(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2023-551154
(P2023-551154A)
(43)公表日　令和5年12月7日(2023.12.7)

(51)国際特許分類　FI
G06F 16/17 (2019.01)　G06F 16/17 100

審査請求　未請求　予備審査請求　未請求　（全37頁）

(21)出願番号　特願2023-530194(P2023-530194)
(86)(22)出願日　令和3年11月11日(2021.11.11)
(85)翻訳文提出日　令和5年5月18日(2023.5.18)
(86)国際出願番号　PCT/EP2021/081320
(87)国際公開番号　WO2022/106289
(87)国際公開日　令和4年5月27日(2022.5.27)
(31)優先権主張番号　63/116,203
(32)優先日　令和2年11月20日(2020.11.20)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　63/226,791
(32)優先日　令和3年7月29日(2021.7.29)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA
最終頁に続く

(71)出願人　590000248
コーニンクレッカ フィリップス エヌヴェ
Koninklijke Philips N.V.
オランダ国 5656 アーヘー アインドーフェン ハイテック キャンパス 52
High Tech Campus 52, 5656 AG Eindhoven, Netherlands
(74)代理人　110001690
弁理士法人M&Sパートナーズ
(72)発明者　チャン イー ヒム
オランダ国 5656 アーエー アインドーフェン ハイ テック キャンパス 5
最終頁に続く

(54)【発明の名称】　自動構成可能なデータ圧縮フレームワークのためのシステム及び方法

(57)【要約】
データファイルを圧縮し、展開するための方法（１００）であって、（ｉ）複数の異なる属性を含む圧縮のためのデータファイルを受信すること（１２０）と、（ｉｉ）複数の異なる属性の第１の属性を識別すること（１３０）と、（ｉｉｉ）複数の圧縮タイプ及び／又は構成を選択すること（１４０）と、（ｉｖ）選択された複数の圧縮タイプ及び／又は構成の各々を使用して、識別された第１の属性のための受信されたデータファイルからのデータの少なくとも一部を圧縮すること（１５０）と、（ｖ）選択された複数の圧縮タイプ及び／又は構成のうちのいずれが圧縮のために最も好適であるかを決定すること（１６０）と、（ｖｉ）選択された複数の圧縮タイプ及び／又は構成の識別を含む圧縮パラメータデータ構造を生成すること（１７０）と、（ｖｉｉ）圧縮されたデータファイルを生成するために、第１の属性のための受信されたデータファイルからのデータを圧縮すること（１８０）と、（ｖｉｉｉ）圧縮パラメータデータ構造と圧縮されたデータファイルとを記憶すること（１９０）とを含む、データファイルを圧縮し、展開するため

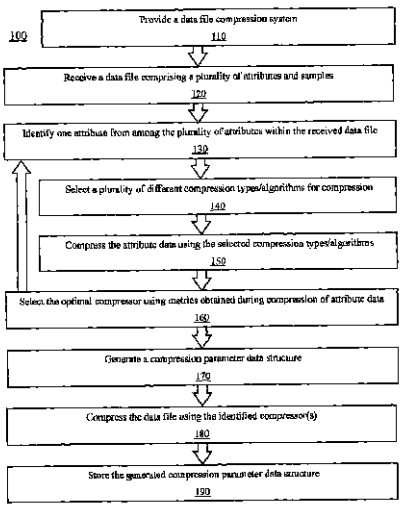


FIG. 1

【特許請求の範囲】
【請求項１】
ファイル圧縮／展開システムを使用してデータファイルを圧縮／展開するための方法であって、前記方法は、
複数の異なる属性を含む、圧縮のためのデータファイルを受信するステップと、
前記複数の異なる属性のうちの第１の属性を識別し、前記識別された第１の属性の仕様を属性パラメータデータ構造に記憶するステップと、
前記識別された第１の属性に基づいて、事前決定された複数の圧縮タイプと圧縮構成とから複数の圧縮タイプ及び／又は圧縮構成を選択するステップと、
前記選択された複数の圧縮タイプ及び／又は圧縮構成の各々を使用して、前記識別された第１の属性のための前記受信されたデータファイルからの前記データの少なくとも一部を個別に圧縮するステップであって、１つ又は複数のメトリクスが各個別の圧縮について測定される、ステップと、
前記個別の圧縮の各々の前記１つ又は複数のメトリクスに基づいて、前記選択された複数の圧縮タイプ及び／又は圧縮構成のうちのいずれが圧縮のために最も好適であるかを決定するステップと、
前記選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、前記識別された第１の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の識別を有する、圧縮パラメータデータ構造を生成するステップと、
圧縮されたデータファイルを生成するために、前記識別された圧縮タイプを使用して、前記第１の属性のための前記受信されたデータファイルからの前記データを圧縮するステップと、
前記属性パラメータデータ構造と前記圧縮パラメータデータ構造と前記圧縮されたデータファイルとを記憶するステップであって、前記記憶された圧縮パラメータデータ構造が、前記圧縮されたデータファイルの展開のために取り出され、使用されるように構成された、ステップと
を有する、データファイルの圧縮／展開のための方法。
【請求項２】
前記複数の異なる属性のうちの少なくとも第２の属性のために、前記識別するステップ、前記選択するステップ、前記圧縮するステップ、及び、前記決定するステップを繰り返すステップをさらに有し、前記圧縮パラメータデータ構造が、前記選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、前記識別された第２の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の仕様を含む、請求項１に記載の方法。
【請求項３】
前記記憶された圧縮されたデータファイルを展開するステップをさらに有し、前記展開するステップが、前記第１の属性のための展開タイプを識別するために前記圧縮パラメータデータ構造を調べるステップを有する、請求項１に記載の方法。
【請求項４】
前記識別された第１の属性のための前記受信されたデータファイルからの前記データを圧縮することが可能であるので、前記選択された複数の圧縮タイプ及び／又は圧縮構成の各々が選択される、請求項１に記載の方法。
【請求項５】
前記データファイルがゲノムデータファイルである、請求項１に記載の方法。
【請求項６】
前記１つ又は複数のメトリクスが前記圧縮のための圧縮比と処理時間とのうちの１つ又は複数を含む、請求項１に記載の方法。
【請求項７】
前記対応する圧縮が前記複数の選択された圧縮タイプ及び／又は圧縮構成のうちの他の圧縮タイプ及び／又は圧縮構成に対してより高い圧縮比及び／又はより速い処理時間を有するときの圧縮タイプが最も好適である、請求項１に記載の方法。

【請求項８】
ユーザインターフェースを介して、前記生成された圧縮パラメータデータ構造及び／又は前記圧縮されたデータファイルについての情報を与えるステップをさらに有する、請求項１に記載の方法。
【請求項９】
ユーザが、属性を圧縮するために適用されるべき圧縮器タイプ及び／又は構成を選択することができるように、前記識別された圧縮タイプ及び／又は圧縮構成のうちの１つ又は複数についての情報を、ユーザインターフェースを介して前記ユーザに与えるステップをさらに有する、請求項１に記載の方法。
【請求項１０】
複数の異なる属性を含む、圧縮のためのデータファイルと、
プロセッサと
を含む、データファイルを圧縮／展開するためのシステムであって、前記プロセッサは、
（ｉ）前記複数の異なる属性のうちの第１の属性を識別し、前記識別された第１の属性の仕様を属性パラメータデータ構造に記憶することと、
（ｉｉ）前記識別された第１の属性に基づいて、事前決定された複数の圧縮タイプ及び／又は圧縮構成から複数の圧縮タイプ及び／又は圧縮構成を選択することと、
（ｉｉｉ）前記選択された複数の圧縮タイプ及び／又は圧縮構成の各々を使用して、前記識別された第１の属性のための前記受信されたデータファイルからの前記データの少なくとも一部を個別に圧縮することであって、１つ又は複数のメトリクスが各個別の圧縮について測定される、前記データの少なくとも一部を個別に圧縮することと、
（ｉｖ）前記個別の圧縮の各々の１つ又は複数のメトリクスに基づいて、前記選択された複数の圧縮タイプ及び／又は圧縮構成のうちのいずれが圧縮のために最も好適であるかを決定することと、
（ｖ）前記選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、前記識別された第１の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の識別を有する、圧縮パラメータデータ構造を生成することと、
（ｖｉ）圧縮されたデータファイルを生成するために、前記識別された圧縮タイプを使用して、前記第１の属性のための前記受信されたデータファイルからの前記データを圧縮することと、
（ｖｉｉ）前記属性パラメータデータ構造と前記圧縮パラメータデータ構造と前記圧縮されたデータファイルとを記憶することであって、前記記憶された圧縮パラメータデータ構造が、前記圧縮されたデータファイルの展開のために取り出され、使用されるように構成された、記憶することと
を行うように構成された、データファイルを圧縮／展開するためのシステム。
【請求項１１】
前記プロセッサが、前記複数の異なる属性のうちの少なくとも第２の属性のために、前記識別することと、前記選択することと、前記圧縮することと、前記決定することとを繰り返すように構成され、前記圧縮パラメータデータ構造が、前記選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、前記識別された第２の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の識別を含む、請求項１０に記載のシステム。
【請求項１２】
前記プロセッサが、前記記憶された圧縮されたデータファイルを展開するように構成され、展開することが、前記第１の属性のための展開タイプを識別するために前記圧縮パラメータデータ構造を調べることを有する、請求項１０に記載のシステム。
【請求項１３】
前記識別された第１の属性のための前記受信されたデータファイルからの前記データを圧縮することが可能であるので、前記選択された複数の圧縮タイプ及び／又は圧縮構成の各々が選択される、請求項１０に記載のシステム。
【請求項１４】

前記１つ又は複数のメトリクスが前記圧縮のための圧縮比と処理時間とのうちの１つ又は複数を含む、請求項１０に記載のシステム。
【請求項１５】
前記システムが、前記生成された圧縮パラメータデータ構造及び／又は前記圧縮されたデータファイルについての情報を与えるように構成されたユーザインターフェースを含む、請求項１０に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
[0001] 本開示は、一般に、ファイル圧縮／展開システムを使用してデータファイルを圧縮するための方法及びシステムを対象とする。
【背景技術】
【０００２】
[0002] 多くのデータファイルのサイズが増大し続けていることを考えると、効率的なファイル圧縮方法及びシステムはますます重要である。単一のゲノムデータファイルは、例えば、注釈及び他のデータ属性とともに、多数のサンプルのゲノムにわたる数百万又は何十億ものロケーションについての配列決定結果を含み得る。バリアントコールフォーマット（ＶＣＦ）ファイルは、例えば、データ属性を含む何百万もの行と多数の列とをもつ遺伝子配列変異を含み得る。これらのデータファイルは、長期間にわたって記憶されなければならず、１００ギガバイト又は１０００ギガバイトの極めて大きいサイズになることがある。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
[0003] データファイルの圧縮は一般に１つ又は複数のデータ圧縮器を含む。データファイルに直接適用され得る多数の一般的な圧縮器があるが、それらは、データ冗長性の効果的な削除のために、データセット内の個々のフィールドのデータの類似性又は予測可能性を十分に活用しないので、それらの圧縮性能は、通常、準最適である。圧縮比を改善するための１つの手法は、したがって、最初に、データセットを、別個の特性をもつフィールドに分割し、次いで、フィールドの特定のデータモデルを最大限に利用することができる圧縮器を用いて、フィールドの各々を個別に圧縮することである。しかしながら、これは、一般に、データフィールドとそれらの関連する圧縮器とが１つ１つのファイルタイプについて事前定義され、符号化及び復号プロセスが、圧縮及び展開のために使用されるソフトウェアにおいて相応してハードコーディングされることを必要とする。そのような手法の柔軟性のなさは、出現するデータタイプに適応すること、又は新しいアルゴリズムを採用することを困難にする。
【課題を解決するための手段】
【０００４】
[0004] データファイルの効率的で適応可能な圧縮及び展開のための方法及びシステムの継続的な必要がある。本開示は、圧縮性能に基づいてデータ圧縮器の自動化された識別及び／又は構成を使用してデータファイルを圧縮するための発明的方法及びシステムを対象とする。ファイル圧縮／展開システムは、複数の異なるサンプルの各々についての複数の異なる属性を含む、圧縮のためのデータファイルを受信する。本システムは、複数の異なる属性のうちの第１の属性を識別し、第１の属性に基づいて、事前定義された複数の圧縮タイプ及び構成から、複数の異なる圧縮タイプ及び／又は圧縮構成を選択する。本システムは、選択された複数の圧縮タイプ及び／又は圧縮構成の各々を使用して、識別された第１の属性のための受信されたゲノムデータファイルからのデータを個別に圧縮し、選択された複数の圧縮タイプ及び／又は圧縮構成のうちのいずれが圧縮のために最も好適であるかを決定する。次いで、本システムは、選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、識別された第１の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧

