・ファイルと目的ファイルとを照合するため階層識別システムを利用することができる。例えば、ＳＡＭファイルでは、本ゲノムベース差分圧縮プログラムは、クエリ・テンプレート名（ＱＮＡＭＥ）、ビット単位フラグ（ＦＬＡＧ）、ＲＥＦ、ＰＯＳ（すなわち、第一～第四コラム）に関連付けられたコラムを利用することが可能で、ＶＣＦファイルでは、本ゲノムベース差分圧縮プログラムは、＃ＣＨＲＯＭ、ＰＯＳ、ＲＥＦ、および代替塩基（群）（ＡＬＴ）（すなわち、第一～第二コラム、および第四～第五コラム）を利用することが可能である。

【００３１】

少なくとも１つの実施形態によれば、ソース記録と目的記録との間の一致が判断された後、特定のフィールド（例えば、ＳＡＭファイルでは、任意もしくは非必須のフィールド（ＯＰＴＩＯＮＡＬ）または第十二コラム、ＶＣＦファイルでは、付加情報（ＩＮＦＯ）または第八コラム）に対してフィールド毎の対比を行うことができる。これら対比されたフィールドに対し、フィールド中に不一致があった場合、その特定のフィールドは「置き換え」と標識されてよく、差分ファイルにその新規の値を加えることができる。対比から除外されたファイル（例えば、ＯＰＴＩＯＮＡＬおよびＩＮＦＯ）は、いくつかのキー－値ペア、またはキー単体および値単体を含むことが可能で、これらは、セミコロン、空白スペース、または別のマークで区切られてよい。対比から除外されたこれらのフィールドは、次いで、目的およびソース・ファイルの両方に対して、それらそれぞれのサブフィールドに構文解析することができる。ソースに存在し、目的に不在のキーまたは値は、差分ファイル中に「削除」と標識されてよい。目的に存在し、ソースに不在のキーまたは値は、「挿入」と標識されてよく、差分ファイル中に加えられてよい。ソース記録に対するこれら操作が決められ差分ファイルに加えられた後、本ゲノムベース差分圧縮プログラムは、ソース・ファイル中の次の記録の読み取りに進むことができ、ソースおよび目的に対し、ファイルの終了点（ＥＯＦ：ｅｎｄ－ｏｆ－ｆｉｌｅ）に達するまで、このプロセスを繰り返すことができる。生成された差分ファイルは、高度に圧縮可能であり得、アーカイブのため圧縮されてよい。

【００３２】

少なくとも１つの実施形態によれば、あらゆるソース記録に対し、本ゲノムベース差分圧縮プログラムは、目的ファイルから新しい記録を読み取る前に、まずバッファ中の一致を検索することができる。バッファは、検索領域が変更される都度、毎回クリアすることが可能である。全ての不一致記録は、「挿入」とタグされ、差分に加えられてよい。

【００３３】

図１を参照すると、少なくとも１つの実施形態による、典型的なネットワーク化コンピ

ュータ環境１００が描かれている。このネットワーク化コンピュータ環境１００は、プロセッサ１０４と、ソフトウェア・プログラム１０８の実行が可能にされたデータ・ストレージ・デバイス１０６と、ゲノムベース差分圧縮プログラム１１０ａとを備えたコンピュータ１０２を含んでよい。また、ネットワーク化コンピュータ環境１００は、データベース１１４および通信ネットワーク１１６と相互作用することが可能なゲノムベース差分圧縮プログラム１１０ｂの実行を可能にされたサーバ１１２も含んでよい。ネットワーク化コンピュータ環境１００は、複数のコンピュータ１０２およびサーバ１１２を含むことができ、そのうちの１つのみが示されている。通信ネットワーク１１６は、広域ネットワーク（ＷＡＮ）、ローカル・エリア・ネットワーク（ＬＡＮ）、電気通信ネットワーク、ワイヤレス・ネットワーク、公衆交換ネットワーク、もしくは衛星ネットワーク、またはこれらの組合せなど、様々な種類の通信ネットワークを含んでよい。当然のことながら、図１は、１つの実装の単なる例を提示しており、各種の実施形態が実装可能な環境に関し、いかなる限定をも意味するものではない。設計および実装上の要件に基づいて、描かれた環境に多くの修改を加えることが可能である。

【００３４】

クライアント・コンピュータ１０２は、通信ネットワーク１１６を介してサーバ・コンピュータ１１２と通信することが可能である。通信ネットワーク１１６は、有線、ワイヤレス通信リンク、または光ファイバ・ケーブルなどの接続手段を含んでよい。図９を参照しながら説明するように、サーバ・コンピュータ１１２は、内部コンポーネント９０２ａおよび外部コンポーネント９０４ａをそれぞれ含んでよく、クライアント・コンピュータ１０２は、内部コンポーネント９０２ｂおよび外部コンポーネント９０４ｂをそれぞれ含んでよい。また、サーバ・コンピュータ１１２は、サービスとしてのソフトウェア（ＳａａＳ：Ｓｏｆｔｗａｒｅ ａｓ ａ Ｓｅｒｖｉｃｅ）、サービスとしてのプラットフォーム（ＰａａＳ：Ｐｌａｔｆｏｒｍ ａｓ ａ Ｓｅｒｖｉｃｅ）、またはサービスとしてのインフラストラクチャ（ＩａａＳ：Ｉｎｆｒａｓｔｒｕｃｔｕｒｅ ａｓ ａ Ｓｅｒｖｉｃｅ）など、クラウド・コンピューティング・サービス・モデル中で運用することも可能である。また、サーバ１１２は、プライベート・クラウド、コミュニティ・クラウド、パブリック・クラウド、またはハイブリッド・クラウドなど、クラウド・コンピューティング展開モデル中に配置されてもよい。クライアント・コンピュータ１０２は、例えば、携帯デバイス、電話、携帯情報端末、ネットブック、ラップトップ・コンピュータ、タブレット・コンピュータ、デスクトップ・コンピュータ、または、プログラムを実行し、ネットワークにアクセスし、データベース１１４にアクセスする能力のある任意の種類のコンピューティング・デバイスであってよい。本実施形態の様々な実装によれば、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、以下に限らないが、コンピュータ／携帯デバイス１０２、ネットワーク・サーバ１１２、またはクラウド・ストレージ・サービスなどの様々なストレージ・デバイス中に内蔵することが可能なデータベース１１４とやり取りすることができる。

【００３５】

本実施形態によれば、クライアント・コンピュータ１０２またはサーバ・コンピュータ１１２を用いるユーザは、ゲノム・データ・ファイルに対する差分ファイルを圧縮するためにゲノムベース差分圧縮プログラム１１０ａ、１１０ｂを（それぞれ）使用することが可能である。ゲノムベース差分圧縮の方法については、以降で図２～８に関連してさらに詳しく説明する。

【００３６】

ここで図２を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａおよび１１０ｂによって用いられる、ゲノム・データ・ファイルに対する典型的な圧縮プロセス２００を示すオペレーション・フローチャートが描かれている。

【００３７】

２０２で、ゲノム・データ・ファイルが、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂへのインプットとして受信される。ゲノム・データ・ファイルは、ユーザのデバ

イス（例えば、ユーザのコンピュータ１０２）上のソフトウェア・プログラム１０８を用い、通信ネットワーク１１６を介してゲノムベース差分圧縮プログラム１１０ａ、１１０ｂ中に伝送することができる。ゲノム・データは、或る生体に関連するゲノム・データを含んでよく、このゲノム・データのファイルは、デオキシリボ核酸（ＤＮＡ：ｄｅｏｘｙｒｉｂｏｎｕｃｌｅｉｃ　ａｃｉｄ）配列のリード（すなわち、ＤＮＡ分子内のヌクレオチドの正確な順序）、リファレンス・ゲノムへのマッピング、およびリファレンス・ゲノムに対して検出された変異を含んでよい。また、ゲノム・データ・ファイルは、少なくとも２つのファイル、ソース（すなわち、リファレンス）および目的ファイルに対し、各々が固定数のコラム（すなわち、フィールド）を含む行（すなわち、記録）のセットを含んでよい。ソース・ファイルは、生体のリファレンス・ゲノムから直接導出することが可能なデータである。各ファイルは、そのファイルが目的ファイルまたはソース・ファイルのどちらかによって標識されてよい（例えば、「ｔａｒｇｅｔ＿ｎａｍｅ．ｆｏｒｍａｔ　ｔｙｐｅ（目的＿名前．フォーマット・タイプ）」または「ｓｏｕｒｃｅ＿ｎａｍｅ．ｆｏｒｍａｔ　ｔｙｐｅ（ソース＿名前．フォーマット・タイプ）」）。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的ファイルを再構築するためにソース・ファイルを利用することが可能である。

【００３８】

例えばゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ゲノム解析ツールキット（ＧＡＴＫ：ｇｅｎｏｍｅ　ａｎａｌｙｓｉｓ　ｔｏｏｌｋｉｔ）から伝送されてくるソース・ファイル（すなわち、ｓｏｕｒｃｅ＿ｌｉｖｅｒＨＳ１５６７．ｓａｍ）および目的ファイル（すなわち、ｔａｒｇｅｔ＿ｌｉｖｅｒＨＳ１５６７．ｓａｍ）を受信する。ソース・ファイルは、或る健康な人の肝細胞標本に関連付けられたリファレンス・ゲノムからのデータを含み、目的ファイルは、肝がんを患っている人からの肝細胞標本に関連付けられたデータを含む。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３０２で、インプットとしてソースおよび目的ファイルを受信するステップを含め、ゲノム・データ・ファイルに対する典型的な圧縮プロセス３００については、図３に関連して後記でもっと詳しく説明することとする。

【００３９】

本実施形態において、ゲノム・データ・ファイルは、ＧＡＴＫなどのゲノム解析パイプライン実行からのアウトプットとして生成されてよい。パイプラインのスクリプトは、目的およびソース・ファイルにゲノムベース差分圧縮プログラム１１０ａ、１１０ｂを実行するための特定のコマンドを使って充実させることが可能である。さらに、圧縮されたアウトプット・ファイル（例えば、ＳＡＭ（ＢＡＭ）およびＶＣＦ（ＢＣＦ）ファイルのバイナリ・バージョン）の一部は、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂを開始する前に、外部のエンジンを利用して解凍することができる。

【００４０】

本実施形態において、ソースおよび目的ファイルは、アウトプット・ファイルが生成され格納される場所であるフォルダまたはディレクトリ（すなわち、データベース１１４）中の、パイプラインの実行中に生成された諸ファイルから読み出すことが可能である。

