(19)日本国特許庁(JP)

# (12)特許公報(B2)

(11)特許番号
特許第7568276号
(P7568276)

(45)発行日　令和6年10月16日(2024.10.16)

(24)登録日　令和6年10月7日(2024.10.7)

(51)国際特許分類
*Ｇ１６Ｂ　40/00　(2019.01)*

ＦＩ
Ｇ１６Ｂ　40/00

請求項の数　13　（全50頁）

(21)出願番号　特願2020-562540(P2020-562540)
(86)(22)出願日　令和1年12月27日(2019.12.27)
(86)国際出願番号　PCT/JP2019/051564
(87)国際公開番号　WO2020/138479
(87)国際公開日　令和2年7月2日(2020.7.2)
審査請求日　令和4年11月29日(2022.11.29)
(31)優先権主張番号　特願2018-247959(P2018-247959)
(32)優先日　平成30年12月28日(2018.12.28)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　504176911
国立大学法人大阪大学
大阪府吹田市山田丘１番１号
(74)代理人　100136629
弁理士　鎌田 光宜
(74)代理人　100080791
弁理士　高島 一
(74)代理人　100118371
弁理士　▲駒▼谷 剛志
(72)発明者　今野 雅允
大阪府吹田市山田丘１番１号 国立大学法人大阪大学内
(72)発明者　石井 秀始
大阪府吹田市山田丘１番１号 国立大学法人大阪大学内

最終頁に続く

(54)【発明の名称】　個体の形質情報を予測するためのシステムまたは方法

(57)【特許請求の範囲】

【請求項１】

個体の形質情報を予測するためのシステムであって、

複数の個体の遺伝情報と、該複数の個体の形質情報とを格納する格納部であって、該遺伝情報は、少なくとも２種類の情報を含む、格納部と、

該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習するように構成されている学習部と、

該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する、計算部と

を備える、システムであって、

ここで、該学習部が、該複数の個体の遺伝情報を画像化して生成した画像を分割して、画像の各領域と形質情報との関連を学習し、各領域から形質情報の判別能力を有するモデルを生成可能な領域を選択して、画像の各領域から形質情報を予測するモデルを生成するように構成されている、システム。

【請求項２】

前記学習部が、さらに、各領域において、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない領域における遺伝子から、形質情報と相関する変異を有する遺伝子を特定するように構成され、

該計算部は、該形質情報と相関する変異を有する遺伝子の情報に基づいて該個体の形質情報を予測するように構成されている、請求項１に記載のシステム。

【請求項３】
前記発現情報に基づいて形質情報が予測可能かの判定が、
前記画像の各領域に含まれる遺伝子の各発現量を基に前記複数の個体についてクラスタリング分析を行うことと、
前記複数の個体を形質情報に従って群に分割することと、
該群と、クラスタリング分析によって分割されたクラスターとの同一性を算出することと、
該同一性が所定の閾値を超える場合に、発現情報に基づいて形質情報が予測可能であると判定することと
によって行われる、
請求項２に記載のシステム。
【請求項４】
前記学習部が、発現情報に基づいて形質情報が予測可能かを判定した後に、発現情報に基づいて形質情報が予測可能である領域をさらに分割し、各分割領域について、発現情報に基づいて形質情報が予測可能かをさらに判定するように構成され、遺伝子発現量情報のみで判別できる領域から、形質情報と相関する変異を有する遺伝子を特定するように構成されている、
請求項２または３に記載のシステム。
【請求項５】
コンピュータを用いて、形質に関与する遺伝子の変異を同定するための方法であって、
該コンピュータに、複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該コンピュータが、該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該コンピュータが、形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該コンピュータが、該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該コンピュータが、該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、方法。
【請求項６】
形質に関与する遺伝子の変異を同定するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、プログラム。
【請求項７】
形質に関与する遺伝子の変異を同定するための方法をコンピュータに実行させるプログラムを格納したコンピュータ読み取り可能な記録媒体であって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、