縮構成の識別を有する圧縮パラメータデータ構造を生成し、圧縮されたゲノムデータファイルを生成するために、識別された圧縮タイプ及び／又は圧縮構成を使用して、第１の属性のための受信されたゲノムデータファイルからのデータを圧縮する。圧縮パラメータデータ構造及び圧縮されたゲノムデータファイルは、次いで、記憶され、展開のために使用され得る。

【０００５】

［0005］ 一般に、一態様では、ファイル圧縮／展開システムを使用してゲノムデータファイルを圧縮／展開するための方法が提供される。本方法は、（ｉ）データファイルが複数の異なる属性を含む、圧縮のためのデータファイルを受信することと、（ｉｉ）複数の異なる属性のうちの第１の属性を識別し、識別された第１の属性の仕様を属性パラメータデータ構造に記憶することと、（ｉｉｉ）識別された第１の属性に基づいて、事前定義された複数の圧縮タイプ及び／又は圧縮構成から複数の圧縮タイプ及び／又は圧縮構成を選択することと、（ｉｖ）選択された複数の圧縮タイプ及び／又は圧縮構成の各々を使用して、識別された第１の属性のための受信されたデータファイルからのデータの少なくとも一部を個別に圧縮することであって、１つ又は複数のメトリクスが各個別の圧縮について測定される、圧縮することと、（ｖ）個別の圧縮の各々の１つ又は複数のメトリクスに基づいて、選択された複数の圧縮タイプ及び／又は圧縮構成のうちのいずれが圧縮のために最も好適であるかを決定することと、（ｖｉ）選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、識別された第１の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の識別を有する、圧縮パラメータデータ構造を生成することと、（ｖｉｉ）圧縮されたデータファイルを生成するために、識別された圧縮タイプを使用して、第１の属性のための受信されたデータファイルからのデータを圧縮することと、（ｖｉｉｉ）属性パラメータデータ構造と圧縮パラメータデータ構造と圧縮されたデータファイルとを記憶することであって、記憶された圧縮パラメータデータ構造が、圧縮されたデータファイルの展開のために取り出され、使用されるように構成された、記憶することとを含む。

【０００６】

［0006］ 一実施形態によれば、本方法は、さらに、複数の異なる属性のうちの少なくとも第２の属性のために、識別するステップ、選択するステップ、圧縮するステップ、及び、決定するステップを繰り返すことを含み、圧縮パラメータデータ構造が、選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、識別された第２の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の仕様を含む。

【０００７】

［0007］ 一実施形態によれば、本方法は、記憶された圧縮されたファイルを展開することをさらに含み、展開することは、第１の属性のための展開タイプを識別するために圧縮パラメータデータ構造を調べることを有する。

【０００８】

［0008］ 一実施形態によれば、選択された複数の圧縮タイプ及び／又は圧縮構成の各々は、それが、識別された第１の属性のための受信されたデータファイルからのデータを圧縮することが可能であるので選択される。

【０００９】

［0009］ 一実施形態によれば、データファイルはゲノムデータファイルである。

【００１０】

［0010］ 一実施形態によれば、１つ又は複数のメトリクスは前記圧縮のための圧縮比と処理時間とのうちの１つ又は複数を含む。

【００１１】

［0011］ 一実施形態によれば、対応する圧縮が複数の選択された圧縮タイプ及び／又は圧縮構成のうちの他の圧縮タイプ及び／又は圧縮構成に対してより高い圧縮比及び／又はより速い処理時間を有するときに、圧縮タイプは最も好適である。

【００１２】

［0012］ 一実施形態によれば、本方法は、ユーザインターフェースを介して、生成さ

れた圧縮パラメータデータ構造及び／又は圧縮されたデータファイルについての情報を与えることをさらに含む。

【００１３】

［0013］ 一実施形態によれば、本方法は、ユーザが、属性を圧縮するために適用されるべき圧縮器タイプ及び／又は構成を選択することができるように、識別された圧縮タイプ及び／又は圧縮構成のうちの１つ又は複数についての情報を、ユーザインターフェースを介してユーザに与えることをさらに含む。

【００１４】

［0014］ 第２の態様は、データファイルを圧縮／展開するためのシステムである。本システムは、複数の異なる属性を含む、圧縮のためのデータファイルと、プロセッサとを含み、プロセッサは、（ｉ）複数の異なる属性のうちの第１の属性を識別し、識別された第１の属性の仕様を属性パラメータデータ構造に記憶することと、（ｉｉ）識別された第１の属性に基づいて、事前定義された複数の圧縮タイプ及び／又は圧縮設定から複数の圧縮タイプ及び／又は圧縮設定を選択することと、（ｉｉｉ）選択された複数の圧縮タイプ及び／又は圧縮設定の各々を使用して、識別された第１の属性のための受信されたデータファイルからのデータの少なくとも一部を個別に圧縮することであって、１つ又は複数のメトリクスが各個別の圧縮について測定される、圧縮することと、（ｉｖ）個別の圧縮の各々の１つ又は複数のメトリクスに基づいて、選択された複数の圧縮タイプ及び／又は圧縮設定のうちのいずれが圧縮のために最も好適であるかを決定することと、（ｖ）選択された複数の圧縮タイプ及び／又は圧縮設定のうちの、識別された第１の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮設定の識別を有する、圧縮パラメータデータ構造を生成することと、（ｖｉ）圧縮されたデータファイルを生成するために、識別された圧縮タイプを使用して、第１の属性のための受信されたデータファイルからのデータを圧縮することと、（ｖｉｉ）属性パラメータデータ構造と圧縮パラメータデータ構造と圧縮されたデータファイルとを記憶することであって、記憶された圧縮パラメータデータ構造が、圧縮されたデータファイルの展開のために取り出され、使用されるように構成された、記憶することとを行うように構成される。

【００１５】

［0015］ 一実施形態によれば、システムは、生成された圧縮パラメータデータ構造及び／又は圧縮されたデータファイルについての情報を与えるように構成されたユーザインターフェースを含む。

【００１６】

［0016］ 様々な実装形態では、プロセッサ又はコントローラは、（本明細書で一般的に「メモリ」と呼ぶ、例えば、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、及びＥＥＰＲＯＭ、フロッピーディスク、コンパクトディスク、光ディスク、磁気テープなど、揮発性及び不揮発性コンピュータメモリ）１つ又は複数の記憶媒体に関連付けられ得る。いくつかの実装形態では、記憶媒体は、１つ又は複数のプロセッサ及び／又はコントローラ上で実行されたときに、本明細書で説明した機能のうちの少なくともいくつかを実行する１つ又は複数のプログラムを用いて符号化され得る。本明細書で説明する様々な態様を実装するように、その上に記憶された１つ又は複数のプログラムをプロセッサ又はコントローラにロードすることができるように、様々な記憶媒体がプロセッサ又はコントローラ内に固定され得るか、又は移送可能であり得る。「プログラム」又は「コンピュータプログラム」という用語は、本明細書では、１つ又は複数のプロセッサ又はコントローラをプログラムするために採用することができる任意のタイプのコンピュータコード（例えば、ソフトウェア又はマイクロコード）を指すために、一般的な意味で使用する。

【００１７】

［0017］ 上記の概念、及び以下でより詳細に説明する追加の概念のすべての組合せは、（そのような概念が相互に矛盾しない限り）本明細書で開示する発明的主題の一部であるとして企図されることを理解されたい。特に、本開示の最後に現れるクレームされる主題のすべての組合せは、本明細書で開示する発明的主題の一部であるとして企図される。

また、参照により組み込まれるどの開示にも現れ得る本明細書で明示的に採用される用語は、本明細書で開示する特定の概念に最も一致する意味を与えられるべきであることを理解されたい。

【００１８】

[0018] 様々な実施形態のこれら及び他の態様は、以下で説明する実施形態から明らかになり、以下で説明する実施形態を参照すれば解明されるであろう。

【００１９】

[0019] 図面において、異なるビュー全体にわたって同様の参照符号は一般に同じ部分を指す。また、図面は必ずしも一定の縮尺でなく、むしろ、一般に、様々な実施形態の原理を示すことに強調が置かれる。

【図面の簡単な説明】

【００２０】

【図１】[0020] 一実施形態による、ファイル圧縮／展開システムを使用してゲノムデータファイルを圧縮するための方法のフローチャートである。

【図２】[0021] 一実施形態による、ファイル圧縮／展開システムの概略表現である。

【図３】[0022] 一実施形態による、データファイルの概略表現である。

【発明を実施するための形態】

【００２１】

[0023] 本開示では、データファイルの効率的で適応可能な圧縮のためのシステム及び方法の様々な実施形態について説明する。出願人は、圧縮性能に基づいて、データ圧縮器の自動化された識別及び／又は構成を採用するデータファイル圧縮方法及びシステムを提供することが有益であろうことを認識し、理解している。データファイル圧縮／展開システムは、圧縮のためのゲノムデータファイルを受信し、ゲノムデータファイルは、複数の異なるゲノムサンプルの各々についての複数の異なる属性（又はデータフィールド）を含む。本システムは、複数の異なる属性のうちの第１の属性を識別し、その属性に基づいて、事前定義された複数の圧縮タイプと圧縮構成とから複数の圧縮タイプ及び／又は圧縮構成を選択し、選択された複数の圧縮タイプ及び／又は圧縮構成の各々は、識別された第１の属性のための受信されたゲノムデータファイルからのデータを圧縮することが可能である。データファイル圧縮システムは、選択された複数の圧縮タイプ及び／又は圧縮構成の各々を使用して、識別された第１の属性のための受信されたゲノムデータファイルからのデータの少なくとも一部を個別に圧縮し、１つ又は複数のメトリクスが各個別の圧縮について測定され、１つ又は複数のメトリクスは前記圧縮のための圧縮比と処理時間とのうちの１つ又は複数を含む。個別の圧縮の各々の１つ又は複数のメトリクスに基づいて、本システムは、選択された複数の圧縮タイプ及び／又は圧縮構成のうちのいずれが圧縮のために最も好適であるかを決定する。一実施形態によれば、対応する圧縮が複数の選択された圧縮タイプ及び／又は圧縮構成のうちの他の圧縮タイプ及び／又は圧縮構成に対してより高い圧縮比及び／又はより速い処理時間を有するときの圧縮タイプが最も好適である。本システムは、次いで、選択された複数の圧縮タイプ及び／又は圧縮構成のうちの、識別された第１の属性の圧縮のために最も好適な圧縮タイプ及び／又は圧縮構成の識別、及び／又はそれらの圧縮タイプ及び／又は圧縮構成のための処理命令を含む圧縮パラメータデータ構造を生成し、圧縮されたゲノムデータファイルを生成するために、第１の属性のための受信されたゲノムデータファイルからのデータを圧縮する。圧縮パラメータデータ構造及び圧縮されたゲノムデータファイルは、記憶され、ここで、記憶された圧縮パラメータデータ構造は、圧縮されたゲノムデータファイルの展開のために取り出され、使用されるように構成される。

【００２２】

[0024] 一実施形態によれば、本明細書で説明又はさもなければ想定する新規の方法及びシステムは、圧縮器の性能に基づいて、データフィールドのための圧縮器の自動構成のためのフレームワークを与え、圧縮器を明示的に指定し、圧縮されたデータとともに記憶することを可能にする。属性データにアクセスするために、デコーダは、属性に関連付