【００４１】

本実施形態において、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂ中にインプットとして１つのゲノム・データ・ファイルしか投入されなかった場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、派生差分ファイルを計算することはできない。代わりに、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ユーザにエラー・メッセージを返すことができる。ユーザは、同じインプット・ファイルを再投入することが可能であるが、但し、そのファイルは、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが派生差分ファイルを計算するための、対応するソースまたは目的ファイルと共に投入される必要がある。

【００４２】

次に２０４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって、ゲノム

・データ・ファイルの行がソートされる。ソースおよび目的ファイルの行は、リファレンス・ゲノム中のそれらのマッピング位置（ＰＯＳ）に基づいて昇順に配置することができ、これにより、ソース・ファイルと目的ファイルとの間での行の照合のための検索領域を絞りこみ、バッファ占拠域を最小化し、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂの実効時間を最適化することが可能になる。ゲノム・データ・ファイルの行に対する行群をソートするために、既存のソート・アルゴリズムを含む既存のオープン・ソース・ソート・ツールを利用することができる。

【００４３】

目的およびソース・ファイルのフォーマットの型によっては、これらファイルは、パイプラインにおいて既にソートされていることがある。ソースおよび目的ファイルがパイプラインにおいてソート済みの場合、それら目的およびソース・ファイルは、２０２で、インプットとしてゲノムベース差分圧縮プログラム１１０ａ、１１０ｂ中に投入されるとき、それらそれぞれの行のソートされた順序を保持し得る。

【００４４】

前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＧＡＴＫから得られたソースおよび目的ファイルが既にソートされており、したがって、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂがマッピング位置（ＰＯＳ）に基づいて行をソートする必要はない、と判断する。

【００４５】

本例の目的の上で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＳＡＭまたはＶＣＦフォーマットのファイルにだけ対応可能である。ファイルがなんらかの他のフォーマットである場合、ユーザには、エラー・メッセージが提示されることになり、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、対応可能なソースおよび目的ファイルがインプットとして投入されるまで、実行を停止する。本例では、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルはＳＡＭフォーマットであると判断し、これは、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが対応可能である。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂはバッファ（すなわち、目的ファイルがソース・ファイルに一致しなかった場合に、その目的ファイルを格納できる、メモリの小部分）を起動する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂがフォーマットを識別するためファイルをチェックし、バッファを初期化する（ＩＮＩＴ：ｉｎｉｔｉａｌｉｚｉｎｇ）ステップを含め、ゲノム・データ・ファイルに対する差分ファイルの圧縮のための典型的なプロセスについては、図３と関連して後記でもっと詳しく説明することとする。

【００４６】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、インプットとしてファイルを受信する前にソースおよび目的ファイルがソートされていなかった場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、それらのＰＯＳ（すなわち、各ファイルの行の第四コラム）に基づいてオープン・ソース・ソート・ツールを利用することによって、ソースおよび目的ファイルの各々の行を昇順でソートすることができる。

【００４７】

本実施形態において、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ファイルのフォーマット型（例えば、ＳＡＭ、ＢＡＭ、ＶＣＦ、ＢＣＦ）の対応可能性をチェックすることが可能である。ファイルのフォーマット型が、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって対応可能である場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは先に進んでよい。また一方、ファイルのフォーマット型が、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂにとって対応可能でない場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、実行を停止してよく、ユーザにエラー・メッセージを提示することができる。

【００４８】

本実施形態において、ゲノム・データ・ファイルの行群を通貫してソートするために用

いることが可能なソート・アルゴリズムは、ゲノム・データ・ファイルの複雑さの如何によっては、部分的に変更をする必要があるかもしれない。

【００４９】

次いで２０６で、ソースおよび目的ファイルが精査される。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルを逐次的に精査する（すなわち、読み取る）することが可能で、このとき、ゲノム・データ・ファイル中の各ラインを一行として判断してよい。インプット・ファイルは、改行コード（すなわち、ラインの終了と新ラインの開始を表す特殊な文字または文字列）に遭遇するまで、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって読み取られてよい。さらに、これらの行は、コラム（すなわち、フィールド）に分割することができる。２つのファイルが同調されていることを確認するために、これら行の値が対比されてよい。両方のファイルは、これらファイルのコンテンツの間の差を、時間およびメモリ効率的仕方で計算するために同調させて精査することができる。

【００５０】

前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルの各行を精査する。ソースおよび目的ファイルは、各行に３つのデータ片を備える４つの行を含む。２つのゲノム・データ・ファイルを対比するブロック図について、図６に関連して後記でもっと詳しく説明することとする。ソース・ファイルの各行を精査した後、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファが空かどうかを判断する（すなわち、バッファが空の場合、目的ファイルは、バッファの代わりにファイルから読み取ることが可能である）。本例では、バッファは空である。したがって、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ファイルからの各目的行を精査すればよい。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂがソースおよび目的ファイルの各行を精査した後、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルが同調されていることを確認する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂがソースおよび目的ファイルの行を読み取るステップを含め、ゲノム・データ・ファイルに対する差分ファイルを圧縮するための典型的なプロセスについて、図３に関連して後記でもっと詳しく説明することとする。

【００５１】

次いで２０８で、目的ファイルとソース・ファイルとの行が対比される。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルと目的ファイルとの間の一切の差を判断するために、ソースおよび目的ファイル上の対応する行（すなわち、同じマッピング位置を有する行）を対比し、派生差分ファイル（すなわち、派生Ｄファイル）を生成することができる。

【００５２】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、部分一致している行の階層的フィールド単位の対比を（例えば、各ファイルに対し、特定のコラムが対比される）、および一致している行に対しては行の全面的対比を行う（例えば、両方のファイル中のさらなるコラムが対比される）ことが可能である。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂを用いて、対比される行が一致しているかどうかによりどちらの対比を使えばよいかを判断することができる。目的およびソース・ファイルからのそれぞれの行の対比は、特定のコラムに対するフィールド単位の対比によって開始されてよい。階層的フィールド単位の対比の過程での各対比によって、中間的な一致または不一致を記録することができる。階層的フィールド単位の対比の終了点で、行の間の一致または不一致のいずれかが生じ得る。この一致または不一致は、対比された行が同調しているかしていないかを判断し、さらに、不一致があるかどうかを判断するための手っ取り早い方法であり得る。これら２つの行の間に一致がある場合、行全体の対比（すなわち、全コラムが対比される）を行うことができる。この対比の結果は、派生差分ファイルに取り込むことができ、このファイルはゲノムベース差分圧縮プログラム１１０ａ、１１０ｂのアウトプットである。

【００５３】

前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、最初に、目的およびソース・ファイル内のそれぞれの行の階層的フィールド単位の対比を行う。このソースおよび目的ファイルはＳＡＭフォーマットなので、１１の固定コラムがある。しかしながら、階層的フィールド単位の対比は、４つのコラム（すなわち、ＰＯＳ、ＲＥＦ、ＱＮＡＭＥ、およびＦＬＡＧ）だけを対比する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的およびソース・ファイル中の第一行のＰＯＳおよびＲＥＦコラムを最初に対比し、それらＰＯＳおよびＲＥＦコラムが一致するかどうかを判断する。目的およびソース・ファイル中の第一行のＰＯＳおよびＲＥＦコラムが一致したので、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは中間的な一致があると判断し、次いで、同じ行のＱＮＡＭＥコラムが対比される。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、同じ行のＱＮＡＭＥコラムが一致しており、中間的な一致があると判断する。次いで、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、同じ行のＦＬＡＧ（ビット単位）コラムを対比する。これらの行のＦＬＡＧ（ビット単位）コラムが一致しないので、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、これらの行のＦＬＡＧ（ビット単位）の間の差が０ｘ４００（すなわち、光学的デュープリケート距離）以下かどうかを判断する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、この差が０ｘ４００に等しかったと判断する。したがって、目的ファイルとソース・ファイルとの第一行は一致する。ソース・ファイルと目的ファイルとの対比によって生成されたＤファイルの第一行に「m」が含められる。次いで、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイル中の次の行の階層的フィールド単位の対比に進んでよい。ＳＡＭフォーマットのゲノム・データ・ファイルの階層的構造を識別するための典型的なプロセスについて、図７に関連して後記でもっと詳しく説明することとする。

【００５４】

部分一致の行のフィールド単位の対比が行われたとき、以下の結果が生成され、派生Ｄファイルを生成するために用いられた。

ｆｉｅｌｄ（フィールド）＿ｉ（Ｓ）がｆｉｅｌｄ＿ｉ（Ｔ）と異なる場合：Ｄ←（ｆｉｅｌｄ＿ｉ（Ｔ），ｉ）

【００５５】

階層的フィールド単位の対比で対比された行が一致する場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、行全体の対比を行うことになり、この対比は、目的およびソース・ファイルの行中の、ＲＮＡＭＥ（すなわち、第三コラムであり、アラインメントの名前を表す）と、ＭＡＰＱ（すなわち、第五コラムであり、品質をマッピングしている）と、ＣＩＧＡＲ（すなわち、第六コラムであり、配列が一致しているかどうかに関連する文字列演算を表す）と、ＲＮＥＸＴ（すなわち、第七コラムであり、次のリードのリファレンス名を表す）と、ＰＮＥＸＴ（すなわち、第８コラムであり、次のリードの位置を表す）と、ＴＬＥＮ（すなわち、第九コラムであり、観測されたテンプレートの長さを表す）と、ＳＥＱ（すなわち、第十コラムであり、部分配列を表す）と、ＱＵＡＬ（すなわち、第十一コラムであり、塩基のエラー確率によって塩基の品質を表す）とを含む、ＯＰＴＩＯＮＡＬコラム（すなわち、残りの任意のまたは非必須の７つのコラム、または第十二コラム）の対比、およびソース・ファイルと目的ファイルとの行の間の差を計算するためにＯＰＴＩＯＮＡＬコラムのサブフィールド全体を通貫して構文解析することを含む。次いで、これらコラムの差が派生ファイルに加えられ、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルから次の行を読み取る。ＳＡＭフォーマットのファイルに対する階層的識別のためのプロセスを含むゲノム・データ・ファイルの典型的圧縮プロセス３００について、図４に関連して後記でもっと詳しく説明することとする。