該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、記録媒体。
【請求項８】
前記複数の個体の遺伝情報の画像化が、以下の方法：
（Ａ）
コンピュータを用いて、複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子集団の発現データを画像化する方法であって、
該コンピュータが、該遺伝因子集団の配列データおよび該遺伝因子集団の発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有し、該遺伝因子の発現量を、該遺伝因子に対応する領域内の一定領域における色情報および／または該領域中のある色を有する領域の面積の情報に変換することを含む工程
を含む、画像化方法、または
（Ｂ）
コンピュータを用いて、遺伝情報を画像化する方法であって、該遺伝情報は、複数の遺伝因子を含む遺伝因子集団の配列データおよび／または発現データを含み、該方法は、
該コンピュータが、該遺伝因子集団の配列データおよび／または発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有し、
該工程は、該複数の遺伝因子のそれぞれを、該画像データ中の領域に対応付けることを含み、各遺伝因子に対応する領域は、各遺伝因子の相関重みが強いものが近接するように配置されることを特徴とする、工程
を含む、画像化方法
によって行われるように構成されている、請求項１～４のいずれか一項に記載のシステム。
【請求項９】
前記学習部が、データ構造を有するデータを学習に用いるように構成されている、請求項１～４のいずれか一項に記載のシステムであって、該データ構造は、複数の遺伝因子を含む遺伝因子集団の配列情報および複数の遺伝因子を含む遺伝因子集団の発現情報を表す画像データのデータ構造であり、
該画像データは、該複数の遺伝因子に対応付けられた複数の領域を有し、
遺伝因子の配列中の各位置が、該遺伝因子に対応付けられた該領域内の位置に対応付けられており、
該遺伝因子の配列中の各位置における置換、欠失および／または挿入の情報が、該位置に対応する位置における色情報として格納され、
該遺伝因子の発現データが、該領域中のある領域における色情報として、および／または該領域中のある色を有する領域の面積の情報として格納されている、システム。
【請求項１０】
学習部が、
（Ｃ）
コンピュータを用いて、画像と、該画像に対応する情報との関連を予測するモデルを作成するための方法であって、
該コンピュータに、複数の画像および該複数の画像に対応する複数の情報のセットを提供する工程と、

該コンピュータが、該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該コンピュータが、該複数の分割学習データを統合し、該画像と、該画像に対応する情報との関連を予測するモデルを生成する工程と
を含む、方法によって、前記遺伝情報と形質情報との関連を学習するように構成されている、請求項１～４のいずれか一項に記載のシステム。
【請求項１１】
前記複数の分割学習データを得る工程において、各分割学習データの判別能力を検証し、判別力のある分割学習データを選択して統合に供することを特徴とする、請求項１０に記載のシステム。
【請求項１２】
前記統合する工程が、ＨＤＤ上でのＲｅａｄ－Ｗｒｉｔｅファイルの利用、ＣＰＵメモリを最大限利用できるような非線形最適化処理アルゴリズムを最適化することを含む、請求項１０または１１に記載のシステム。
【請求項１３】
前記非線形最適化処理アルゴリズムが、必要なデータを随時メモリに移して計算し、計算結果をＨＤＤに戻すことによって、データサイズに非依存的に計算可能なアルゴリズムである、請求項１２に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、データ解析の分野に関する。より詳細には、個体の遺伝子情報のデータから、個体の形質情報を予測する技術に関する。
【背景技術】
【０００２】
近年の測定技術の発展により、個体の遺伝子情報について、より多岐にわたる情報を大量に収集することが可能となっている。例えば、ゲノム配列を含めた核酸配列、遺伝子の発現情報、非コーディング核酸の発現情報、核酸のエピジェネティック修飾などの情報を収集することが可能である。個体の形質は、遺伝子情報に基づいて画定されているということを前提とすれば、原理上は、遺伝子情報を網羅的に取得することができれば、個体の形質について予め予測することができるはずである。しかしながら、個体の遺伝子情報は非常に膨大な情報量を有し、また、形質への寄与はさまざまな因子の複合的な影響を受けるものであるため、このような予測は未だ困難である。
【発明の概要】
【課題を解決するための手段】
【０００３】
本開示の１つの態様において、個体の形質情報を予測するためのシステムまたはそれを用いる方法、プログラムおよび記録媒体が提供される。本開示のこの態様は、複数の個体の情報から学習することによって、ある個体の遺伝子情報から、当該個体の形質情報を予測し、予測結果を表示することを可能とすることを企図している。例えば、複数の個体の遺伝情報と、当該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習することができ、とりわけ、遺伝情報として、複数の遺伝情報（例えば、遺伝因子の配列情報（例えば、変異情報）、発現情報、および修飾情報（例えば、メチル化情報）など）を用いて学習し、その学習に基づいて予測し、その結果を表示することが可能である。
【０００４】
本開示の１つの実施形態において、学習は、複数の個体の遺伝情報を画像化して学習することを含み得る。そのような画像化は、例えば、本明細書の他の部分に詳述されるようにして行うことができる。また、画像化されたデータは、本明細書の他の部分に詳述されるようなデータ形式を有するものであり得る。これは、複数種の遺伝情報に関する大量のデータを同時に人工知能によって学習する際の人工知能のパフォーマンスを最大化させ得