けられた圧縮器構成において指定された処理ステップ及びパラメータに基づいて、展開及び逆変換アルゴリズムを実行することができる。本明細書で説明又はさもなければ想定する新規の方法及びシステムは多数の利点を与える。例えば、本方法及びシステムは、ファイルのための圧縮器構成をカスタマイズするためのフレキシビリティを与え、新しい変換及び圧縮アルゴリズムの採用を容易にする。第２に、本方法及びシステムは、圧縮器構成を識別し、与えることによって任意の新しいファイルタイプに適応する。処理命令は圧縮器構成に含まれており、ソフトウェアにおいてハードコーディングされていないので、デコーダは、コードの修正の必要なしにファイルを容易に処理することができる。第３に、圧縮器設定の自動構成は、ユーザ知識又は経験を必要とすることなしに圧縮性能を改善することを助ける。本方法及びシステムは多くの他の利点をも与える。

【００２３】

[0025] 一実施形態によれば、圧縮器構成フレームワークシステムは、以下の構成要素、すなわち、（ｉ）データ変換のシーケンスと、圧縮ステップと、それらのデフォルトパラメータとからなる圧縮プロセスなど、圧縮器を指定するためのユニフォームデータインターフェースと、（ｉｉ）データ属性を指定し、カスタマイズされたパラメータを用いてデータ属性を圧縮器に関連付けるためのユニフォームデータインターフェースと、（ｉｉｉ）関連付けられた圧縮器構成において指定された処理ステップに基づいて属性データを符号化するためのアルゴリズムと、（ｉｖ）関連付けられた圧縮器構成において指定された処理ステップに基づいて属性データを復号するためのアルゴリズムと、（ｖ）性能に基づく圧縮器構成の自動決定のための機構とのうちの１つ又は複数を含む。

【００２４】

[0026] 図１を参照すると、一実施形態は、ファイル圧縮／展開システムを使用してゲノムデータファイルを圧縮するための方法１００のフローチャートである。図に関して説明する方法は、単に例として与えられ、本開示の範囲を限定しないと理解されるべきである。ファイル圧縮／展開システムは、本明細書で説明又はさもなければ想定するシステムのいずれかであり得る。ファイル圧縮／展開システムは単一のシステム又は複数の異なるシステムであり得る。

【００２５】

[0027] 本方法のステップ１１０において、ファイル圧縮／展開システム２００が与えられる。図２に示されているファイル圧縮／展開システム２００の実施形態を参照すると、例えば、本システムは、１つ又は複数のシステムバス２１２を介して相互接続された、プロセッサ２２０と、メモリ２３０と、ユーザインターフェース２４０と、通信インターフェース２５０と、ストレージ２６０とのうちの１つ又は複数を含む。図２は、いくつかの点で、アブストラクションを構成していること、及びシステム２００の構成要素の実際の編成は異なり、示されているよりも複雑であり得ることが理解されよう。さらに、ファイル圧縮／展開システム２００は、本明細書で説明又はさもなければ想定するシステムのいずれかであり得る。ファイル圧縮／展開システム２００の他の要素及び構成要素は本明細書の他の場所で開示及び／又は想定される。

【００２６】

[0028] 一実施形態によれば、ファイル圧縮／展開システム２００は、ファイルデータベース２７０を含むか、又はファイルデータベース２７０と直接的又は間接的に通信している。ファイルデータベースは、可能なデータの中でも、ファイル圧縮／展開システム２００を使用して圧縮され得るか、又は圧縮された、１つ又は複数のファイルを含む。一実施形態によれば、１つ又は複数のファイルはゲノムファイルを含み、ゲノムファイルは、潜在的に多くの異なるゲノムサンプルのための、データ属性を含む何百万もの行と多数の列とをもつ遺伝子配列変異を含む。ファイルデータベースは、限定はしないが、本明細書で説明又はさもなければ想定するデータベース及びシステムを含む、何らかのそのようなデータベースであり得る。

【００２７】

[0029] 一実施形態によれば、ファイル圧縮／展開システム２００は、圧縮及び展開

アルゴリズムデータベース２８０を含むか、又は圧縮及び展開アルゴリズムデータベース２８０と直接的又は間接的に通信している。圧縮及び展開アルゴリズムデータベースは、異なる圧縮アルゴリズムと展開アルゴリズムとのライブラリを含む。これらの圧縮及び展開アルゴリズムは、データファイルシステムによってデータファイルを圧縮及び展開するために利用され得る。データベースに新しい圧縮アルゴリズム及び展開アルゴリズムをいつでも追加することができる。圧縮及び展開アルゴリズムデータベースは、限定はしないが、本明細書で説明又はさもなければ想定するデータベース及びシステムを含む、何らかのそのようなデータベースであり得る。

【００２８】

［0030］ 本方法のステップ１２０において、ファイル圧縮／展開システムは、圧縮のための圧縮されていないデータファイルを受信し、データファイルは、複数の異なるサンプルの各々についての複数の異なる属性を含む。例として、ゲノムデータファイルの場合、ゲノムデータファイルは、複数の異なるゲノムサンプルの各々についての複数の異なる属性を含む。データファイルは、限定はしないが、それ自体がローカル又はリモートデータベースであり得る、ファイルデータベース２７０を含む、任意のソースからローカル又はリモートで受信され得る。

【００２９】

［0031］ 方法のステップ１３０において、ファイルデータシステムは、データファイル中の複数の異なる属性のうちの第１の属性を識別する。属性はデータファイルからの任意の情報であり得る。図３を参照すると、データファイルの非限定的な一例では、各列が属性である、複数の列３１０を含むゲノムデータファイルの部分３００である。ファイルデータシステムは、データファイル内の知られている属性を認識するように構成され得るか、又は、データファイル内の新しい又は認識されていない属性を予測又は推定又はさもなければ決定するように構成され得る。したがって、ファイルデータシステムは、多種多様な異なる方法を使用してデータファイル中の複数の異なる属性のうちの第１の属性を識別するように構成され得る。一実施形態によれば、属性を識別した後に、本システムは、（本明細書で説明する属性パラメータセットなど）属性パラメータデータ構造において属性を指定する。

【００３０】

［0032］ ステップ１４０において、データファイル圧縮／展開システムは、事前定義された複数の圧縮タイプとアルゴリズムと構成とから複数の圧縮タイプ又はアルゴリズム（「圧縮器」）及び／又は圧縮構成を選択する。圧縮器又は圧縮構成は、例えば、圧縮アルゴリズムのシーケンスと、圧縮アルゴリズムに関連付けられたパラメータとを含み得る。圧縮タイプ、アルゴリズム、及び圧縮構成は、限定はしないが、それ自体がローカル又はリモートデータベースであり得る、圧縮及び展開アルゴリズムデータベース２８０を含む、任意のソースからローカル又はリモートで選択され得る。一実施形態によれば、複数の圧縮タイプとアルゴリズムと圧縮構成とのうちの１つ又は複数は、識別された属性に少なくとも部分的に基づいて、事前定義された複数の圧縮タイプとアルゴリズムと圧縮構成とから選択される。例えば、圧縮器又は構成は、識別された第１の属性のための受信されたデータファイルからのデータを圧縮する圧縮器の能力に基づいて選択され得る。したがって、本システムは、第１の属性データを圧縮することが可能な圧縮器又は構成を優先的に識別し得、及び／又は第１の属性データを圧縮することが不可能な圧縮器を優先的に除外し得る。

【００３１】

［0033］ ステップ１５０において、ファイルデータシステムは、識別された圧縮器を使用してデータを圧縮する。一実施形態によれば、ファイル圧縮／展開システムは第１の属性データを個別に圧縮する。複数の圧縮器によるこの個別の圧縮は連続的に又は同時に実行され得る。一実施形態によれば、各圧縮器は第１の属性データのサブセットのみを圧縮し、サブセットは、複数の圧縮器の中での圧縮メトリクスのより客観的な比較を可能にするために各圧縮器について同等のサブセットである。別の実施形態によれば、各圧縮器

はすべての第１の属性データを圧縮する。
【００３２】
［0034］ 実施形態によれば、ファイル圧縮／展開システムは、各圧縮器による各個別の圧縮についての１つ又は複数の圧縮メトリクスを測定又はさもなければ取得又は識別する。これにより、複数の圧縮器の中での圧縮性能の客観的な比較と、第１の属性データのための好ましい圧縮器の選択とが可能になる。一実施形態によれば、圧縮メトリクスは、圧縮比、処理時間、圧縮精度、及び／又は任意の他の圧縮メトリックであり得る。
【００３３】
［0035］ 方法のステップ１６０において、ファイル圧縮／展開システムは、測定された又は取得されたメトリクスを使用して、選択された複数の圧縮タイプ、アルゴリズム、又は圧縮構成のいずれが圧縮のために最も好適であるかを決定する。したがって、ファイル圧縮／展開システムは、各個別の圧縮器に対応するメトリクスを比較、ランク付け、又はさもなければ利用し、好ましい圧縮器を選択する。例えば、圧縮又は構成が、複数の選択された圧縮タイプ又は構成のうちの他の圧縮タイプ及び／又は圧縮構成に対して、より高い又は最良の圧縮比及び／又はより速い処理時間を有するとき、圧縮器又は構成は、圧縮のために最も好適に又は好ましくなり得る。第１の実施形態によれば、ファイル圧縮／展開システムは、最良の圧縮比をもつ圧縮器を選択するように構成される。第２の実施形態によれば、ファイル圧縮／展開システムは、最も速い圧縮又は処理時間をもつ圧縮器を選択するように構成される。第３の実施形態によれば、ファイル圧縮／展開システムは、圧縮比と圧縮又は処理時間との最良の組合せをもつ圧縮器を選択するように構成される。これらの構成は、オプションの中でも、事前プログラムされ得るか、学習され得るか、又はシステムのユーザによって決定され得る。
【００３４】
［0036］ 一実施形態によれば、ファイル圧縮／展開システムは、１つ又は複数の追加の属性のために、識別するステップ、選択するステップ、圧縮するステップ、及び、決定するステップを繰り返す。例えば、本システムは、データファイル内で見つけられた一部の又は全部の属性のための好ましい圧縮器を識別するように構成され得る。
【００３５】
［0037］ 本方法のステップ１７０において、ファイル圧縮／展開システムは、圧縮パラメータデータファイルなど、圧縮パラメータデータ構造を生成する。一実施形態によれば、生成された圧縮パラメータデータ構造は、少なくとも第１の識別された属性のための選択された圧縮器及び／又は圧縮構成の識別、及び／又はそのような圧縮器及び／又は圧縮構成のための仕様、及び、随意に、データファイル中の複数の属性のために選択された圧縮器及び／又は圧縮構成の識別（例えば、複数の属性のための圧縮器の使用を指定すること）を含む。圧縮パラメータデータファイルは、圧縮器設定及び他の情報を含む、圧縮器を利用するために必要な任意の情報を含むことができる。
【００３６】
［0038］ 方法のステップ１８０において、ファイル圧縮／展開システムは、識別された圧縮器及び／又は圧縮構成を使用して、圧縮されたファイルを生成するために、受信されたデータファイルからのデータを圧縮する。したがって、これは、第１の属性データか、複数の属性についてのデータか、又はファイル中のすべての属性についてのデータの圧縮を含むことができる。システムが圧縮されたファイルを生成すると、圧縮されたファイルを送る又はさもなければ送信する又は使用するなど、ファイルはすぐに利用され得るか、又は圧縮されたファイルは、本方法のさらなるステップにおいて使用するためにローカル又はリモートストレージに記憶され得る。多くの他のオプションが可能である。
【００３７】
［0039］ 方法のステップ１９０において、生成された圧縮パラメータデータ構造は、記憶されたファイルを圧縮されたデータファイルの展開のために取り出し、使用することができるように記憶される。圧縮パラメータデータ構造は、ファイル圧縮／展開システム２００の構成要素であるか、又はさもなければファイル圧縮／展開システム２００と通信

し得る、リモート又はローカルデータベースに記憶され得る。例えば、圧縮パラメータデータ構造は、圧縮されたデータファイルと一緒に記憶されるか、又は圧縮されたデータファイルへの又は圧縮されたデータファイルからの参照とともに記憶され得る。

【００３８】

［0040］ 図２を参照すると、一実施形態では、ファイル圧縮／展開システム２００の概略表現である。システム２００は、本明細書で説明又はさもなければ想定するシステムのいずれかであり得、本明細書で説明又はさもなければ想定する構成要素のいずれかを備え得る。図２は、いくつかの点で、アブストラクションを構成していること、及びシステム２００の構成要素の実際の編成は異なり、示されているよりも複雑であり得ることが理解されよう。