【００５６】

行が不一致なので、行の全面的対比は行わなくてもよい。また一方、行の全面的対比が行われた場合は、以下の結果を生成することができる。

一致（Ｄ←「m」）、削除（Ｄ←「ｄ」）、挿入（Ｄ←Ｔ（ｉ））

【００５７】

また一方、ソースおよび目的ファイルがＶＣＦフォーマットであった場合は、各行内に８つの固定で必須のコラムがある。この階層的フィールド単位の対比では、４つのコラムだけ（すなわち、＃ＣＨＲＯＭ、ＰＯＳ、ＲＥＦ、およびＡＬＴ）を対比する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、最初に目的およびソース・ファイル中の第一行の＃ＣＨＲＯＭおよびＰＯＳコラムを対比し、それら＃ＣＨＲＯＭおよびＰＯＳコラムが一致するかどうかを判断する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが＃ＣＨＲＯＭおよびＰＯＳコラムが中間的な一致または不一致のどちらであると判断したかによって、同じ行のＲＥＦおよびＡＬＴコラムが対比されることになる。ＶＣＦフォーマットのゲノム・データ・ファイルの階層構造を識別するための典型的なプロセスについて、図８に関連して後記でもっと詳しく説明することとする。

【００５８】

階層的フィールド単位の対比で対比された行が一致すれば、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、行の全面的対比を行うことができ、この対比は、目的およびソース・ファイルの当該行中のＩＮＦＯコラム（すなわち、付加情報を表す第八コラム）の対比と、ソース・ファイルと目的ファイルとの行の間の差を計算して派生ファイルを生成するためにＩＮＦＯコラムのサブフィールドを通貫して構文解析することとを含む。次いで、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂはソース・ファイルから次の行を読み取る。階層的識別のためのプロセスを含め、ＶＣＦフォーマットのファイルのゲノム・データ・ファイルに対する典型的な圧縮プロセス３００について、図５に関連して後記でもっと詳しく説明することとする。

【００５９】

次いで２１０で、派生差分ファイルは、既存の汎用ファイル・コンプレッサを使って圧縮される。派生差分ファイル（すなわち、派生Ｄファイル）の汎用ファイル・コンプレッサへの伝送は、派生Ｄファイルの一時的変更に対するコードを使って自動化することができる。派生Ｄファイルの圧縮が完了した後、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ユーザのデバイス（例えば、ユーザのコンピュータ１０２）のメモリに保存されてよい。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ユーザに、派生Ｄファイルの圧縮が完了したことを（例えば、ダイアログ・ボックスを介して）プロンプトすることができる。このダイアログ・ボックスには、例えば、圧縮派生Ｄファイルが完了したことのメッセージ、およびダイアログ・ボックスの下方に「詳細を見る」ボタンを含めてよい。ユーザが「詳細を見る」ボタンをクリックすると、このダイアログ・ボックスは消え、圧縮派生Ｄファイルを含む別のダイアログ・ボックスをユーザに提示することができる。この派生Ｄファイルは、目的ファイルを再生するために、ソース・ファイルと併せ用いることが可能である。

【００６０】

前述の例を続けると、ソース・ファイルと目的ファイルとの対比によって生成された（すなわち、ソース・ファイルが、目的ファイルを派生Ｄファイルとして再構成するために用いられた）派生Ｄファイルは、．ｚｉｐによって圧縮される。派生差分ファイルは、当初のサイズの５２％に低減されており、圧縮派生Ｄファイルは、ユーザのコンピュータ（すなわち、ユーザ・デバイス１０２）のメモリに保存される。ユーザは、ダイアログ・ボックスによって、圧縮派生Ｄファイルが完了したことをプロンプトされ、ユーザはメッセージの下方の「詳細を見る」ボタンをクリックする。ユーザが「詳細を見る」ボタンをクリックすると、このダイアログ・ボックスは消え、圧縮派生Ｄファイルをユーザに提示する別のダイアログ・ボックスが現れる。

【００６１】

ここで図３を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａおよび１１０ｂによって用いられるゲノム・データ・ファイルに対する、典型的な圧縮プロセス３００を示すオペレーション・フローチャートが描かれている。図示のように、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、Ｄファイルを生成するためにソースおよび目的ファイルを用い、派生Ｄファイルを圧縮するために汎用フ

ァイル・コンプレッサを用いる。

【００６２】

３０２で、ソースおよび目的ファイルが、インプットとしてゲノムベース差分圧縮プログラム１１０ａ、１１０ｂ中に投入される。ソースおよび目的ファイルは、ユーザのデバイス（例えば、ユーザのコンピュータ１０２）上のソフトウェア・プログラム１０８を使い、通信ネットワーク１１６を介して、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂ中に、インプットとして伝送することが可能である。各ファイルは、そのファイルがソース・ファイルかまたは目的ファイルなのかを示すために標識されてよい。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、伝送されてきたファイルの種類（すなわち、そのファイルがソース・ファイルかまたは目的ファイルのいずれか）を識別することを可能にできる。

【００６３】

例えば、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ゲノム解析ツールキット（ＧＡＴＫ）から、ソース・ファイルおよび目的ファイルを受信する。このソース・ファイルは、健全なヒトスジシマ蚊に関連付けられたリファレンス・ゲノムからのデータを含み、目的ファイルは、西ナイル・ウイルス（すなわち、フラビウイルス）に感染したヒトスジシマ蚊のゲノムに関連付けられたデータを含む。

【００６４】

次に３０４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、フォーマットを識別するためにファイルをチェックし、バッファ（ＩＮＩＴバッファ）を初期化する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルが対応可能なフォーマット（例えば、ＳＡＭまたはＶＣＦ）であることをチェックすることができる。次いで、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファを初期化することができ、該バッファは、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによってソース・ファイルから一致する行が見出されるまで、目的ファイルからの行を格納することが可能なメモリの小部分であってよい。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソースおよび目的ファイルがＶＣＦフォーマットであることを確認した。加えて、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファを起動する。

【００６５】

次いで３０６で、ソース・ファイルから一行が読み取られる。フォーマットの型、およびファイルがパイプラインから読み出されているかどうかによって、それぞれのファイルの行が既にソートされている可能性があり、その場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルの精査（または読み取り）に進んでよい。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルが、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂに伝送される前にパイプラインにおいて既にソートされていたと判断する。しかして、受信されたソースおよび目的ファイルは、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが受信ソース・ファイルの行を読み取る前に、ソートを行う必要はない。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルの第一行の読み取りに進むことができる。

【００６６】

３０８で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、バッファが空である（例えば、バッファ内に目的ファイルがない）と判断した場合、３１２で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂはファイルから目的行を読み取ってよい。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファを検索し、バッファ中に何らかの目的行があるかどうかを判断する。バッファ中にいかなる目的行もない場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファが空であると判断する。

【００６７】

また一方、３０８で、バッファが空でない（例えば、バッファが少なくとも１つの目的

行を含む）場合、３１０で、その目的行がバッファから読み取られる。バッファ中に目的行が存在する場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファが空でないと判断する。

【００６８】

前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、バッファが空でなく３つの目的行を含むと判断する。しかして、それら目的行の１つがバッファから読み取られる。

【００６９】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３１０でバッファから目的行を読み取るか、または３１２でファイルから目的行を読み取ったとき、３１４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、当該ソースおよび目的ファイルに対するフォーマットがＳＡＭかまたはＶＣＦかを判断する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂがソースおよび目的ファイルの両方を読み取った後、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、派生Ｄファイルの生成および圧縮に進み、これらはフォーマットの型によって異なる。３１４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、ファイルはＳＡＭフォーマットであると判断した場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは図４に進む。また一方、３１４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、ファイルはＶＣＦフォーマットであると判断した場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは図５に進む。

【００７０】

前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルの両方がＶＣＦフォーマットであることを確認し、次いでゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは図５に進む。

【００７１】

ここで図４を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａおよび１１０ｂによって用いられる、ＳＡＭフォーマットのゲノム・データ・ファイルに対する典型的な圧縮プロセス３００を示すオペレーション・フローチャートが描かれている。ＳＡＭフォーマットのソース・ファイルと目的ファイルとの行内の諸フィールドの対比が示されている。図４では、ソースおよび目的ファイルはＳＡＭフォーマットである。

【００７２】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３１６で、ＰＯＳおよびＲＥＦは同じでないと判断した場合、３１８で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルのＰＯＳおよびＲＥＦが目的ファイルのものより小さいかどうかを判断する。ＰＯＳおよびＲＥＦが同じかどうかを判断するために、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって、ＳＡＭフォーマットのソースおよび目的ファイルのそれぞれの行を評価することができる。

【００７３】

前述の例を続けると、ソースおよび目的ファイルがＳＡＭフォーマットであった場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルの第一行を評価し、それぞれのファイルのＰＯＳおよびＲＥＦが同じかどうかを判断することが可能である。ソースおよび目的ファイルが同じである場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルと目的ファイルとのＱＮＡＭＥが同じであるかどうかを評価してよい。また一方、ソース・ファイルと目的ファイルとのＰＯＳおよびＲＥＦが同じでない場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルのＰＯＳおよびＲＥＦが、目的ファイルより小さいかどうかを判断することができる。

【００７４】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３１８で、ソース・ファイルのＰＯＳおよびＲＥＦは、目的ファイルより大きいと判断した場合、３２０で、ゲノムベ

ース差分圧縮プログラム１１０ａ、１１０ｂは、目的の記録を差分に挿入し、３０６のソースから行を読み取る、に戻る。したがって、ソース・ファイル中の行のＰＯＳおよびＲＥＦが目的ファイル中のものより大きい場合、その目的の記録を派生差分ファイルに含めることができる。前述の例を続けると、ソース・ファイルが目的ファイルよりも大きい場合、その目的の記録は差分中に挿入されてよい。したがって、派生差分ファイルの当該行中に、その目的の記録を挿入することが可能である。

【００７５】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３１８で、ソース・ファイルのＰＯＳおよびＲＥＦが目的ファイル中の対応する行よりも小さいと判断した場合、３２２で、差分中の対応する行は削除され、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、３０６のソース・ファイルから次の行を読み取る、に戻る。前述の例を続けると、ソース・ファイルが目的ファイルよりも小さい場合、そのソースの記録は、派生差分から除去されてよい。しかして、その差分は、派生差分ファイルの当該行に、ソースの記録の代わりに「ｄ」を有してよい。

【００７６】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３１６で、ソース・ファイル中の行のＰＯＳおよびＲＥＦは目的ファイル中の対応する行と同じであると判断した場合、３２４で、その行は、それぞれのＱＮＡＭＥが同じかどうかを判断するため評価される。前述の例を続けると、受信されたソースおよび目的ファイルの行が同じＰＯＳおよびＲＥＦを有する場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソース・ファイルと目的ファイルとの対応する行のＱＮＡＭＥが同じかどうかを判断することができる。