る。
【０００５】
本開示の１つの実施形態において、学習は、遺伝情報を分割して、部分遺伝情報と形質情報との関連を学習した後に、複数の部分遺伝情報と形質情報との関連を統合し、遺伝情報と形質情報との関連を学習するように行われ得る。これにより、遺伝情報におけるデータの量における制約を解消し得る。
【０００６】
例えば、本開示の例として、以下の項目が挙げられる。
［項目Ａ１］
個体の形質情報を予測するためのシステムであって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを格納する格納部であって、該遺伝情報は、少なくとも２種類の情報を含む、格納部と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習するように構成されている学習部と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する、計算部と
を備える、システム。
［項目Ａ２］
前記学習部が、前記複数の個体の遺伝情報を画像化して学習するように構成されている、前記項目に記載のシステム。
［項目Ａ３］
前記学習部が、前記複数の個体の遺伝情報を分割して、部分遺伝情報と形質情報との関連を学習し、複数の部分遺伝情報と形質情報との関連を統合し、前記遺伝情報と形質情報との関連を学習するように構成されている、前記項目のいずれかに記載のシステム。
［項目Ａ４］
前記遺伝情報が、遺伝因子の配列情報（例えば、変異情報）、発現情報、および修飾情報（例えば、メチル化情報）からなる群から選択される、前記項目のいずれかに記載のシステム。
［項目Ａ５］
前記複数の個体の遺伝情報の画像化が、項目Ｂのいずれかに記載の画像化方法によって行われるように構成されている、前記項目のいずれかに記載のシステム。
［項目Ａ６］
前記学習部が、項目Ｃのいずれかに記載のデータ構造を有するデータを学習に用いるように構成されている、前記項目のいずれかに記載のシステム。
［項目Ａ７］
学習部が、項目Ｄのいずれかに記載の方法によって、前記遺伝情報と形質情報との関連を学習するように構成されている、前記項目のいずれかに記載のシステム。
［項目Ａ８］
前記計算部において予測された形質情報から、前記個体の診断および／または個体に対する治療または予防を分析する、分析部と
を備える、前記項目のいずれかに記載のシステム。
［項目Ａ９］
前記計算部において予測された形質情報を表示する、表示部をさらに備える、前記項目のいずれかに記載のシステム。
［項目Ａ１－１］
個体の形質情報を予測するための方法であって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを提供する情報提供工程であって、該遺伝情報は、少なくとも２種類の情報を含む、工程と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習させる学習工程と、

該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する予測工程と
を含む、方法。
［項目Ａ２－１］
個体の形質情報を予測するための方法であって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを提供する情報提供工程であって、該遺伝情報は、少なくとも２種類の情報を含む、工程と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習させる学習工程と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する予測工程と、
該予測された形質情報を表示する表示工程と
を含む、方法。
［項目Ａ３－１］
前記項目のいずれかまたは複数に記載の特徴をさらに備える、前記項目のいずれかに記載の方法。
［項目Ａ１－２］
個体の形質情報を予測するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝情報と、該複数の個体の形質情報とを提供する情報提供工程であって、該遺伝情報は、少なくとも２種類の情報を含む、工程と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習させる学習工程と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する予測工程と
を含む、プログラム。
［項目Ａ２－２］
前記方法は、前記予測された形質情報を表示する表示工程をさらに含む、前記項目に記載のプログラム。
［項目Ａ３－２］
前記項目のいずれかまたは複数に記載の特徴をさらに備える前記項目のいずれかに記載のプログラム。
［項目Ａ１－３］
個体の形質情報を予測するための方法をコンピュータに実行させるプログラムを格納した記録媒体であって、該方法は、
複数の個体の遺伝情報と、該複数の個体の形質情報とを提供する情報提供工程であって、該遺伝情報は、少なくとも２種類の情報を含む、工程と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習させる学習工程と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する予測工程と
を含む、記録媒体。
［項目Ａ２－３］
前記方法は、前記予測された形質情報を表示する表示工程をさらに含む、前記項目のいずれかに記載の記録媒体。
［項目Ａ３－３］
前記項目のいずれかまたは複数に記載の特徴をさらに備える、前記項目のいずれかに記載の記録媒体。
［項目Ｂ１］
複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子