【００３９】

［0041］ 一実施形態によれば、システム２００は、メモリ２３０又はストレージ２６０に記憶された命令を実行すること、又はさもなければ、例えば、方法の１つ又は複数のステップを実行するためにデータを処理することが可能な、プロセッサ２２０を含む。プロセッサ２２０は１つ又は複数のモジュールから形成され得る。プロセッサ２２０は、限定はしないが、マイクロプロセッサ、マイクロコントローラ、複数のマイクロコントローラ、回路、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、特定用途向け集積回路（ＡＳＩＣ）、シングルプロセッサ、又は複数のプロセッサを含む、任意の好適な形態を取り得る。

【００４０】

［0042］ メモリ２３０は、不揮発性メモリ及び／又はＲＡＭを含む、任意の好適な形態を取ることができる。メモリ２３０は、例えばＬ１、Ｌ２、又はＬ３キャッシュ又はシステムメモリなど、様々なメモリを含み得る。したがって、メモリ２３０は、スタティックランダムアクセスメモリ（ＳＲＡＭ）、ダイナミックＲＡＭ（ＤＲＡＭ）、フラッシュメモリ、読取り専用メモリ（ＲＯＭ）、又は他の同様のメモリデバイスを含み得る。メモリは、とりわけ、オペレーティングシステムを記憶することができる。ＲＡＭはデータの一時的記憶のためにプロセッサによって使用される。一実施形態によれば、オペレーティングシステムは、プロセッサによって実行されたときに、システム２００の１つ又は複数の構成要素の動作を制御するコードを含み得る。プロセッサが本明細書で説明する機能のうちの１つ又は複数をハードウェアにおいて実装する実施形態では、他の実施形態におけるそのような機能に対応するとして説明したソフトウェアは省略され得ることは明らかであろう。

【００４１】

［0043］ ユーザインターフェース２４０は、ユーザとの通信を可能にするための１つ又は複数のデバイスを含み得る。ユーザインターフェースは、情報を伝達及び／又は受信することを可能にする任意のデバイス又はシステムであり得、ユーザコマンドを受信するためのディスプレイ、マウス、及び／又はキーボードを含み得る。いくつかの実施形態では、ユーザインターフェース２４０は、通信インターフェース２５０を介してリモート端末に提示され得るコマンドラインインターフェース又はグラフィカルユーザインターフェースを含み得る。ユーザインターフェースは、システムの１つ又は複数の他の構成要素とともに位置し得るか、又は、システムからリモートに位置し、ワイヤード及び／又はワイヤレス通信ネットワークを介して通信し得る。

【００４２】

［0044］ 通信インターフェース２５０は、他のハードウェアデバイスとの通信を可能にするための１つ又は複数のデバイスを含み得る。例えば、通信インターフェース２５０は、イーサネットプロトコルに従って通信するように構成されたネットワークインターフェースカード（ＮＩＣ）を含み得る。さらに、通信インターフェース２５０はＴＣＰ／ＩＰプロトコルに従った通信のためのＴＣＰ／ＩＰスタックを実装し得る。通信インターフェース２５０のための様々な代替又は追加のハードウェア又は構成が明らかになろう。

【００４３】

［0045］ストレージ２６０は、読取り専用メモリ（ＲＯＭ）、ランダムアクセスメモリ（ＲＡＭ）、磁気ディスク記憶媒体、光記憶媒体、フラッシュメモリデバイス、又は同様の記憶媒体など、１つ又は複数の機械可読記憶媒体を含み得る。様々な実施形態では、ストレージ２６０は、プロセッサ２２０による実行のための命令、又はプロセッサ２２０がそれに対して動作し得るデータを記憶し得る。例えば、ストレージ２６０は、システム２００の様々な動作を制御するためのオペレーティングシステム２６１を記憶し得る。

【００４４】

［0046］ストレージ２６０に記憶されるとして説明される様々な情報が追加又は代替としてメモリ２３０に記憶され得ることが明らかになろう。この点で、メモリ２３０は記憶デバイスを構成するとも考えられ得、ストレージ２６０はメモリと考えられ得る。様々な他の配置が明らかになろう。さらに、メモリ２３０とストレージ２６０は両方とも非一時的機械可読媒体であると考えられ得る。本明細書で使用する際、非一時的という用語は、一時的信号を除外するが、揮発性メモリと不揮発性メモリの両方を含む、すべての形態のストレージを含むことが理解されよう。

【００４５】

［0047］システム２００は、各説明された構成要素のうちの１つを含むとして示されているが、様々な構成要素は様々な実施形態において重複し得る。例えば、プロセッサ２２０は、本明細書で説明する方法を単独で実行するように構成された、又は本明細書で説明する機能を達成するために複数のプロセッサが協働するように、本明細書で説明する方法のステップ又はサブルーチンを実行するように構成された、複数のマイクロプロセッサを含み得る。さらに、システム２００の１つ又は複数の構成要素がクラウドコンピューティングシステムにおいて実装される場合、様々なハードウェア構成要素は別個の物理的システムに属し得る。例えば、プロセッサ２２０は、第１のサーバ中の第１のプロセッサと、第２のサーバ中の第２のプロセッサとを含み得る。多くの他の変形形態及び構成が可能である。

【００４６】

［0048］一実施形態によれば、ファイル圧縮／展開システム２００は、ファイルデータベース２７０を含むか、又はファイルデータベース２７０と直接的又は間接的に通信している。ファイルデータベースは、可能なデータの中でも、ファイル圧縮／展開システム２００を使用して圧縮され得るか、又は圧縮された、１つ又は複数のファイルを含む。一実施形態によれば、１つ又は複数のファイルは、ゲノムファイルを含み、ゲノムファイルは、潜在的に多くの異なるゲノムサンプルのための、データ属性を含む、何百万もの行と多数の列とをもつ遺伝子配列変異を含む。

【００４７】

［0049］一実施形態によれば、ファイル圧縮／展開システム２００は、圧縮及び展開アルゴリズムデータベース２８０を含むか、又は圧縮及び展開アルゴリズムデータベース２８０と直接的又は間接的に通信している。圧縮及び展開アルゴリズムデータベースは、異なる圧縮アルゴリズムと展開アルゴリズムとのライブラリを含む。これらの圧縮アルゴリズム及び展開アルゴリズムは、データファイルシステムによってデータファイルを圧縮及び展開するために利用され得る。新しい圧縮アルゴリズム及び展開アルゴリズムをいつでもデータベースに追加することができる。

【００４８】

［0050］一実施形態によれば、システム２００のストレージ２６０は、本明細書で説明又はさもなければ想定する方法の１つ又は複数の機能又はステップを実行するための１つ又は複数のアルゴリズム、モジュール、及び／又は命令を記憶し得る。例えば、ストレージ２６０は、他の命令又はデータの中でも、属性識別命令２６２、圧縮器フォーミュレーション命令２６３、圧縮器選択命令２６４、記憶命令２６５、及び／又は報告命令２６６を備え得る。

【００４９】

［0051］一実施形態によれば、属性識別命令２６２は、システムによって受信される

データファイル中の複数の異なる属性のうちの１つの属性を識別するようにシステムに指令する。属性はデータファイルからの任意の情報であり得る。ファイルデータシステムは、データファイル内の知られている属性を認識するように構成され得るか、又はデータファイル内の新しい又は認識されていない属性を予測又は推定又はさもなければ決定するように構成され得る。したがって、ファイルデータシステムは、多種多様な異なる方法を使用してデータファイル中の複数の異なる属性のうちの第１の属性を識別するように構成され得る。

【００５０】

［0052］ 一実施形態によれば、圧縮器構築命令２６３は、事前定義された複数の圧縮タイプ、アルゴリズム、及び／又は圧縮構成から複数の圧縮器を選択するようにシステムに指令する。圧縮タイプ、アルゴリズム、及び／又は圧縮構成は、限定はしないが、圧縮及び展開アルゴリズムデータベース２８０を含む任意のソースから、ローカル又はリモートで選択され得る。圧縮器構築命令２６３はまた、識別された圧縮器及び／又は圧縮構成を使用して属性データの一部又は全部を個別に圧縮するようにシステムに指令する。複数の圧縮器によるこの個別の圧縮は連続的に又は同時に実行され得る。一実施形態によれば、各圧縮器は第１の属性データのサブセットのみを圧縮し、サブセットは、複数の圧縮器の中での圧縮メトリクスのより客観的な比較を可能にするために各圧縮器について同等のサブセットである。別の実施形態によれば、各圧縮器はすべての第１の属性データを圧縮する。

【００５１】

［0053］ 実施形態によれば、圧縮器構築命令２６３はまた、各圧縮器による各個別の圧縮について１つ又は複数の圧縮メトリクスを測定又はさもなければ取得又は識別するようにシステムに指令する。これにより、複数の圧縮器及び／又は圧縮構成の中での圧縮性能の客観的な比較と、第１の属性データのための好ましい圧縮器及び／又は圧縮構成の選択とが可能になる。一実施形態によれば、圧縮メトリクスは、圧縮比、処理時間、圧縮精度、及び／又は任意の他の圧縮メトリックであり得る。

【００５２】

［0054］ 一実施形態によれば、圧縮器選択命令２６４は、測定又は取得されたメトリクスを使用して、選択された複数の圧縮タイプ、アルゴリズム、及び／又は圧縮構成のいずれが圧縮のために最も好適であるかを決定するようにシステムに指令する。ファイル圧縮／展開システムは、各個別の圧縮器及び／又は圧縮構成に対応するメトリクスを比較、ランク付け、又はさもなければ利用することができ、好ましい圧縮器及び／又は圧縮構成を選択する。例えば、圧縮が複数の選択された圧縮タイプのうちの他の圧縮タイプに対してより高い又は最良の圧縮比及び／又はより速い処理時間を含むとき、圧縮器及び関連する構成は圧縮のために最も好適又は好ましいものであり得る。一実施形態によれば、属性に適用されるべき圧縮器のマニュアル選択を促進するために、異なる圧縮器構成及び対応するメトリクスを、ユーザインターフェースを介してユーザに提示することができる。

【００５３】

［0055］ 一実施形態によれば、記憶命令２６５は、圧縮パラメータデータファイルなど、圧縮パラメータデータ構造を生成するようにシステムに指令する。一実施形態によれば、生成された圧縮パラメータデータ構造は、少なくとも第１の識別された属性のための選択された圧縮器及び／又は圧縮構成の識別、及び／又はそのような圧縮器及び／又は圧縮構成のための仕様を含む。圧縮パラメータデータファイルは、圧縮器設定及び他の情報を含む、圧縮器を利用するために必要な任意の情報を含むことができる。記憶命令２６５はまた、記憶されたファイルを圧縮されたデータファイルの展開のために取り出し、使用することができるように、生成された圧縮パラメータデータ構造を記憶するようにシステムに指令することができる。圧縮パラメータデータ構造は、ファイル圧縮／展開システム２００の構成要素であるか、又はさもなければファイル圧縮／展開システム２００と通信し得るし、リモート又はローカルデータベースに記憶され得る。例えば、圧縮パラメータデータ構造は、圧縮されたデータファイルと一緒に記憶され得るか、又は圧縮されたデー