【００７７】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３２４で、ソースおよび目的ファイルの行のＱＮＡＭＥは同じでないと判断した場合、３２６で、目的ファイルからの当該行はバッファに保存され、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、３１２のファイルから次の目的行を読み取る、に戻る。前述の例を続けると、目的およびソース・ファイルの対応する行のＱＮＡＭＥが同じでない（すなわち、異なる）場合、その目的行は、該目的行に対するより良好な一致が判断されるまで、バッファ中に保存される。

【００７８】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３２４で、それぞれの行のＱＮＡＭＥが同じであると判断した場合、３２８で、目的およびソース・ファイルの当該行は、ＦＬＡＧが同じであるかどうかを判断するために評価される。前述の例を続けると、ソースおよび目的ファイルの対応する行のＱＮＡＭＥが同じである場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソースおよび目的ファイルの対応する行のＦＬＡＧ（ビット単位）を対比する。

【００７９】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３２８で、それぞれの行のＦＬＡＧが同じでないと判断した場合、３３０で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、その差が閾値（例えば、０ｘ４００）以下かどうかを判断する。３３０で、差が閾値（例えば、０ｘ４００）より大きい場合、３２６で、目的ファイルからの行がバッファに保存され、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ファイルから目的行を読み取るため３１２に戻る。前述の例を続けると、受信されたソースおよび目的ファイルの対応する行のそれぞれのＦＬＡＧ（ビット単位）が相異なる場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、その差が閾値以下かどうかを判断することができる。この閾値は、ユーザが前もって決めておくことが可能で、本例では、０ｘ４００（すなわち、光学的デュープリケート距離）の閾値がユーザによって事前に選択されていた。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＦＬＡＧ（ビット単位）の差が０ｘ４００以下かどうかを判断すればよい。

【００８０】
また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３３０で、この差は閾値（例えば、０ｘ４００）以下であると判断するか、または、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３２８で、目的およびソース・ファイルの行のＦＬＡＧが同じであると判断した場合、３３２で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＯＰＴＩＯＮＡＬフィールドを除き、コラムの差を計算する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、コラムの差を調べ、次いでどのコラムがかかる差を含むかを示すことができる。差のあるコラムの位置およびコラムの差の程度の如何によって、派生差分ファイル内の記録が影響され得る。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、受信されたソースおよび目的ファイルの対応する行の間の差が０ｘ４００以下であると判断するか、または受信されたソースおよび目的ファイルの対応する行のＦＬＡＧ（ビット単位）が同じであると判断した場合、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＯＰＴＩＯＮＡＬフィールドを除いたコラムの差の計算に進んでよい。
【００８１】
次いで３３４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＯＰＴＩＯＮＡＬフィールドをサブフィールドに構文解析する。ＯＰＴＩＯＮＡＬのサブフィールドは、ＳＡＭフォーマット・ファイルのアラインメント記録中に含めることが可能で、これらには、ソース・ファイルと目的ファイルとの対応する行のＯＰＴＩＯＮＡＬのサブフィールドの間の相似点または差を判断するため、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって構文解析することが可能な、２０より多い所定の標準タグを含めることが可能である。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソースおよび目的ファイルの対応する行の間に存在する一切の相似点および差があるかどうかを判断するために、ＯＰＴＩＯＮＡＬフィールドを通貫して、ＯＰＴＩＯＮＡＬフィールド内のサブフィールド中の所定の標準タグを検索することができる。
【００８２】
次いで３３６で、目的ファイルとソース・ファイルとのそれぞれの行のこれらサブフィールドの差が計算される。このとき、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的およびソース・ファイルのそれぞれの行のＯＰＴＩＯＮＡＬのサブフィールドを通貫して差を調べることができ、目的およびソース・ファイルそれぞれの行に対し、どのサブフィールドに差があるかを示すことができる。この差の程度の如何によって、派生差分ファイル内の記録が影響され得る。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソース・ファイルと目的ファイルとの対応する行のＯＰＴＩＯＮＡＬサブフィールドの差を計算することができる。
【００８３】
次いで３３８で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、コラムの差を差分ファイルに加え、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルから次の行を読み取るため３０６に戻る。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、最短ではソースおよび目的ファイルが終わるまで、ゲノム・データ・ファイルに対するこの典型的な圧縮プロセス３００を通してのループを続ける。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、コラムの差を派生差分ファイルに加え、受信されたソース・ファイルの次の対応行を読み取るために３０６に戻る。
【００８４】
ここで図５を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって用いられる、ＶＣＦフォーマットのゲノム・データ・ファイルに対する典型的な圧縮プロセスを示すオペレーション・フローチャートが描かれている。ＶＣＦフォーマットのソース・ファイルと目的ファイルとの行内のフィールドの対比が示されている。図５では、ソースおよび目的ファイルはＶＣＦフォーマットであ

る。
【００８５】
ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３４０で、＃ＣＨＲＯＭおよびＰＯＳが同じでないと判断した場合、３４２で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルの当該行の＃ＣＨＲＯＭおよびＰＯＳが目的ファイルの行より小さいかどうかを判断する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３４２で、ソース・ファイルの行の＃ＣＨＲＯＭおよびＰＯＳは目的ファイルの行より大きいと判断した場合、３２０で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的の記録を差分中に挿入し、ソース・ファイルから行を読み取るために３０６に戻る。
【００８６】
また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３４２で、ソース・ファイルの行の＃ＣＨＲＯＭおよびＰＯＳは目的ファイル中の対応する行より小さいと判断した場合、３２２で、差分中の対応する行は除去され、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルから次の行を読み取るために３０６に戻る。
【００８７】
また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３４０で、ソースおよび目的ファイルのそれぞれの行の＃ＣＨＲＯＭおよびＰＯＳは同じであると判断した場合、３４４で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルのそれぞれの行のＲＥＦおよびＡＬＴが同じであるかどうかを判断する。前述の例を続けると、受信されたソースおよび目的ファイルは、事前にＶＣＦフォーマットであると判断されていた。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、図５に進んでおり、本図では、受信されたソースおよび目的ファイルの第一行が、＃ＣＨＲＯＭおよびＰＯＳが同じかどうかを判断するために評価される。受信されたソース・ファイルと目的ファイルとの対応する行の＃ＣＨＲＯＭは２１であり、それぞれの行に対するＰＯＳは３である。したがって、ソースおよび目的ファイルの対応する行の＃ＣＨＲＯＭおよびＰＯＳは同じであり、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルの行中のＲＥＦ、ＡＬＴが目的ファイル中の対応する行と同じであるかどうかを判断するために３４４に進む。
【００８８】
ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３４４で、ソース・ファイルと目的ファイルとの行のＲＥＦおよびＡＬＴは同じでないと判断した場合、３２６で、目的ファイルからの当該行はバッファに保存され、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ファイルから目的行を読み取るために３１２に戻る。
【００８９】
また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、３４４で、ソースおよび目的ファイルの行のＲＥＦおよびＡＬＴは同じであると判断した場合、３４６で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＩＮＦＯフィールドを除いたコラムの差を計算する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、コラムの差を調べ、どのコラムがかかる差を含むかを示すことができる。差のあるコラムの位置およびコラムの差の程度の如何によって、派生差分ファイル内の記録が影響され得る。前述の例を続けると、両方の行のＲＥＦはＧであり、両方の行のＡＬＴはＣである。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソースおよび目的ファイルの両方の行のＡＬＴおよびＲＥＦは同じであると判断する。ソースおよび目的ファイルの行のＲＥＦおよびＡＬＴが同じなので、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、一切のコラムの差を計算するために、ＶＣＦフォーマットのソースおよび目的ファイル中に含まれる８つの必須の固定コラム全てを通貫して検索する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＩＤおよびＱＵＡＬの２つのコラム中にいくつかの差を生成する。
【００９０】

次いで３４８で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＩＮＦＯフィールドをサブフィールドに構文解析する。ＩＮＦＯフィールド（すなわち、ＶＣＦファイルの第八コラム）は、ソースまたは目的ファイルに関する付加情報を含んでよく、ソース・ファイルと目的ファイルとのそれぞれの行の間の相似点および差を判断するために検索することができる。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＶＣＦフォーマットのソースおよび目的ファイルの諸行のＩＮＦＯまたは第八コラムを通貫して検索する。ソース・ファイルの行中の各ＩＮＦＯサブフィールドのフォーマットは、メタ情報中にＤＰ＝１５４；ＭＱ＝５２；Ｈ２として指定されている。また一方、目的ファイルの行中の各ＩＮＦＯサブフィールドのフォーマットは、ＤＰ＝１５９；ＭＱ＝７９；Ｈ２である。

【００９１】

次いで３５０で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的ファイルとソース・ファイルとのそれぞれの行に対するこれらサブフィールドを対比し、ソース・ファイルから次の行を読み取るため３０６に戻る。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、最短ではソースおよび目的ファイルが終わるまで、３００を通してのループを続ける。前述の例を続けると、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、これらサブフィールドを対比して差を計算し、それらは派生差分ファイル中に挿入される。次いで、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、受信されたソース・ファイルの次の行を読み取るため３０６に戻る。

【００９２】

ここで図６を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａおよび１１０ｂによって用いられる、ゲノム・データ・ファイルを対比する典型的なプロセス４００を示すブロック図が描かれている。図示のように、ゲノム・データに対するファイル４０２および４０４は、行およびコラムのセットに編成することが可能である。

【００９３】

４０２のファイル１はソース・ファイルであり、４０４のファイル２は目的ファイルである。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂがこれら２つのファイルを対比するのに、４０４のマッピング位置に対比される４０２のマッピング位置に基づく行毎のインジケータが利用される。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイル４０６の記録の１つ中に含まれるデータを、目的ファイル４０８中の別の記録と対比したとき、これら２つのファイルの間の差があることを判断することができる。図４は色の差を示しているが、この差は記録内の相異なるデータを含んでよい。しかして、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、派生Ｄファイルに対し以下の結果を書込むことができる。

差分：＜一致行＞
＜第二コラム－置き換え＞＜新規の値＞
＜一致行＞
＜一致行＞

【００９４】

これによれば、４０８中の新規の値を除いて、ソース・ファイル４０２と目的ファイル４０４とは、対比された行４つのうちの３つが同じであり、相互に一致する。

【００９５】

ここで図７を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂによって用いられる、ＳＡＭフォーマットのゲノム・データ・ファイルの階層構造を識別するための典型的なプロセス５００を示すオペレーション・フローチャートが描かれている。図示のように、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＳＡＭの記録に対し、目的ファイルとソース・ファイルとの行を対比する際に、階層的な識別のシステムを利用する。