タファイルへの又は圧縮されたデータファイルからの参照とともに記憶される。
【００５４】
［0056］ 一実施形態によれば、報告命令２６６は、システムのユーザインターフェース２４０を介して、圧縮、データファイル、及び／又は圧縮器についての情報を与えるようにシステムに指令する。与えられる情報は、本明細書で説明又はさもなければ想定する何らかの情報であり得る。本システムは、限定はしないが、視覚表示、可聴通知、テキストメッセージ、電子メール、ページ、又は任意の他の通知方法を含む、任意の機構を介してその情報をユーザに与え得る。
【００５５】
［0057］ 圧縮器を指定するためのユニフォームインターフェース
［0058］ 一実施形態によれば、ファイル圧縮／展開システムは圧縮器パラメータセットを含む。圧縮器パラメータセットは、属性データの符号化と復号とのために必要とされるすべての命令とデフォルトパラメータとを含む、圧縮器構成を含んでいるデータ構造であり、圧縮されたデータとともに記憶されるべきである。各圧縮器構成は、属性との各圧縮器構成の関連を指定するために使用されるべき一意のＩＤと、それぞれ以下の要素のうちの１つ又は複数からなる圧縮器ステップのシーケンスとを搬送する。
【００５６】
［0059］ （ｉ）圧縮器ステップの一意のＩＤ。
【００５７】
［0060］ （ｉｉ）ステップにおいて適用されるべきアルゴリズムのタイプ及びＩＤ。一実施形態によれば、データ表現のフォーマットを変更することと、確率モデリングによって固定長シンボルを圧縮されたビットストリームに変換することとを伴う、２つの広義のタイプのアルゴリズム、すなわち、変換と圧縮とがある。アルゴリズムＩＤは、指定されたタイプの知られているアルゴリズムを参照する事前定義されたコードである。データを処理するために、参照されているすべてのアルゴリズムがインストールされ、エンコーダ又はデコーダにアクセス可能にされなければならない。
【００５８】
［0061］ （ｉｉｉ）アルゴリズムへのデフォルトパラメータのリスト。各パラメータは、パラメータ名と、データタイプと、パラメータ中の値の数と、パラメータ値とから構成される。
【００５９】
［0062］ （ｉｖ）アルゴリズムへの入力変数の数及び入力変数名。変数名は、１つの圧縮器ステップから別の圧縮器ステップに渡されるべきデータオブジェクトを参照するためにのみ使用されることに留意されたい。アルゴリズムに対して、それは、認識された、入力変数名の順序によって指定された、入力データオブジェクトの順序である。第１のステップの場合、入力データは未加工属性値でなければならないので、入力変数名を指定する必要がなく、入力変数の数を０に設定することができる。第２のステップ以降、各入力変数名は前のステップにおける出力変数名と一致しなければならない。
【００６０】
［0063］ （ｖ）アルゴリズムによって生成されたデータオブジェクトの順序付けされたリストを保持するための出力変数の数、並びに各出力変数の状態及び名前。出力変数がさらなる処理を必要としない場合、それの状態は完了であり、データオブジェクトは圧縮器の出力バッファに直接進むべきであるので、それの変数名を省略することができる。出力変数が後続のステップによって処理される必要がある場合、変数状態は不完了であり、変数名を指定しなければならない。
【００６１】
［0064］ （ｖｉ）アルゴリズムを実行するためにそれの値が必要とされる依存性属性の数。依存性属性のＩＤは、関連付けられた属性の構成中に与えられるべきである。
【００６２】
［0065］ 表１を参照すると、上記で説明した圧縮器構成の重要な要素をキャプチャす

る、圧縮器パラメータセットの例示的なシンタックスであり、表の後にはセマンティクスについての説明が与えられている。個々のフィールドの名前、順序付け及びデータタイプは機能への影響なしに変更することができることに留意されたい。データタイプの表記法に関して、ｕ（ｎ）は、ｎビットを使用する符号なし整数を示し、ｕ（ｖ）は、他のシンタックス要素の値に応じたビットの数をもつ符号なし整数を示し、ｆ（ｎ）は、ｎビットを使用する固定パターンビットストリングを示し、ｓｔ（ｖ）は、変数長をもつヌル終端されたストリングを示す。

【００６３】

［0066］ 表１．圧縮器パラメータセットシンタックス

【表１】

シンタックス	タイプ
compressor_parameter_set {	
compressor_ID	u(8)
n_compressor_steps	u(4)
for (i=0; i<n_compressor_steps; i++) {	
compressor_step_ID[i]	u(4)
algorithm_ID[i]	u(5)
use_default_pars[i]	u(1)
if (!use_default_pars[i]){	
algorithm_parameters(algorithm_ID)	
}	
n_in_vars[i]	u(4)
for (j=0; j<n_in_vars[i]; j++){	
in_var_ID[i][j]	u(4)
prev_step_ID[i][j]	u(4)
prev_out_var_ID[i][j]	u(4)
}	
n_completed_out_vars[i]	u(4)
for (j=0; j<n_completed_out_vars[i]; j++) {	
completed_out_var_ID[i][j]	u(4)
}	u(4)
}	
while (!byte_aligned()) {	
nesting_zero_bit	f(1)
}	
}	

【００６４】
［0067］一実施形態によれば、以下の圧縮器パラメータセットセマンティクスである。
【００６５】
［0068］ｃｏｍｐｒｅｓｓｏｒ＿ＩＤは注釈テーブル内の圧縮器の一意識別子である。値０は圧縮なしのために確保されることに留意されたい。
【００６６】
［0069］ｎ＿ｃｏｍｐｒｅｓｓｏｒ＿ｓｔｅｐｓはこの圧縮器構成における処理ステップの総数である。各圧縮器ステップｉについて、以下のフィールドが指定される。（ｉ）ｃｏｍｐｒｅｓｓｏｒ＿ｓｔｅｐ＿ＩＤ［ｉ］は圧縮器ステップの識別子である。（ｉｉ）ａｌｇｏｒｉｔｈｍ＿ＩＤ［ｉ］は、このステップにおいて適用されるべきアルゴリズムの識別子である。各アルゴリズムタイプはそれ自体のＩＤマップを有するべきである。（ｉｉｉ）ｕｓｅ＿ｄｅｆａｕｌｔ＿ｐａｒｓ［ｉ］はフラグであり、１に設定された場合、アルゴリズムのデフォルトパラメータが使用されることを示し、したがってパラメータに対する変更は不要である。
【００６７】
［0070］ｎ＿ｉｎ＿ｖａｒｓ［ｉ］はアルゴリズムのための入力変数の数である。第１のステップ（ｉ＝＝０）の場合、入力は未加工属性値でなければならないので、それの値を０に設定するべきである。
【００６８】
［0071］アルゴリズムのための各入力変数ｊについて、以下のフィールドが指定される。（ｉ）ｉｎ＿ｖａｒ＿ＩＤ［ｉ］［ｊ］は、アルゴリズムに知られている入力変数の事前定義されたＩＤであり、（ｉｉ）ｐｒｅｖ＿ｓｔｅｐ＿ＩＤ［ｉ］［ｊ］、ｐｒｅｖ＿ｏｕｔ＿ｖａｒ＿ＩＤ［ｉ］［ｊ］は、それぞれ、前の圧縮器ステップのＩＤ、及びｉｎ＿ｖａｒ＿ＩＤ［ｉ］［ｊ］によって指定された入力変数を通して現在のアルゴリズムに渡されるべきデータを含んでいる、対応する出力変数のＩＤである。
【００６９】
［0072］ｎ＿ｃｏｍｐｌｅｔｅｄ＿ｏｕｔ＿ｖａｒｓ［ｉ］は、さらなる処理を必要とせず、圧縮器の出力バッファに直接進むことができる、アルゴリズムの完了した出力変数の数である。
【００７０】
［0073］ｃｏｍｐｌｅｔｅｄ＿ｏｕｔ＿ｖａｒ＿ＩＤ［ｉ］［ｊ］は、完了であり、さらなる処理を必要としない、アルゴリズムに知られている出力変数の事前定義されたＩＤを指定する。
【００７１】
［0074］ｎｅｓｔｉｎｇ＿ｚｅｒｏ＿ｂｉｔは、データ構造のバイトアラインメントを保証するために必要に応じて挿入される、０に設定された１ビットである。
【００７２】
［0075］一実施形態によれば、アルゴリズムパラメータは、ａｌｇｏｒｉｔｈｍ＿ＩＤによって参照されるアルゴリズムのパラメータ設定を指定するためのデータ構造である。表２を参照すると、アルゴリズムパラメータのための例示的なシンタックスの表である。
【００７３】
［0076］表２．アルゴリズムパラメータシンタックス

【表２】

シンタックス	タイプ	注
algorithm_parameters(algorithm_ID) {		
n_pars	u(4)	
for (i = 0; i < n_pars; i++) {		
par_ID[i]	u(4)	
par_type[i]	u(8)	
par_num_array_dims[i]	u(2)	
for (j = 0; j < par_num_array_dims; j++) {		
par_array_dims[i][j]	u(8)	
}		
if (par_num_array_dims[i] == 0) {		
par_val[i]	u(v)	パラメータデータタイプに応じたv
}else if (par_num_array_dims[i] == 1) {		
for (j = 0; j < par_array_dims[i][0]; j++) {		
par_val[i][j]	u(v)	パラメータデータタイプに応じたv
}		
} else if (par_num_array_dims[i] == 2) {		
for (j = 0; j < par_array_dims[i][0]; j++) {		
for (k = 0; k < par_array_dims[i][1]; k++) {		
par_val[i][j][k]	u(v)	パラメータデータタイプに応じたv
}		
}		
} else if (par_num_array_dims[i] == 3) {		
for (j = 0; j < par_array_dims[i][0]; j++) {		
for (k = 0; k < par_array_dims[i][1]; k++) {		
for (k = 0; k < par_array_dims[i][2]; k++) {		
par_val[i][j][k][l]	u(v)	パラメータデータタイプに応じたv
}		
}		
}		
}		
}		
}		

【００７４】

［0077］一実施形態によれば、以下のアルゴリズムパラメータセマンティクスである。

【００７５】

［0078］ｎ＿ｐａｒｓは、ａｌｇｏｒｉｔｈｍ＿ＩＤによって参照されるアルゴリズムを実行するために修正される必要があるパラメータの数である。それの値は、アルゴリズムにおいて定義されたパラメータの数よりも大きくなってはならない。

【００７６】

［0079］修正されている各パラメータｉについて、以下のフィールドが必要とされる。（ｉ）ｐａｒ＿ＩＤ［ｉ］は、アルゴリズムにおいて定義されるパラメータのうちの１

つのＩＤである、（ｉｉ）ｐａｒ＿ｔｙｐｅ［ｉ］はパラメータのデータタイプＩＤである、（ｉｉｉ）ｐａｒ＿ｎｕｍ＿ａｒｒａｙ＿ｄｉｍｓ［ｉ］はパラメータの次元の数であり、それがスカラー値である場合、０である、（ｉｖ）ｐａｒ＿ａｒｒａｙ＿ｄｉｍｓ［ｉ］［］は、それぞれ配列次元のサイズを指定する要素を含んでいる。それは、ａｔｔｒｉｂｕｔｅ＿ｎｕｍ＿ａｒｒａｙ＿ｄｉｍｓ＝０である場合に省略される。（ｖ）ｐａｒ＿ｖａｌ［ｉ］［］［］［］はパラメータ値を含んでいる。それの次元数は（ｐａｒ＿ａｒｒａｙ＿ｄｉｍｓ＋１）に等しく、余分な次元を省略することができる。

【００７７】

［0080］ データ属性と関連する圧縮器構成とを指定するためのユニフォームインターフェース

［0081］ 一実施形態によれば、ファイル圧縮／展開システムは属性パラメータセットを含む。属性パラメータセットは、何らかの基本情報とそれの処理のための圧縮器構成とを含む、属性の構成を含んでいるデータ構造である。以下は重要な要素である。（ｉ）属性の一意のＩＤ及び名前、（ｉｉ）データタイプ情報、並びに属性のデフォルト値及び欠損値、（ｉｉｉ）属性の処理のための圧縮器のＩＤ、それは、圧縮器パラメータセットにおいて定義された圧縮器ＩＤのうちの１つでなければならない。（ｉｖ）圧縮器ステップの選択されたパラメータへの修正、及び（ｖ）それのデータがいくつかの圧縮器ステップによる処理のために必要とされる、依存性変数（属性又はデスクリプタ）のＩＤ。

【００７８】

［0082］ 表３を参照すると、上記で説明した属性構成の重要な要素をキャプチャする、属性パラメータセットの例示的シンタックスであり、表の後にはセマンティクスについての説明が与えられている。個々のフィールドの名前、順序付け及びデータタイプは機能への影響なしに変更することができることに留意されたい。データタイプ表記法は表１と同じである。属性パラメータセットは、それらのパラメータと、グループ化と、順序付けと、関連する圧縮器とを含む、追加の属性の定義を含んでいる。

【００７９】

［0083］ 表３．属性パラメータセットシンタックス

【表３】

シンタックス	タイプ
attribute_parameter_set {	
attribute_ID	u(16)
attribute_name_len	u(8)
attribute_name	c(attribute_name_len)
attribute_type	u(8)
attribute_num_array_dims	u(2)
for (i = 0; i < attribute_num_array_dims; i++) {	
attribute_array_dims[i]	u(8)
}	
attribute_default_value	st(v)
attribute_missing_value	st(v)
compressor_ID	u(8)
n_steps_with_dependencies	u(4)
for (i = 0; i < n_steps_with_dependencies; i++) {	
compressor_step_ID[i]	u(4)
n_dependencies[i]	u(4)
for (j = 0; j < n_dependencies[i]; j++) {	
dependency_var_ID[i][j]	u(4)
dependency_is_attribute[i][j]	u(1)
if (dependency_is_attribute[i][j]) {	
dependency_ID[i][j]	u(16)
} else {	
dependency_ID[i][j]	u(7)
}	
}	
}	
while(!byte_aligned()) {	
nesting_zero_bit	f(1)
}	
}	

【００８０】

［0084］一実施形態によれば、以下の属性パラメータセットセマンティクスである。

【００８１】

［0085］ａｔｔｒｉｂｕｔｅ＿ＩＤは、注釈パラメータセット内で一意の、属性の識別子である。

【００８２】

［0086］ａｔｔｒｉｂｕｔｅ＿ｎａｍｅ＿ｌｅｎは属性名の長さである。

【００８３】

［0087］ａｔｔｒｉｂｕｔｅ＿ｎａｍｅは、長さａｔｔｒｉｂｕｔｅ＿ｎａｍｅ＿ｌ

ｅｎをもつ属性の名前である。
【００８４】
［0088］ ａｔｔｒｉｂｕｔｅ＿ｔｙｐｅは属性のデータタイプを指定する。
【００８５】
［0089］ ａｔｔｒｉｂｕｔｅ＿ｎｕｍ＿ａｒｒａｙ＿ｄｉｍｓは属性エントリの配列次元の数を指定し、それがスカラー値である場合、０である。
【００８６】
［0090］ ａｔｔｒｉｂｕｔｅ＿ａｒｒａｙ＿ｄｉｍｓ［］は、それぞれ配列次元のサイズを指定する要素を含んでいる。それは、ａｔｔｒｉｂｕｔｅ＿ｎｕｍ＿ａｒｒａｙ＿ｄｉｍｓ＝０である場合に省略される。
【００８７】
［0091］ ａｔｔｒｉｂｕｔｅ＿ｄｅｆａｕｌｔ＿ｖａｌｕｅは属性のデフォルト値である（すなわち、デフォルトに等しい大部分の値が除外されるとき、それをスパース符号化のために使用することができる）。
【００８８】
［0092］ ａｔｔｒｉｂｕｔｅ＿ｍｉｓｓｉｎｇ＿ｖａｌｕｅは、復号プロセスの後に出力においてヌル値の代わりに使用されるべき属性の欠損値である。
【００８９】
［0093］ ｃｏｍｐｒｅｓｓｏｒ＿ＩＤは、現在の属性の処理のために使用されたｃｏｍｐｒｅｓｓｏｒ＿ｐａｒａｍｅｔｅｒ＿ｓｅｔにおいて定義されるｃｏｍｐｒｅｓｓｏｒ＿ＩＤである。
【００９０】
［0094］ ｎ＿ｓｔｅｐｓ＿ｗｉｔｈ＿ｄｅｐｅｎｄｅｎｃｉｅｓは、依存性変数を伴うｃｏｍｐｒｅｓｓｏｒ＿ＩＤによって参照される、圧縮器におけるステップの数である。
【００９１】
［0095］ 指定される必要がある依存性変数をもつ各圧縮器ステップｉについて、（ｉ）ｃｏｍｐｒｅｓｓｏｒ＿ｓｔｅｐ＿ＩＤ［ｉ］は圧縮器ステップのＩＤであり、（ｉｉ）ｎ＿ｄｅｐｅｎｄｅｎｃｉｅｓ［ｉ］は、圧縮器ステップのアルゴリズムによって必要とされる依存性変数の数である。
【００９２】
［0096］ アルゴリズムの各依存性変数ｊについて、（ｉ）ｄｅｐｅｎｄｅｎｃｙ＿ｖａｒ＿ＩＤ［ｉ］［ｊ］は、アルゴリズムに知られている依存性変数の事前定義されたＩＤであり、（ｉｉ）ｄｅｐｅｎｄｅｎｃｙ＿ｉｓ＿ａｔｔｒｉｂｕｔｅ［ｉ］［ｊ］はフラグであり、１に設定された場合、依存性データは属性に記憶されることを示し、他の場合、依存性データはデスクリプタとして記憶され、（ｉｉｉ）ｄｅｐｅｎｄｅｎｃｙ＿ＩＤ［ｉ］［ｊ］は、依存性データを含んでいる属性又はデスクリプタのＩＤを指定する。ＩＤの長さは、属性の場合は１６ビットであり、デスクリプタの場合は８ビットであることに留意されたい。
【００９３】
［0097］ ｎｅｓｔｉｎｇ＿ｚｅｒｏ＿ｂｉｔは、データ構造のバイトアラインメントを保証するために必要に応じて挿入される、０に設定された１ビットである。
【００９４】
［0098］ 圧縮器構成に基づく符号化アルゴリズム
［0099］ 一実施形態によれば、未加工属性データの符号化は、関連する圧縮器構成において定義された圧縮器ステップの逐次実行によって行われる。符号化アルゴリズムは以下のように概説される。
【００９５】
［00100］ １．処理されている属性とそれの関連する圧縮器との構成（パラメータセット）を取り出す。

【００９６】
［00101］　２．属性構成における属性固有のパラメータ値を用いて圧縮器構成からのデフォルトパラメータ値をオーバーライドすることによって、属性のためにカスタマイズされた圧縮器構成を得る。
【００９７】
［00102］　３．各圧縮器ステップを最初から最後まで連続的に実行する。各圧縮器ステップｉについて、
ａ．アルゴリズムＩＤを使用して変換／圧縮アルゴリズムをルックアップする。
ｂ．アルゴリズムＩＤを見つけることができない場合、安全な伝送チャネルを通して、認証されたソフトウェアリポジトリから対応するアルゴリズムをダウンロードし、インストールする許可をユーザに促す。
ｃ．ステップについて定義された依存性変数がある場合、ｄｅｐｅｎｄｅｎｃｙ＿ｉｓ＿ａｔｔｒｉｂｕｔｅ［ｉ］［］とｄｅｐｅｎｄｅｎｃｙ＿ＩＤ［ｉ］［］とによって指定された、対応する属性又はデスクリプタからの依存性変数の復号されたデータを得る。
ｄ．ステップに関連付けられたパラメータ値と依存性変数のデータオブジェクトとを使用して、入力変数のデータオブジェクトに変換／圧縮アルゴリズムを適用する。それが第１のステップである場合、未加工属性データは処理されている唯一の入力データオブジェクトである。生成された出力は、データオブジェクトとそれらの対応する変数ＩＤとのリストからなる。
ｅ．出力変数が完了したとして指定される場合は、それのデータオブジェクトを圧縮器の出力バッファ中に配置するべきである。他の場合は、さらなる処理のために後続のステップによって、それのステップと変数ＩＤとを通してルックアップされるように、そのデータオブジェクトをデータマップに記憶するべきである。各符号化されたデータブロック／オブジェクトは、それのデータタイプと配列形態とを保存するために、本明細書で指定するように、ｔｙｐｅｄ＿ｄａｔａ構造を使用して記憶されることに留意されたい。
【００９８】
［00103］　表４．タイプされたデータシンタックス

【表４】

シンタックス	タイプ	注

typed_array_data {
data_type_ID
num_array_dims
n_elements = 1
for (i = 0; i < num_array_dims; i++) {
array_dims[i]
n_elements = n_elements * array_dims[i]
}
for (i = 0; i < n_elements; i++) {
data_block[i]
}
}
	
	u(8)
	u(2)
	
	u(32)
	
	u(v)
		
		データタイプのID
		配列次元数、スカラー値の場合0
		
		i番めの配列次元のサイズ
		配列中の要素の数
		
		vはデータタイプＩＤのビットの数に応じる

ｆ．最後のステップの場合、符号化されたブロックの連接である、出力バッファのコンテンツを戻す。

【００９９】

［00104］ 一実施形態による、関連する圧縮器構成に基づく上記の属性データ符号化アルゴリズムの擬似コードが表５に与えられている。

【０１００】

［00105］ 表５．属性データ符号化アルゴリズムの擬似コード

【表５】

```
encode_attribute(attribute_ID) {
  data_map = new Map()                    //中間データオブジェクトのリポジトリ
  out_buffer = new List ()      //出力データバッファ

  //圧縮器ステップを最初から最後まで実行する。
  for (i=0; i<n_compressor_steps; i++) {
    if (i == 0) {
      //それが第１のステップである場合、未加工属性データが
      //処理されるべき唯一の入力データオブジェクトである。
      in_data[0] = attribute_stream
    } else {
      //第２のステップ以降の場合、処理されるべき入力データオブジェクトは
      //前のステップによって生成されたデータオブジェクトから来る。オブジェクトは
      //データマップ中に記憶され、それらの変数名によってルックアップされ得る。
      in_data = NULL
      for (j=0; j<n_in_vars[i]; j++) {
        key = {prev_step_ID[i][j], prev_out_var_ID[i][j]}
        in_data[j] = data_map.get(key)
        data_map.remove(key)
      }
    }

    //依存性変数のデータオブジェクトがある場合、それを取り出す。
    dependency_data[] = NULL
    for (j=0; j<n_dependencies[i]; j++) {
      dependency_data[j] = get_decoded_data(
          dependency_is_attribute[i][j], dependency_ID[i][j])
    }

    //par_ID[i][]とpar_val[i][][][][]とによって指定されたパラメータ値と、
```

```
    // dependency_var_ID[i][]とdependency_data[]とによって指定された必要とされる依存性データオブジェクトとを使用して、
    //algorithm_ID[i]によって指定されたアルゴリズムを
    //in_var_ID[i][]とin_data[]とによって指定された入力データオブジェクトに適用する。
    {out_var_ID[], out_data[]} = apply(algorithm_ID[i], in_var_ID[i][], in_data[],
        par_ID[i][], par_val[i][][][][], dependency_var_ID[i][], dependency_data[])

    //さらなる処理を必要としない、完了した出力データオブジェクトを
    //出力バッファに書き込む。
    for (j=0; j<n_completed_out_vars[i]; j++) {
      for (k=0; k < size(out_var_ID); k++) {
        if (out_var_ID[k] == completed_out_var_ID[i][j]) {
          out_buffer.add(out_data[k])
        }
      }
    }

    //さらなる処理を必要とする不完了した出力データオブジェクトを
    //それのキーとして{compressor_step_ID[i], out_var_ID[j]}をもつdata_mapに記憶する。
    for (j=0; j<size(out_var_ID); j++) {
      out_var_complete = 0
      for (k = 0; k<n_completed_out_vars[i]; k++) {
        if (out_var_ID[j] == completed_out_var_ID[i][k]) {
          out_var_complete = 1
          Break
        }
      }
      if (out_var_complete == 0) {
        key = {compressor_step_ID[i], out_var_ID [j]}
        data_map.add(key, out_data[j])
      }
```

```
    }
  }

  //符号化されたデータは、出力バッファのシリアライズされたバイトストリームによって
与えられる。
  encoded_data = out_buffer.stream()
  return encoded_data
}
```

【０１０１】

［００１０６］　圧縮器構成に基づく復号アルゴリズム

［００１０７］　一実施形態によれば、ファイル圧縮／展開システムは属性復号プロセスを含む。圧縮された属性データの復号は、最後のステップから開始し第１のステップまでアルゴリズムの段階的な逆演算によって、関連する圧縮器構成において定義された圧縮器ステップの逆を伴う。復号アルゴリズムは以下のように概説される。

【０１０２】

［００１０８］　１．処理されている属性とそれの関連する圧縮器との構成（パラメータセット）を取り出す。

【０１０３】

［００１０９］　２．圧縮された属性ストリームから、それぞれ圧縮器ステップにおいて定義された完了した出力変数に対応する、１つ又は複数の符号化されたブロックを抽出する。各符号化されたブロックは、それのデータタイプと配列形態とを保存するために、表４において指定されたｔｙｐｅｄ＿ｄａｔａ構造を使用して記憶される。