【００９６】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、５０２で、当該行のＰＯＳおよびＲＥＦが一致しない（すなわち、不一致）と判断した場合、５０４で、目的ファイルとソース・ファイルとは不一致であると判定され、目的ファイルとソース・ファイルとの対比は、ソース・ファイル中の次の行に移る。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的ファイルとソース・ファイルとのそれぞれの行のＰＯＳおよびＲＥＦを対比し、ＰＯＳ、ＲＥＦが同じかどうかを判断する。

【００９７】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、５０２で、ＰＯＳおよびＲＥＦが一致すると判断した場合、５０６で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソースおよび目的ファイルの行は中間的な一致であると判定し、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、５０８でソースおよび目的ファイルの次の行のＱＮＡＭＥを対比するために行を下降する。

【００９８】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、５０８で、これらの行のＱＮＡＭＥが一致しないと判断した場合、５１０で、目的ファイルとソース・ファイルとは不一致であると判定され、目的ファイルとソース・ファイルとの行の対比を終了する。また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、５０８で、ＱＮＡＭＥが一致すると判断した場合、５１２で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的およびソース・ファイルの行は中間的な一致であると判定し、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、５１４でソースおよび目的ファイルのＦＬＡＧ（ビット単位）を対比するために行を下降する。

【００９９】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、５１４で、当該行のＦＬＡＧ（ビット単位）が一致すると判断した場合、５１６で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的およびソース・ファイルは一致であると判定し、図４中のより複雑なオペレーション・フローチャートに進む。

【０１００】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、５１４で、ＦＬＡＧ（ビット単位）が一致しないと判断した場合、５１８で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ソース・ファイルと目的ファイルとの行は中間的不一致であると判定し、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、５２０で、その差が閾値（例えば、０ｘ４００）以下かどうかを判断することができる。５２０で、差が閾値以下であった場合、５２２で、目的およびソース・ファイルの当該行は一致と見なされ、一致に対する「m」が派生差分ファイルに書込まれる。また一方、５２０で、差が閾値よりも大きかった場合、５２４で、ソース・ファイルと目的ファイルとの行は不一致であると判断され、このソースおよび目的ファイルの行の階層的な識別は終了する。

【０１０１】

ここで図８を参照すると、少なくとも１つの実施形態による、ゲノムベース差分圧縮プログラム１１０ａおよび１１０ｂによって用いられる、ＶＣＦフォーマットのゲノム・データ・ファイルの階層構造を識別するための簡単で典型的なプロセス６００を示すオペレーション・フローチャートが描かれている。図示のように、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ＶＣＦの記録に対する目的およびソース・ファイルの行を対比する際に、階層的な識別のシステムを利用する。

【０１０２】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、６０２で、当該行の＃ＣＨＲＯＭおよびＰＯＳが一致しない（すなわち、不一致）と判断した場合、６０４で、目的ファイルとソース・ファイルとは不一致であるとして判定され、目的およびソース・ファイルの対比は終了する。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的およびソース・ファイル中の各行の＃ＣＨＲＯＭおよびＰＯＳを対比する。

【０１０３】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、６０２で、＃ＣＨＲＯＭおよびＰＯＳは一致すると判断した場合、６０６で、目的およびソース・ファイルの行は中間的な一致として判定され、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、６０８でソースおよび目的ファイルの次の行のＲＥＦおよびＡＬＴを対比するために行を下降する。

【０１０４】

ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、６０８で、当該行のＲＥＦおよびＡＬＴは一致すると判断した場合、６１２で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的およびソース・ファイルの行は一致すると判定し、目的ファイルとソース・ファイルとの行の対比は終了する。一致に対する「ｍ」が派生差分ファイル中に書込まれる。

【０１０５】

また一方、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂが、６０８で、当該行のＲＥＦおよびＡＬＴは一致しないと判断した場合、６１０で、ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、目的ファイルとソース・ファイルとの行は不一致であると判定し、ソースおよび目的ファイルの階層的識別は終了する。

【０１０６】

当然のことながら、図２～８は、１つの実施形態の単なる例示を提供しており、各種の実施形態がどのように実装できるかに関して、いかなる限定をも意味するものではない。設計および実装上の要件に基づいて、描かれた（諸）実施形態に多くの修改を加えることが可能である。

【０１０７】

図９は本発明の或る例示的な実施形態による、図１に描かれたコンピュータの内部および外部の諸コンポーネントのブロック図９００である。当然のことながら、図９は、単に一実装の例示を提供するものであり、環境に関するいかなる限定を意味するものでもなく、この環境には種々の実施形態を実装することができる。設計および実装上の要件に基づいて、描かれた環境に多くの修改を加えることが可能である。

【０１０８】

データ処理装置９０２、９０４は、マシン可読のプログラム命令を実行する能力のある任意の電子デバイスを表している。データ処理装置９０２、９０４は、スマート・フォン、コンピュータ・システム、ＰＤＡ、または他の電子デバイスを表していてよい。データ処理装置９０２、９０４によって表されてよい、コンピューティング・システム、環境、もしくは構成体またはこれらの組合せには、以下に限らないが、パーソナル・コンピュータ・システム、サーバ・コンピュータ・システム、シン・クライアント、シック・クライアント、ハンドヘルドまたはラップトップ・デバイス、マルチプロセッサ・システム、マイクロプロセッサベースのシステム、ネットワークＰＣ、ミニコンピュータ・システム、および前述のシステムまたはデバイスのいずれかを含む分散型クラウド・コンピューティング環境が含まれる。

【０１０９】

ユーザ・クライアント・コンピュータ１０２およびネットワーク・サーバ１１２は、図９に示された内部コンポーネント９０２ａ、ｂおよび外部コンポーネント９０４ａ、ｂのそれぞれのセットを含んでよい。内部コンポーネント９０２ａ、ｂのセットの各々は、１つ以上のバス９１２上の、１つ以上のプロセッサ９０６、１つ以上のコンピュータ可読ＲＡＭ９０８、および１つ以上のコンピュータ可読ＲＯＭ９１０、ならびに１つ以上のオペレーティング・システム９１４および１つ以上のコンピュータ可読有形ストレージ・デバイス９１６を含む。１つ以上のオペレーティング・システム９１４、ソフトウェア・プログラム１０８、ならびにクライアント・コンピュータ１０２中のゲノムベース差分圧縮プログラム１１０ａおよびネットワーク・サーバ１１２中のゲノムベース差分圧縮プログラム１１０ｂは、１つ以上のＲＡＭ９０８（これは、通常、キャッシュ・メモリを含む）を介しての、１つ以上のプロセッサ９０６による実行のために、１つ以上のコンピュータ可

読有形ストレージ・デバイス９１６に格納することができる。図９に示された実施形態では、コンピュータ可読有形ストレージ・デバイス９１６の各々は内部ハード・ドライブの磁気ディスク・ストレージ・デバイスである。あるいは、コンピュータ可読有形ストレージ・デバイス９１６の各々は、ＲＯＭ９１０、ＥＰＲＯＭ、フラッシュ・メモリなどの半導体ストレージ・デバイス、またはコンピュータ・プログラムおよびデジタル情報を格納できる任意の他のコンピュータ可読有形ストレージ・デバイスである。

【０１１０】

また、内部コンポーネント９０２ａ、ｂの各セットは、ＣＤ－ＲＯＭ、ＤＶＤ、メモリ・スティック、磁気テープ、磁気ディスク、光ディスク、または半導体ストレージ・デバイスなどの、１つ以上の携帯型コンピュータ可読有形ストレージ・デバイス９２０から読み取りおよび書込みをするためのＲ／Ｗドライブまたはインターフェース９１８を含む。ソフトウェア・プログラム１０８、およびゲノムベース差分圧縮プログラム１１０ａおよび１１０ｂなどのソフトウェア・プログラムは、それぞれの携帯型コンピュータ可読有形ストレージ・デバイス９２０の１つ以上に格納し、それぞれのＲ／Ｗドライブまたはインターフェース９１８を介して読み取り、それぞれのハード・ドライブ９１６中にロードすることができる。

【０１１１】

また、内部コンポーネント９０２ａ、ｂの各セットは、ＴＣＰ／ＩＰアダプタ・カード、ワイヤレスＷｉ－Ｆｉインターフェース・カード、または３Ｇもしくは４Ｇワイヤレス・インターフェース・カード、または他の有線またはワイヤレス通信リンクなどの、ネットワーク・アダプタ（もしくはスイッチ・ポート・カード）またはインターフェース９２２も含んでよい。クライアント・コンピュータ１０２中のソフトウェア・プログラム１０８およびゲノムベース差分圧縮プログラム１１０ａ、ならびにネットワーク・サーバ・コンピュータ１１２中のゲノムベース差分圧縮プログラム１１０ｂは、ネットワーク（例えば、インターネット、ローカル・エリア・ネットワーク、またはその他の広域ネットワーク）と、それぞれのネットワーク・アダプタまたはインターフェース９２２とを介して、外部のコンピュータ（例えば、サーバ）からダウンロードすることが可能である。ソフトウェア・プログラム１０８、ならびにクライアント・コンピュータ１０２中のゲノムベース差分圧縮プログラム１１０ａ、およびネットワーク・サーバ１１２中のゲノムベース差分圧縮プログラム１１０ｂは、ネットワーク・アダプタ（もしくはスイッチ・ポート・アダプタ）またはインターフェース９２２から、それぞれのハード・ドライブ９１６の中にロードすることができる。このネットワークは、銅線、光ファイバ、ワイヤレス送信、ルータ、ファイアウォール、交換機、ゲートウェイ・コンピュータ、もしくはエッジ・サーバ、またはこれらの組合せを含んでよい。

【０１１２】

外部コンポーネント９０４ａ、ｂの各セットには、コンピュータ・ディスプレイ・モニタ９２４、キーボード９２６、およびコンピュータ・マウス９２８を含めることができる。また、外部コンポーネント９０４ａ、ｂは、タッチ・スクリーン、仮想キーボード、タッチ・パッド、ポインティング・デバイス、および他の人間用のインターフェース・デバイスも含んでよい。内部コンポーネント９０２ａ、ｂの各セットは、コンピュータ・ディスプレイ・モニタ９２４、キーボード９２６、およびコンピュータ・マウス９２８にインターフェース接続するためのデバイス・ドライバ９３０をさらに含む。デバイス・ドライバ９３０、Ｒ／Ｗドライブもしくはインターフェース９１８、およびネットワーク・アダプタもしくはインターフェース９２２は、ハードウェア、および（ストレージ・デバイス９１６もしくはＲＯＭ９１０またはその両方に格納された）ソフトウェアを含む。

【０１１３】

本開示は、クラウド・コンピューティングの詳細な説明を含むが、前もって当然のことながら、本明細書で述べる教示の実装は、クラウド・コンピューティング環境に限定されない。それどころか、本開示の諸実施形態は、現在知られた、または今後開発される任意の他の種類のコンピューティング環境に合わせて実装することができる。

【０１１４】
クラウド・コンピューティングは、最小の管理作業またはサービスのプロバイダとのやり取りで、迅速に立ち上げてリリースすることができる、構成可能なコンピューティング・リソース（例えば、ネットワーク、ネットワーク帯域幅、サーバ、処理、メモリ、ストレージ、アプリケーション、仮想マシン、およびサービス）の共用のプールへの、便利でオンデマンドのネットワーク・アクセスを可能にするためのサービス・デリバリのーモデルである。このクラウド・モデルは、少なくとも５つの特徴と、少なくとも３つのサービス・モデルと、少なくとも４つの展開モデルとを含んでよい。
【０１１５】
特徴は次の通りである。
オンデマンド・セルフサービス：クラウド・コンシューマは、サービスのプロバイダとの人間のやり取りなしに、必要な場合は自動的に、サーバ時間およびネットワーク・ストレージなどのコンピューティング能力を一方的にセットアップすることができる。
広範なネットワーク・アクセス：諸能力はネットワークを介して利用可能で、異機種から成るシンまたはシック・クライアント・プラットフォーム（例えば、携帯電話、ラップトップ、およびＰＤＡ）による使用を推進する標準的なメカニズムを通してアクセスされる。
リソースのプール化：プロバイダのコンピューティング・リソースは、マルチテナント・モデルを用いる複数のコンシューマにサービスするために、デマンドに従って動的に割り当ておよび再割り当てされる各種の物理的および仮想のリソースとしてプール化される。一般にコンシューマは提供されたリソースの正確な場所の制御または知識を持たない、という点で、場所無依存性の感覚があるが、抽象化のより高位レベルでは場所（例えば、国、州、またはデータセンタ）を特定することを可能にする。
敏速な伸縮性：諸能力は、迅速にスケール・アウトするため、場合によっては自動的に、敏速且つ伸縮自在にセットアップでき、迅速にスケール・インするために敏速にリリースされる。コンシューマには、セットアップのため利用可能な諸能力が多くの場合無制限に見え、いつでもどのような量でも購入が可能である。
計量されるサービス：クラウド・システムは、サービスの種類（例えば、ストレージ、処理、帯域幅、および有効なユーザ・アカウント）に適した抽象化のいずれかのレベルで、計量機能を利用することによって、リソース使用を自動的に管理し、最適化する。リソース利用は、プロバイダおよび利用されるサービスのコンシューマの双方に透明性を提供しながら、モニタし、管理し、報告することができる。
【０１１６】
サービス・モデルは次のとおりである。
サービスとしてのソフトウェア（ＳａａＳ：Ｓｏｆｔｗａｒｅ ａｓ ａ Ｓｅｒｖｉｃｅ）：コンシューマに提供される能力は、クラウド・インフラストラクチャ上で実行されているプロバイダのアプリケーションを使うことである。これらのアプリケーションは、様々なクライアント・デバイスから、ウェブ・ブラウザ（例えば、ウェブベースのｅメール）などのシン・クライアント・インターフェースを介してアクセス可能である。コンシューマは、限られたユーザ固有のアプリケーション構成設定のあり得る例外を除いて、ネットワーク、サーバ、オペレーティング・システム、ストレージ、または個別のアプリケーション機能でさえも含め、根底にあるクラウド・インフラストラクチャを管理または制御はしない。
サービスとしてのプラットフォーム（ＰａａＳ：Ｐｌａｔｆｏｒｍ ａｓ ａ Ｓｅｒｖｉｃｅ）：コンシューマに提供される能力は、クラウド・インフラストラクチャ上に、プロバイダによってサポートされるプログラミング言語およびツールを使って生成された、コンシューマ生成の、またはコンシューマ取得のアプリケーションを展開することである。コンシューマは、ネットワーク、サーバ、オペレーティング・システム、またはストレージを含め、根底にあるクラウド・インフラストラクチャを管理または制御することはないが、これら展開されるアプリケーション、およびおそらくはアプリケーションのホステ

ィング環境設定に対する制御を有する。
サービスとしてのインフラストラクチャ（ＩａａＳ：Ｉｎｆｒａｓｔｒｕｃｔｕｒｅ ａｓ ａ Ｓｅｒｖｉｃｅ）：コンシューマに提供される能力は、コンシューマが、処理、ストレージ、ネットワーク、およびオペレーティング・システムおよびアプリケーションを含み得る、任意のソフトウェアを展開し実行することが可能な他の基本的コンピューティング・リソースをセットアップすることである。コンシューマは、根底にあるクラウド・インフラストラクチャを管理または制御することはないが、オペレーティング・システム、ストレージ、展開されるアプリケーションに対する制御、およびおそらくは選択したネットワーク・コンポーネント（例えば、ホストのファイアウォール）の限定された制御を有する。

【０１１７】
展開モデルは次のとおりである。
プライベート・クラウド：このクラウド・インフラストラクチャは、一組織のためだけに運営される。これは、その組織または第三者によって管理されてよく、自組織の構内に所在しても自組織の構外に所在してもよい。
コミュニティ・クラウド：このクラウド・インフラストラクチャは、いくつかの組織によって共有され、共有の利害関係（例えば、任務、安全要件、指針、およびコンプライアンス配慮事項）を有する特定のコミュニティをサポートする。これは、これらの組織または第三者によって管理されてよく、これら組織の構内に所在してもこれら組織の構外に所在してもよい。
パブリック・クラウド：このクラウド・インフラストラクチャは、一般公衆または大きな産業グループが利用可能なようにされており、クラウド・サービスを販売する組織によって所有されている。
ハイブリッド・クラウド：このクラウド・インフラストラクチャは、独自のエンティティに留まりながら、データおよびアプリケーションの可搬性（例えば、クラウド間の負荷バランスのためのクラウド・バースティング）を可能にする標準化されたまたは独自の技術によって一緒に結ばれた２つ以上のクラウド（プライベート、コミュニティ、またはパブリック）の合成体である。

【０１１８】
クラウド・コンピューティング環境は、ステートレスネス、弱連結、モジュール性、および意味相互運用性に焦点を合わせて方向付けられたサービスである。クラウド・コンピューティングの中心には、相互接続されたノードのネットワークを含むインフラストラクチャがある。

【０１１９】
図１０を参照すると、例示的なクラウド・コンピューティング環境１０００が描かれている。図示のように、クラウド・コンピューティング環境１０００は、例えば、携帯情報端末（ＰＤＡ：ｐｅｒｓｏｎａｌ ｄｉｇｉｔａｌ ａｓｓｉｓｔａｎｔ）もしくはセルラ電話１０００Ａ、デスクトップ・コンピュータ１０００Ｂ、ラップトップ・コンピュータ１０００Ｃ、もしくは車載コンピュータ・システム１０００Ｎ、またはこれらの組合せなど、クラウド・コンシューマによって使用されるローカルのコンピューティング・デバイスが通信可能な１つ以上のクラウド・コンピューティング・ノード１００を含む。ノード１００は、相互に通信することができる。これらは、前述したプライベート、コミュニティ、パブリック、またはハイブリッド・クラウド、またはこれらの組合せなど、１つ以上のネットワークに物理的にまたは仮想的にグループ化する（図示せず）ことが可能である。これは、クラウド・コンピューティング環境１０００が、クラウド・コンシューマがローカル・コンピューティング・デバイス上にリソースを維持する必要のない、サービスとしてのインフラストラクチャ、プラットフォーム、もしくはソフトウェア、またはこれらの組合せを提供することを可能にする。当然のことながら、図１０に示されたコンピューティング・デバイス１０００Ａ～Ｎの種類は例示だけを意図されたものであり、コンピューティング・ノード１００およびクラウド・コンピューティング環境１０００は、任意の

種類のネットワークもしくは（例えば、ウェブ・ブラウザを使って）ネットワーク・アドレス指定が可能な接続またはその両方を介して、任意の種類のコンピュータ化デバイスと通信することができる。

【０１２０】

ここで図１１を参照すると、クラウド・コンピューティング環境１０００によって設けられる機能的抽象化層のセット１１００が示されている。前もって当然のことながら、図１１に示されたコンポーネント、層、および機能は例示だけを意図されたものであり、本発明の諸実施形態はこれに限定されない。図示のように、以下の層および対応する機能が設けられている。

【０１２１】

ハードウェアおよびソフトウェア層１１０２は、ハードウェアおよびソフトウェア・コンポーネントを含む。ハードウェア・コンポーネントの例には、メインフレーム１１０４、ＲＩＳＣ（Ｒｅｄｕｃｅｄ Ｉｎｓｔｒｕｃｔｉｏｎ Ｓｅｔ Ｃｏｍｐｕｔｅｒ（縮小命令セット・コンピュータ））アーキテクチャ・ベースのサーバ１１０６、サーバ１１０８、ブレード・サーバ１１１０、ストレージ・デバイス１１１２、ならびにネットワークおよびネットワーキング・コンポーネント１１１４が含まれる。いくつかの実施形態において、ソフトウェア・コンポーネントは、ネットワーク・アプリケーション・サーバ・ソフトウェア１１１６およびデータベース・ソフトウェア１１１８を含む。

【０１２２】

仮想化層１１２０は、仮想サーバ１１２２、仮想ストレージ１１２４、仮想プライベート・ネットワークを含む仮想ネットワーク１１２６、仮想アプリケーションおよびオペレーティング・システム１１２８、ならびに仮想クライアント１１３０、の仮想エンティティの諸例を設けることが可能な抽象化層を提供する。