【０１０４】

［００１１０］　３．最後のステップから開始し第１のステップに到達するまで、各圧縮器ステップの逆演算を逆順に実行する。各圧縮器ステップｉについて、

ａ．上記ステップ２における符号化されたブロック、又はそれのアップストリーム復号ステップの出力変数のいずれかから、現在のステップのすべての入力変数データを収集する。

ｂ．現在のステップについて定義された依存性変数がある場合、ｄｅｐｅｎｄｅｎｃｙ＿ｉｓ＿ａｔｔｒｉｂｕｔｅ［ｉ］［］とｄｅｐｅｎｄｅｎｃｙ＿ＩＤ［ｉ］［］とによって指定された、対応する属性又はデスクリプタからの依存性変数の復号されたデータを得る。

ｃ．圧縮器パラメータセットから、現在のステップについてのデフォルトパラメータ設定に対する変化を取り出す。

ｄ．関連するパラメータ設定と任意の依存性変数データとを使用して、現在のステップについて指定されたアルゴリズムの逆演算を入力変数データに適用する。

ｅ．それが第１のステップに到達していない場合、ルックアップのためのキーとしてステップＩＤと変数ＩＤとのペアを使用して、後続の復号ステップによってさらに処理されるべき出力変数データを記憶する。

ｆ．それが第１のステップに到達している場合、欠損値を埋め、圧縮器によって復号された属性データとして逆演算の出力変数データを戻す。

【０１０５】

［００１１１］　関連する圧縮器構成に基づく属性復号プロセスのステップが以下の表に与えられている。ａｔｔｒｉｂｕｔｅ＿ｓｔｒｅａｍの初期値は、それのインデックスによって参照されるタイル中の符号化された属性データであることに留意されたい。コード内

で宣言又は初期化されない他の変数は、属性に関連付けられた圧縮器ＩＤによって参照される圧縮器パラメータセットにおいて定義される。

【０１０６】

［００１１２］　表６．属性復号プロセス

【表６】

```
decode_attribute_payload(attribute_ID) {
    encoded_block[] = extract_encoded_blocks(attribute_stream)
    ptr = size(encoded_block)  //encoded_block[]へのポインタ
    data_map = new Map()                    //中間データオブジェクトのリポジトリ

    //最後から最初まで、逆順に
    //圧縮器ステップにおいて定義されたアルゴリズムの逆演算を実行する。
    for (i=(n_compressor_steps – 1); i>=0; i--) {
      //現在のステップのための入力データを収集する。
      idx = 0
      for (j = i + 1; j < n_compressor_steps; j++) {
        for (k = 0; k < n_in_vars[j]; k++) {
          if (prev_step_ID[j][k] == i) {
            curr_in_var_ID[idx] = prev_out_var_ID[j][k]
            key = {j, in_var_ID[j][k]}
            curr_in_var_data[idx] = data_map.get(key)
            idx++
          }
        }
      }
      ptr -= n_complete_out_vars[i]
      for (j = 0; j < n_complete_out_vars[i]; j++) {
        curr_in_var_ID[idx] = complete_out_var_ID[i][j]
        curr_in_var_data[idx] = encoded_block[ptr + j]
        idx++
      }

      //attribute_IDによって参照される属性パラメータセットから
      //依存性データを取り出す。
      idx = 0
```

```
        n_steps_with_dependencies
        for (j = 0; j < n_steps_with_dependencies; j++) {
          if (compressor_step_ID[j] == i) {
            for (k = 0; k < n_dependencies[j]; j++) {
              curr_depend_var_ID[idx] = dependency_var_ID[j][k]
              curr_depend_var_data[idx] = get_decoded_data(
                      dependency_is_attribute[j][k], dependency_ID[j][k])
            }
          }

        //属性に関連付けられた圧縮器ＩＤによって参照される圧縮器パラメータセットから
        //現在のステップのためのパラメータデータを取り出す。
        [curr_par_ID[], curr_par_data[]] = get_parameters(compressor_ID, i)

        //取り出された入力データと依存性データとパラメータデータとを使用して
        //algorithm_ID[i]によって参照されるアルゴリズムの逆演算を適用する。
        [curr_out_var_ID[], curr_out_var_data[]] = apply_inverse(algorithm_ID[i],
              curr_in_var_ID, curr_in_var_data, curr_depend_var_ID,
              curr_depend_var_data, curr_par_ID, curr_par_data)

        if (i == 0) {
          //それが第１の圧縮器ステップに到達している場合、逆演算によって
          //生成されたデータオブジェクトは復号された属性である。
          attribute_stream = curr_out_var_data[0]
        } else{
          //それが第１の圧縮器ステップに到達していない場合、
          //逆演算によって生成された各データオブジェクトcurr_out_var_data[j]は、
          //後続のステップによるルックアップのためのそれのキーとして
          //現在のステップＩＤとcurr_out_var_ID[j]とのペアをもつdata_map中に配置される。
          for (j = 0; j < size(curr_out_var_ID); j++) {
```

```
                key = {i, curr_out_var_ID[j]}
                data_map.add(key, curr_out_var_data[j])
            }
        }
    }

    //出力シンボル中の欠損値を埋める」
    decoded_symbols[31][attribute_ID][tile_index_1][tile_index_2] =
        fill_missing_values(attribute_stream, attribute_missing_value)
}
```

【０１０７】

［００１１３］　圧縮器構成の自動決定のための機構

［００１１４］　知られている普及したファイルタイプの場合、圧縮器構成をデータ圧縮エキスパートのグループがあらかじめ定義し、コミュニティによる使用のために採用することができるが、圧縮器構成を指定する必要があるとき、例えば、新しい又は排他的ファイルタイプを処理するとき、又は知られているタイプのファイルが予定外の属性の包含を可能にするときがある。１つの可能な方法は、場合によっては、異なるオプションと説明とを提供するグラフツールの助けを借りて、ユーザに圧縮器構成を手動で指定させることである。この手法は、しかしながら、データ圧縮に関する多くのユーザ知識及び経験を必要とし、しばしば準最適な構成を生じる。

【０１０８】

［００１１５］　したがって、性能に基づく圧縮器構成の自動決定のための機構が提供される。本システムは、選択のために利用可能な、利用可能な変換及び圧縮アルゴリズムのリストを含むことができ、各々について、アルゴリズムが処理することができるデータのタイプと、アルゴリズム間の可能な接続とが知られている。機構は以下のように働くことができる。ファイル中の各属性について、

［００１１６］　１．利用可能なアルゴリズムと、圧縮器ステップの最大許容数、及びステップのシーケンスにおける循環の禁止など、追加の制約とがわかっていれば、属性のためのすべての可能な圧縮器構成を識別する。

［００１１７］　２．圧縮比及び処理時間などのメトリクスを用いて各圧縮器構成の性能を評価する。それは、圧縮されているファイルからのデータのサブセットに対して、又は特定のファイルタイプのための圧縮器構成テンプレートを構築するためのトレーニングファイルに対して行われ得る。プロセスをスピードアップするために、評価を並列に実行することができる。

［００１１８］　３．最も高い圧縮比と、すべての構成の中で最初の４分の１に入る処理時間とをもつ圧縮器構成など、性能基準に基づいて属性のための最良の圧縮器構成を選択する。

［００１１９］　各属性について最良の圧縮器構成を選択した後に、

［００１２０］　４．ファイルの処理のために必要とされるすべての命令とパラメータとを含んでいる属性と圧縮器パラメータセットデータ構造を生成する。

［００１２１］　５．生成された属性及び圧縮器構成が同じタイプの他のファイルに適用されるものである場合、それらのデータ構造をリポジトリに記憶し、ファイルタイプ又はファイル拡張子による将来における使用のためにルックアップすることができる。

［００１２２］　６．ファイルを圧縮するために、エンコーダは、生成された構成を使用し

て、各属性に上述の符号化アルゴリズムを適用し、圧縮された属性データとともに構成データ構造を記憶することができる。

［００１２３］　７．圧縮されたファイルを復号するために、デコーダは、ファイルに付随する構成データを使用して、各属性に上述の復号アルゴリズムを適用することができる。

【０１０９】

［００１２４］　実施例１－スパース表現データのための圧縮器構成

［００１２５］　以下は、スパースシングルセルＲＮＡ表現データの圧縮のための属性及び圧縮器構成を適用する非限定的な例である。２つの圧縮器構成を評価した。

【０１１０】

［００１２６］　表７－シングルセル表現データの圧縮のための圧縮器構成１

【表７】

ステップ	アルゴリズム	in_var_name	n_out_vars	out_var_name	out_var_complete
1	スパース変換	(入力：未加工属性値の行列)	3	[“row_idx”, “col_idx”, “value”]	[0, 0, 0]
2	BSC圧縮	[“row_idx”]	1	[“”]	[1]
3	BSC圧縮	[“col_idx”]	1	[“”]	[1]
4	BSC圧縮	[“value”]	1	[“”]	[1]

【０１１１】

［００１２７］　表７に要約された圧縮器構成１の場合、表現値の入力行列は、最初に、非０値をもつセルのみを登録する、３つの出力変数、すなわち、ｒｏｗ＿ｉｄｘとｃｏｌ＿ｉｄｘとｖａｌｕｅとを生成するスパース変換を受ける。次いで、ステップ１からの出力変数の各々にＢＳＣ圧縮が単独で適用され、その結果、圧縮器ステップにおいてそれらが定義された順序で連接された３つのデータブロックからなる符号化されたデータストリームが生じる。

【０１１２】

［００１２８］　表８－シングルセル表現データの圧縮のための圧縮器構成２

【表８】

ステップ	アルゴリズム	in_var_name	n_out_vars	out_var_name	out_var_complete
1	スパース変換	(入力：未加工属性値の行列)	3	[“row_idx”, “col_idx”, “value”]	[0, 0, 0]
2	デルタ変換	[“row_idx”]	1	[“row_idx_delta”]	[0]
3	BSC圧縮	[“row_idx_delta”]	1	[“”]	[1]
4	RLE変換	[“col_idx”]	2	[“col_idx_value”, “col_idx_count”]	[0, 0]
5	BSC圧縮	[“col_idx_val”]	1	“”	[1]
6	BSC圧縮	[“col_idx_count”]	1	“”	[1]
7	BSC圧縮	[“value”]	1	“”	[1]

【０１１３】

［００１２９］　表８に要約された圧縮構成２の場合、第１の構成のように、表現値の入力行列は、最初に、３つの出力変数、すなわち、ｒｏｗ＿ｉｄｘとｃｏｌ＿ｉｄｘとｖａｌｕｅとを生成するスパース変換を受ける。変数ｒｏｗ＿ｉｄｘ＿ｄｅｌｔａに記憶されるデルタ値を生成するために、デルタ変換がｒｏｗ＿ｉｄｘに適用され、デルタ値は、次いでＢＳＣアルゴリズムによって圧縮される。２つの出力変数、すなわち、ｃｏｌ＿ｉｄｘ＿ｖａｌとｃｏｌ＿ｉｄｘ＿ｃｏｕｎｔとを生成するために、ランレングス符号化（ＲＬＥ）がｃｏｌ＿ｉｄｘに適用され、ｃｏｌ＿ｉｄｘ＿ｖａｌ及びｃｏｌ＿ｉｄｘ＿ｃｏｕｎｔは、次いでＢＳＣによって単独で圧縮される。ステップ１からのｖａｌｕｅ変数中のデータはＢＳＣによって直接圧縮される。これにより、圧縮器ステップにおいてそれらが定義されたように連接された４つのデータブロックからなる符号化されたデータストリームが生じる。

【０１１４】

［００１３０］　性能を評価するために、圧縮前に１５２ＭＢのサイズをもつ、Ｍａｒｋｅｔ　Ｍａｔｒｉｘフォーマットのファイルに記憶されたシングルセルＲＮＡ表現データに２つの圧縮器構成を適用した。Ｇｚｉｐ及びＢＳＣによる直接圧縮の後に、ファイルサイズはそれぞれ４７ＭＢ及び４２ＭＢになった。