【０１２３】

一例において、管理層１１３２には以下に記載の機能を備えることが可能である。リソース・セットアップ１１３４は、クラウド・コンピューティング環境内でタスクを実行するために利用されるコンピューティング・リソースおよび他のリソースの動的な調達を提供する。計量および価格計算１１３６は、クラウド・コンピューティング環境内でリソースが利用されるのに応じて、これらリソースの消費に対するコストの追跡、および請求書または納入書の作成を提供する。一例において、これらのリソースには、アプリケーション・ソフトウェア・ライセンスを含めてよい。セキュリティは、クラウド・コンシューマおよびタスクに対する身元検証、ならびにデータおよび他のリソースに対する保護を提供する。ユーザ・ポータル１１３８は、コンシューマおよびシステム管理者に対し、クラウド・コンピューティング環境へのアクセスを提供する。サービス・レベル管理１１４０は、要求されるサービス・レベルが満たされるように、クラウド・コンピューティング・リソースの割り当ておよび管理を提供する。サービス・レベル合意（ＳＬＡ：Ｓｅｒｖｉｃｅ Ｌｅｖｅｌ Ａｇｒｅｅｍｅｎｔ）計画および達成１１４２は、ＳＬＡに沿って将来の必要性が予期されるクラウド・コンピューティング・リソースに対する事前配置およびそれらリソースの調達を提供する。

【０１２４】

作業負荷層１１４４は、クラウド・コンピューティング環境で使用できる機能の諸例を提供する。この層から提供できる作業負荷および機能の例には、マッピングおよびナビゲーション１１４６、ソフトウェア開発およびライフサイクル管理１１４８、仮想教室教育配信１１５０、データ解析処理１１５２、トランザクション処理１１５４、およびゲノムベース差分圧縮１１５６が含まれる。ゲノムベース差分圧縮プログラム１１０ａ、１１０ｂは、ゲノム・データ・ファイルに対する差分ファイルを圧縮する方法を提供する。

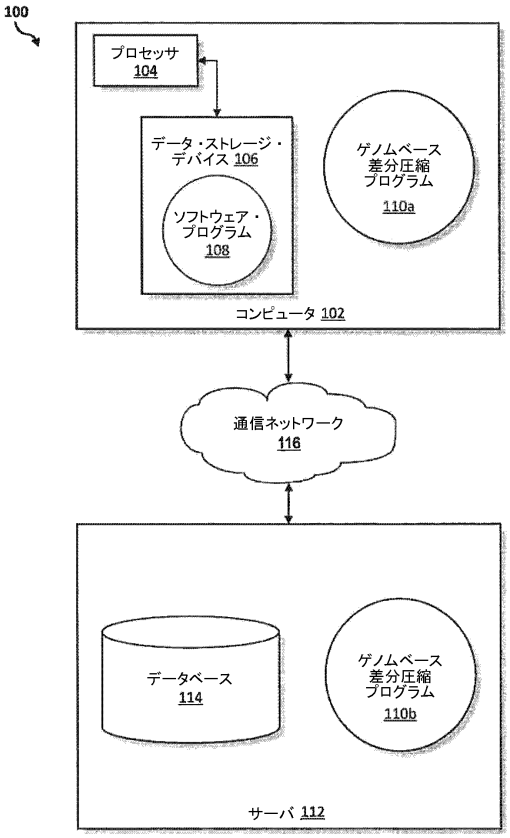
【０１２５】

本発明の様々な実施形態の説明は、例示目的で提示されたもので、網羅的であることも、または開示された態様に限定することも意図されていない。当業者には、記載された諸態様の範囲から逸脱することのない多くの修改および別形が明白であろう。本明細書で用

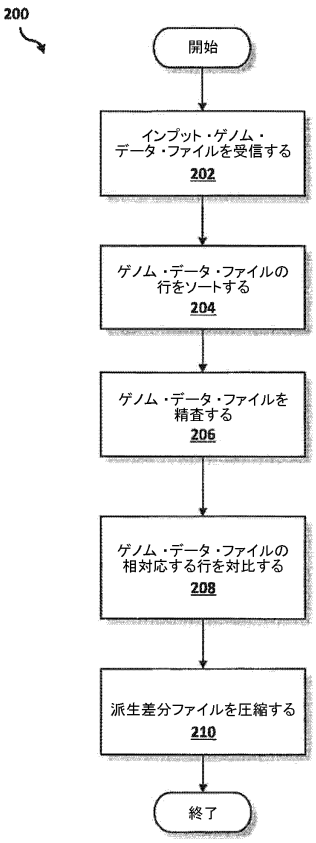
いられた用語は、諸実施形態の原理、実際上の応用、または市販の技術の技術的な改良を最善に説明し、または他の当業者が本明細書に開示された諸実施形態を理解できるように選択されたものである。

【図面】

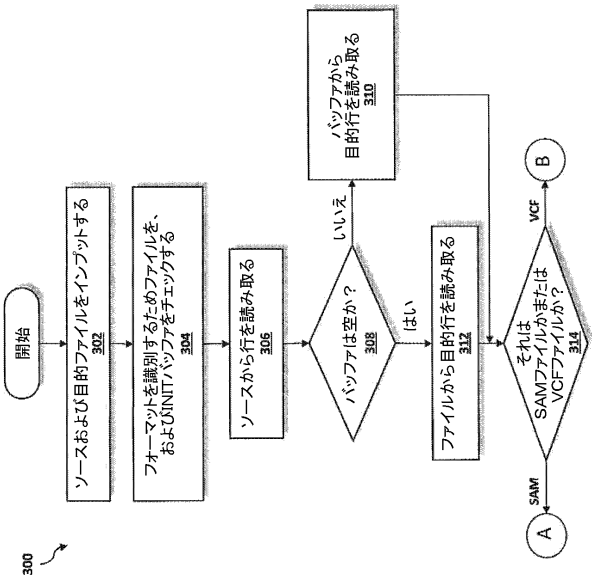
【図１】



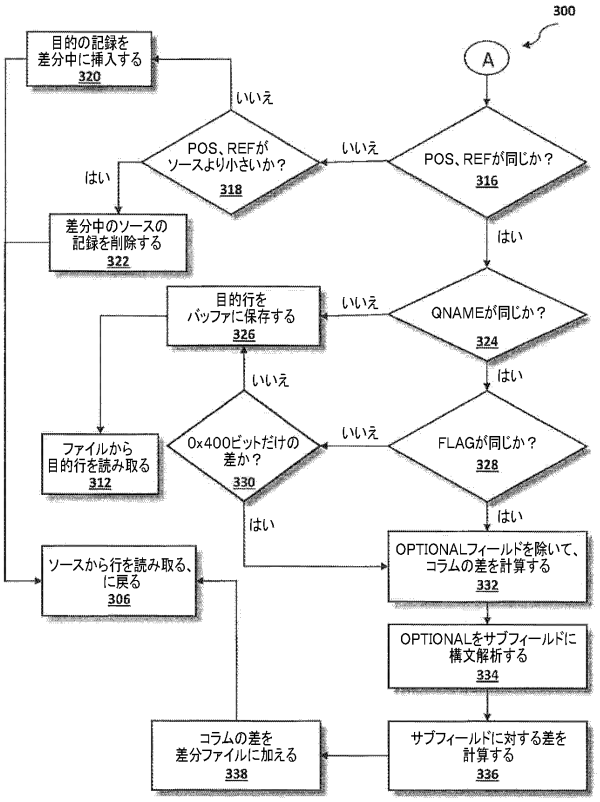
【図２】



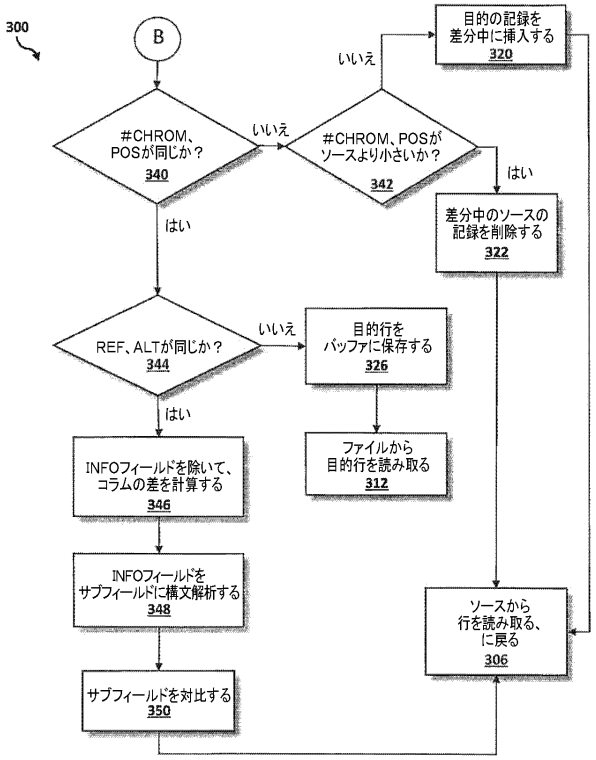
【図３】


開始
ソースおよび目的ファイルをインプットする
302
フォーマットを識別するためファイルを、
およびINITバッファをチェックする
304
ソースから行を読み取る
306
バッファは空か？
308
いいえ
はい
バッファから
目的行を読み取る
310
ファイルから目的行を読み取る
312
それは
SAMファイルかまたは
VCFファイルか？
314
VCF
SAM
B
A
300

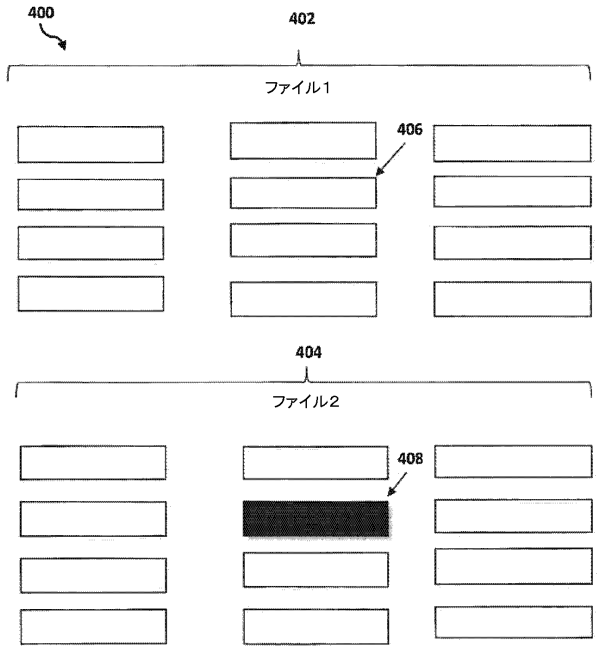
【図４】


300
A
目的の記録を
差分中に挿入する
320
いいえ
POS、REFが
ソースより小さいか？
318
いいえ
POS、REFが同じか？
316
はい
差分中のソースの
記録を削除する
322
はい
目的行を
バッファに保存する
326
いいえ
QNAMEが同じか？
324
はい
いいえ
ファイルから
目的行を読み取る
312
0x400ビットだけの
差か？
330
いいえ
FLAGが同じか？
328
はい
はい
ソースから行を読み取る、
に戻る
306
OPTIONALフィールドを除いて、
コラムの差を計算する
332
OPTIONALをサブフィールドに
構文解析する
334
コラムの差を
差分ファイルに加える
338
サブフィールドに対する差を
計算する
336