【０１１５】

［００１３１］　表９は、符号化アルゴリズムを使用して圧縮器構成１及び２を適用した後

の様々なデータサイズを要約している。ＢＳＣによる直接圧縮の後の４２ＭＢのファイルサイズに関して、圧縮器構成１及び２の圧縮比はそれぞれ５．１９及び３．６８である。圧縮器構成１は圧縮比に関してより良い性能を有するので、圧縮器構成は、一般にスパース表現データの圧縮に使用するために選択されるべきである。

【０１１６】

［００１３２］　表９－圧縮器構成１及び２についての圧縮データサイズ

【表９】

圧縮器構成	ステップ	変数名	BSC圧縮後のブロックサイズ	BSC圧縮後の総データサイズ
1	2	row_idx	5.1 MB	8.1 MB
	3	col_idx	27 KB	
	4	value	3.1 MB	
2	3	row_idx_delta	8.3 MB	11.4 MB
	5	col_idx_val	<1 KB	
	6	col_idx_count	12 KB	
	7	value	3.1 MB	

【０１１７】

［００１３３］　本実施例は、本方法が、（ＢＳＣによる直接圧縮など）従来技術の方法に対して、圧縮されたデータファイルのサイズを著しく低減することを証明する。これは、ファイル圧縮／展開システムの機能及び効率を改善し、コンピュータシステム又はデータベースの記憶能力に革命を起こす。さらに、１つ又は複数の属性の各々について数百万個又は何十億個ものデータ点を含むファイルを処理するとき、圧縮は、人間の心がペンと鉛筆とを用いても実行する能力を備えていないことである、数百万又は何十億もの計算を含む。

【０１１８】

［００１３４］　本明細書で定義し、使用するすべての定義は、辞書定義、参照により組み込まれる文書における定義、及び／又は定義された用語の通常の意味を統制することを理解されるべきである。

【０１１９】

［００１３５］　明細書及び特許請求の範囲において本明細書で使用される、単数形の要素は、明らかに反対に示されていない限り、「少なくとも１つ」を意味することを理解されるべきである。

【０１２０】

［００１３６］　明細書及び特許請求の範囲において本明細書で使用される、「及び／又は」というフレーズは、そのように結合された要素、すなわち、ある場合には接続的に存在する要素、及び他の場合には離接的に存在する要素の、「いずれか又は両方」を意味することを理解されるべきである。「及び／又は」を用いて記載される複数の要素は、同様に、すなわち、そのように結合された要素のうちの「１つ又は複数」と解釈されるべきである。「及び／又は」節によって明確に識別される要素以外の他の要素は、明確に識別されるそれらの要素に関係があってもなくても、随意に存在し得る。

【０１２１】
［００１３７］明細書及び特許請求の範囲において本明細書で使用する際、「又は」は、上記で定義した「及び／又は」と同じ意味を有することを理解されるべきである。例えば、リスト中の項目を分離するとき、「又は」又は「及び／又は」は包括的であるとして解釈される、すなわち、いくつかの要素又は要素のリスト、及び、随意に、追加の記載されていないいくつかの項目又はそのような項目のリストの少なくとも１つを含むが、それらのうちの２つ以上をも含むとして解釈されるものとする。「のうちのただ１つ」又は「のうちの厳密に１つ」、又は、特許請求の範囲において使用されるとき、「からなる」など、明らかに反対に示されている用語のみが、いくつかの要素又は要素のリストのうちの厳密に１つの要素を含むことを指す。一般に、本明細書で使用する「又は」という用語は、「いずれか」、「のうちの１つ」、「のうちのただ１つ」又は「のうちの厳密に１つ」など、排他性の用語が後続するときに、排他的な代替を示す（すなわち「一方又は他方であるが、両方ではない」）としてのみ解釈されるものとする。
【０１２２】
［００１３８］明細書及び特許請求の範囲において本明細書で使用する際、「少なくとも１つの」というフレーズは、１つ又は複数の要素のリストに関して、要素のリスト中の要素のうちのいずれか１つ又は複数から選択される少なくとも１つの要素を意味するが、必ずしも要素のリスト内に明確に記載されたありとあらゆる要素のうちの少なくとも１つを含み、要素のリスト中の要素の任意の組合せを除外しないとは限らないことを理解されるべきである。この定義はまた、要素が、「少なくとも１つの」というフレーズがそれを指す、要素のリスト内で明確に識別される要素以外に、明確に識別されるそれらの要素に関係があってもなくても、随意に、存在し得ることを可能にする。
【０１２３】
［００１３９］また、明らかに反対に示されていない限り、２つ以上のステップ又は行為を含む、本明細書でクレームされるいかなる方法においても、ステップの順序又は方法の行為は、必ずしもステップ又は方法の行為が具陳されている順序に限定されるとは限らないことを理解されるべきである。
【０１２４】
［００１４０］特許請求の範囲において、並びに上記の明細書において、「有する」、「含む」、「搬送する」、「有している」、「含んでいる」、「伴う」、「保持する」、「から構成される」などのすべての移行句はオープンエンドである、すなわち、限定はしないが…を含む、を意味することが理解されるべきである。「からなる」及び「本質的に…からなる」という移行句のみは、それぞれクローズド又はセミクローズド移行句である。
【０１２５】
［００１４１］いくつかの発明的実施形態について本明細書で説明し、図示したが、当業者は、本明細書で説明する機能を実行するための、及び／又は本明細書で説明する結果及び／又は本明細書で説明する利点のうちの１つ又は複数を得るための、様々な他の手段及び／又は構造を容易に想定するであろう、及び、そのような変更及び／又は改変の各々は、本明細書で説明した発明的実施形態の範囲内であると考えられる。より一般的には、当業者は、本明細書で説明するすべてのパラメータ、寸法、材料、及び構成は例示的なものであること、及び実際のパラメータ、寸法、材料、及び／又は構成は、発明的教示がそれのために使用される特定の適用例又は複数の適用例に依存することを容易に理解するであろう。当業者は、ルーチン実験を使用するだけで、本明細書で説明した特定の発明的実施形態の多くの等価物を認識するか、又は確認することが可能であろう。したがって、上記の実施形態は単に例として提示されること、及び、添付の特許請求の範囲及びそれの等価物の範囲内で、具体的に説明し、クレームされるもの以外の発明的実施形態が実施され得ることが理解されるべきである。本開示の発明的実施形態は、本明細書で説明する各個々の特徴、システム、物品、材料、キット、及び／又は方法を対象とする。さらに、２つ又はそれ以上のそのような特徴、システム、物品、材料、キット、及び／又は方法の任意の組合せは、そのような特徴、システム、物品、材料、キット、及び／又は方法が相互に矛

盾しない場合、本開示の発明的範囲内に含まれる。

【図面】

【図１】

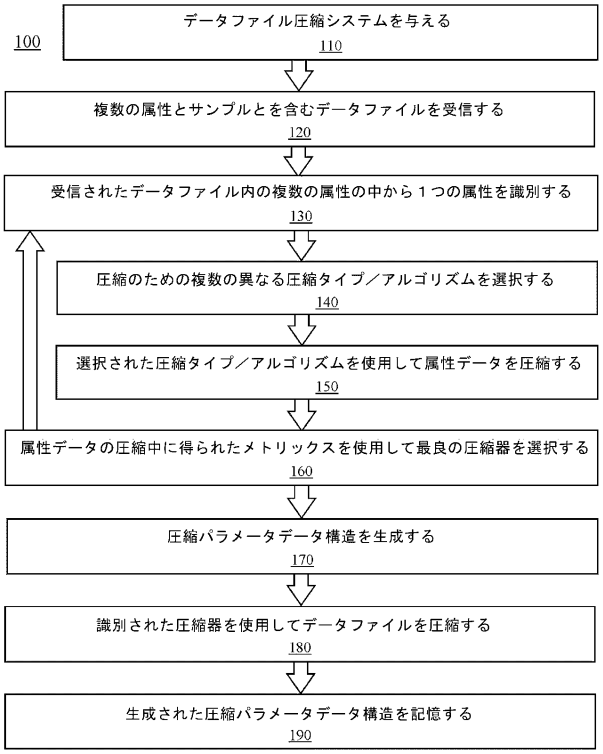


図１

【図２】

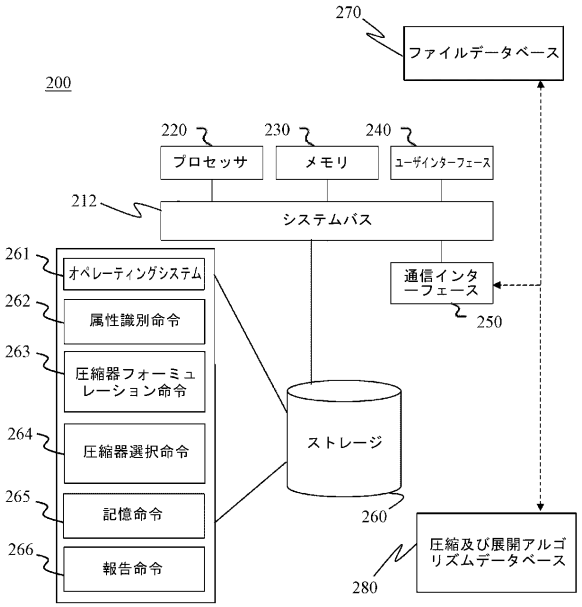


図２

【図３】

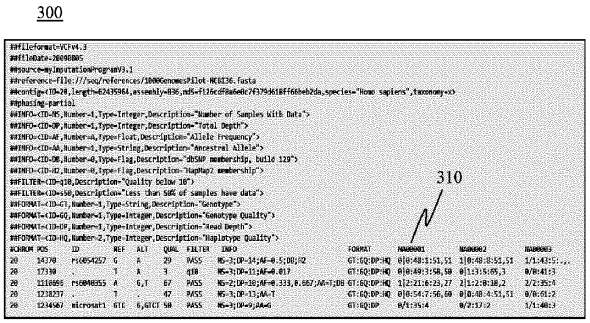


FIG. 3

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2021/081320

A. CLASSIFICATION OF SUBJECT MATTER
INV. H03M7/30
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H03M G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 3 229 444 A1 (HUAWEI TECH CO LTD [CN]) 11 October 2017 (2017-10-11) paragraph [0007]; figure 1 paragraph [0021] - paragraph [0038]	1-15
X	EP 2 381 579 A1 (ITRON METERING SOLUTIONS UK LTD [GB]) 26 October 2011 (2011-10-26) paragraph [0010] - paragraph [0011]	1-15
X	CN 108 306 650 A (XIAMEN JIYUAN TECH CO LTD) 20 July 2018 (2018-07-20) paragraph [0052] - paragraph [0057] paragraph [0099] - paragraph [0107]	1-15

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 February 2022	01/03/2022

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Belardinelli, Carlo

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2021/081320

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Anonymous: "Data compression", Wikipedia, 19 November 2020 (2020-11-19), pages 1-18, XP055887736, Retrieved from the Internet: URL:https://en.wikipedia.org/w/index.php?title=Data_compression&oldid=989504324 [retrieved on 2022-02-04]	1,10
A	page 1 - page 2	2-9, 11-15

X	Anonymous: "Image compression", Wikipedia, 9 September 2020 (2020-09-09), pages 1-5, XP055887122, Retrieved from the Internet: URL:https://en.wikipedia.org/w/index.php?title=Image_compression&oldid=977550343 [retrieved on 2022-02-03]	1,10
A	page 1 - page 2	2-9, 11-15

3

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2021/081320

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 3229444 A1	11-10-2017	CN 107172886 A	15-09-2017
		EP 3229444 A1	11-10-2017
		JP 6344486 B2	20-06-2018
		JP 2018503882 A	08-02-2018
		SG 11201703157Y A	30-08-2017
		US 2018191370 A1	05-07-2018
		WO 2017113124 A1	06-07-2017
EP 2381579 A1	26-10-2011	NONE	
CN 108306650 A	20-07-2018	NONE	

Form PCT/ISA/210 (patent family annex) (April 2005)

フロントページの続き

,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

フィリップス インターナショナル ビー． ヴィ． インテレクチュアル プロパティー アンド スタンダーズ

【要約の続き】

の方法（１００）。