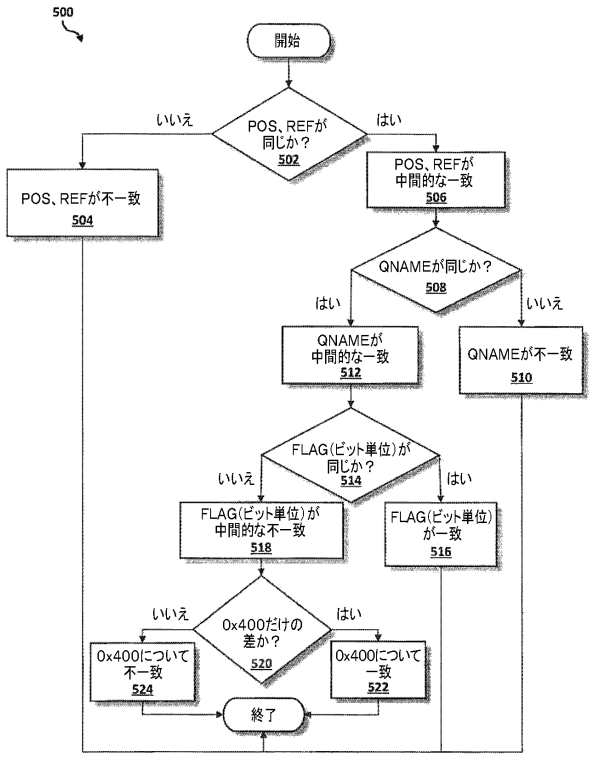
【図５】


300
B
#CHROM、
POSが同じか？
340
いいえ
#CHROM、POSが
ソースより小さいか？
342
いいえ
目的の記録を
差分中に挿入する
320
はい
差分中のソースの
記録を削除する
322
はい
REF、ALTが同じか？
344
いいえ
目的行を
バッファに保存する
326
はい
INFOフィールドを除いて、
コラムの差を計算する
346
ファイルから
目的行を読み取る
312
INFOフィールドを
サブフィールドに構文解析する
348
ソースから
行を読み取る、
に戻る
306
サブフィールドを対比する
350

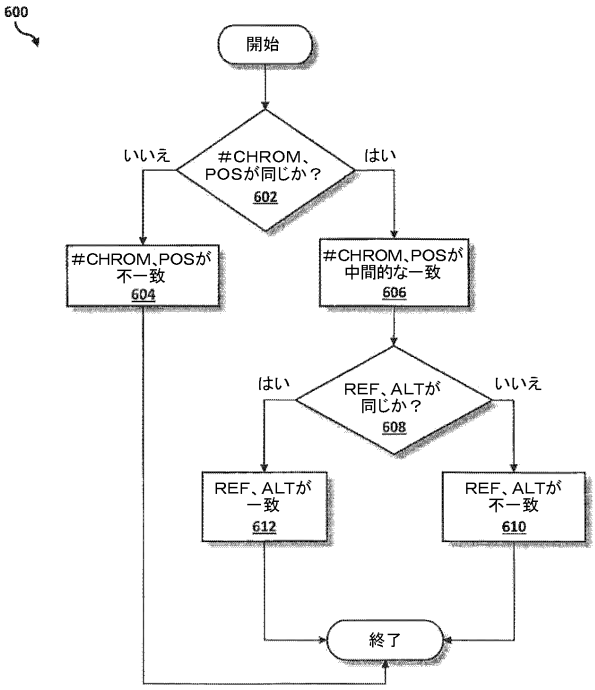
【図６】


400
402
ファイル1
406
404
ファイル2
408

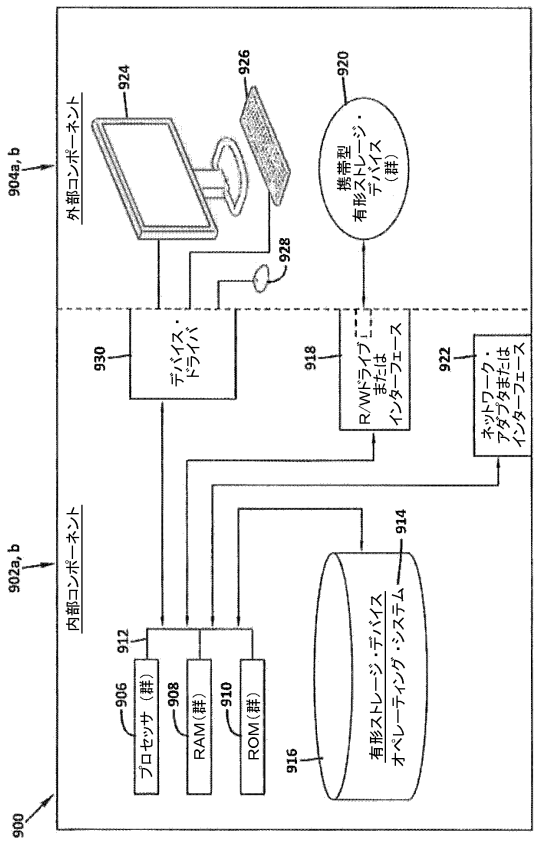
【図７】


500
開始
POS、REFが同じか？
502
いいえ
はい
POS、REFが不一致
504
POS、REFが中間的な一致
506
QNAMEが同じか？
508
はい
いいえ
QNAMEが中間的な一致
512
QNAMEが不一致
510
FLAG(ビット単位)が同じか？
514
いいえ
はい
FLAG(ビット単位)が中間的な不一致
518
FLAG(ビット単位)が一致
516
0x400だけの差か？
520
いいえ
はい
0x400について不一致
524
0x400について一致
522
終了

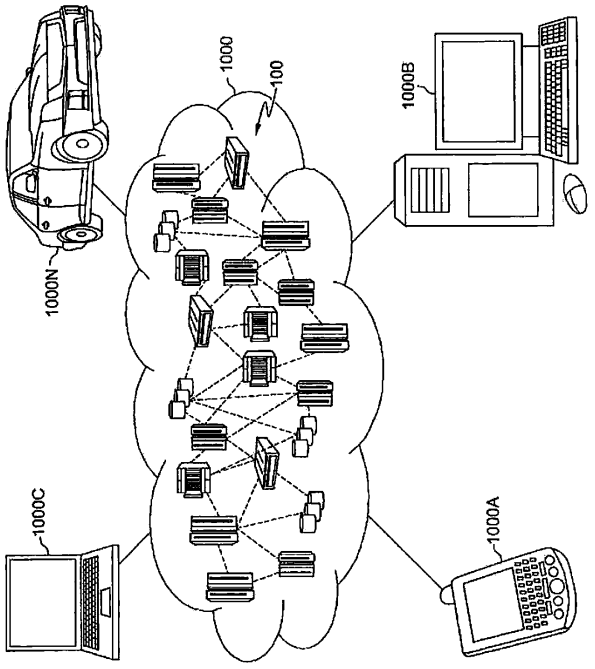
【図８】


600
開始
#CHROM、POSが同じか？
602
いいえ
はい
#CHROM、POSが不一致
604
#CHROM、POSが中間的な一致
606
REF、ALTが同じか？
608
はい
いいえ
REF、ALTが一致
612
REF、ALTが不一致
610
終了

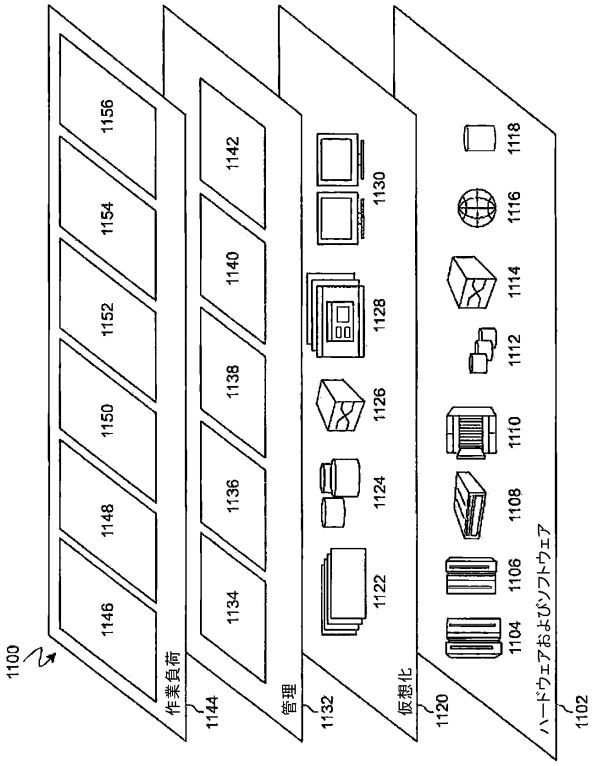
【図９】


900
902a, b
内部コンポーネント
904a, b
外部コンポーネント
912
プロセッサ（群）
906
RAM(群)
908
ROM(群)
910
有形ストレージ・デバイス・オペレーティング・システム
914
916
デバイス・ドライバ
930
R/Wドライブまたはインターフェース
918
ネットワーク・アダプタまたはインターフェース
922
携帯型有形ストレージ・デバイス（群）
920
924
926
928

【図１０】


1000
100
1000B
1000N
1000C
1000A

【図１１】


1100
作業負荷
1144
管理
1132
仮想化
1120
ハードウェアおよびソフトウェア
1102
1104
1106
1108
1110
1112
1114
1116
1118
1122
1124
1126
1128
1130
1134
1136
1138
1140
1142
1146
1148
1150
1152
1154
1156

フロントページの続き

(72)発明者　マハラナ、アジャシャ
アメリカ合衆国９５２１－６０９９ カリフォルニア州 サンノゼ ハリー・ロード ６５０
(72)発明者　コンスタンチネスキュ、ミハイル、コルネリウ
アメリカ合衆国９５２１－６０９９ カリフォルニア州 サンノゼ ハリー・ロード ６５０
審査官　山内 裕史
(56)参考文献　米国特許出願公開第２０１３／０１３２３５３（ＵＳ，Ａ１）
米国特許出願公開第２００８／００７７６０７（ＵＳ，Ａ１）
米国特許出願公開第２０１１／０１１９２４０（ＵＳ，Ａ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ　５／００ － ９９／００