集団の発現データを画像化する方法であって、
該遺伝因子集団の配列データおよび該遺伝因子集団の発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有する、工程
を含む、方法。
［項目Ｂ２］
前記複数の遺伝因子のそれぞれが、前記画像データ中の領域に対応付けられており、前記画像データを生成する工程が、
前記遺伝因子の発現量を、該遺伝因子に対応する領域内の一定領域における色情報および／または該領域中のある色を有する領域の面積の情報に変換する工程を含む、
前記項目に記載の方法。
［項目Ｂ２－１］
複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子集団の発現データを画像化する方法をコンピュータに実行させるプログラムであって、該方法は
該遺伝因子集団の配列データおよび該遺伝因子集団の発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有する、工程
を含む、プログラム。
［項目Ｂ３］
遺伝情報を画像化する方法であって、該遺伝情報は、複数の遺伝因子を含む遺伝因子集団の配列データおよび／または発現データを含み、該方法は、
該遺伝因子集団の配列データおよび／または発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有し、
該工程は、該複数の遺伝因子のそれぞれを、前記画像データ中の領域に対応付けることを含み、各遺伝因子に対応する領域は、各遺伝因子の相関重みが強いものが近接するように配置されることを特徴とする、工程
を含む、方法。
［項目Ｂ４］
前記画像データを生成する工程が、前記遺伝因子について必要な画像データ中の領域の面積を算出することをさらに含む、前記項目に記載の方法。
［項目Ｂ４－１］
遺伝情報を画像化する方法をコンピュータに実行させるプログラムであって、該遺伝情報は、複数の遺伝因子を含む遺伝因子集団の配列データおよび／または発現データを含み、該方法は、
該遺伝因子集団の配列データおよび／または発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有し、
該工程は、該複数の遺伝因子のそれぞれを、前記画像データ中の領域に対応付けることを含み、各遺伝因子に対応する領域は、各遺伝因子の相関重みが強いものが近接するように配置されることを特徴とする、工程
を含む、プログラム。
［項目Ｂ５］
前記相関重みが、
遺伝因子間の相関解析から強い相関を有する遺伝因子の組み合わせを抽出し、
各遺伝因子についての強い相関遺伝因子を抽出し、
抽出された該遺伝因子を用いた変数選択重回帰を行い、
該変数選択重回帰の結果から相関重みを算出すること
によって算出される、前記項目のいずれかに記載の方法。

［項目Ｂ６］
前記遺伝因子集団の配列データが、親細胞から娘細胞に遺伝形質を伝搬するイベントに関わる因子の配列データを含む、前記項目のいずれかに記載の方法。
［項目Ｂ７］
前記遺伝因子集団の発現データが、当世代のみの情報伝達に関わる因子の発現データを含む、前記項目のいずれかに記載の方法。
［項目Ｂ８］
前記配列データおよび発現データが、同一の個体の遺伝因子のものである、前記項目のいずれかに記載の方法。
［項目Ｂ９］
前記複数の遺伝因子のそれぞれが、前記画像データ中の領域に対応付けられており、前記画像データを生成する工程が、
ある遺伝因子の配列における変異の位置および型の情報を、該遺伝因子に対応する領域内の位置および色情報に変換する工程
を含む、前記項目のいずれかに記載の方法。
［項目Ｂ１０］
前記画像データを生成する工程が、
ある遺伝因子の配列における修飾の情報を、該遺伝因子に対応する領域内の位置および色情報に変換する工程
をさらに含む、前記項目のいずれかに記載の方法。
［項目Ｂ１１］
前記遺伝因子集団の発現データが、転写ユニットの発現データを含む、前記項目のいずれかに記載の方法。
［項目Ｂ１２］
前記遺伝因子集団の発現データが、ｍＲＮＡの発現データを含む、前記項目のいずれかに記載の方法。
［項目Ｂ１３］
前記ｍＲＮＡの発現データが、ｍＲＮＡの発現量、スプライシング、転写開始点、および／またはエピジェネティック修飾のデータを含む、前記項目のいずれかに記載の方法。
［項目Ｂ１４］
前記遺伝因子集団の発現データが、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、および／または長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡの発現データを含む、前記項目のいずれかに記載の方法。
［項目Ｂ１５］
前記遺伝因子集団の発現データが、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、および／または長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡの発現量、スプライシング、転写開始点、および／またはエピジェネティック修飾のデータを含む、前記項目のいずれかに記載の方法。
［項目Ｂ１６］
個体の遺伝因子の配列情報および発現情報から該個体の形質情報を予測するモデルを作成するための方法であって、
複数の個体の遺伝因子の配列情報および発現情報を前記項目のいずれかのいずれか１項に記載の方法によって画像化し、画像データを提供する工程と、
該複数の個体の形質情報を提供する工程と、
該画像データおよび該形質情報から、深層学習により、形質と相関する画像中の特徴表現を抽出する工程と
を含む、方法。
［項目Ｂ１－１］
複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子集団の発現データを画像化する方法をコンピュータに実行させるプログラムであって、該

方法は、
該遺伝因子集団の配列データおよび該遺伝因子集団の発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有する、工程
を含む、プログラム。
［項目Ｂ１－２］
複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子集団の発現データを画像化する方法をコンピュータに実行させるプログラムを格納した記録媒体であって、該方法は、
該遺伝因子集団の配列データおよび該遺伝因子集団の発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有する、工程
を含む、記録媒体。
［項目Ｂ１－３］
複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子集団の発現データを画像化する方法を実行するシステムであって、該システムは、
該遺伝因子集団の配列データおよび該遺伝因子集団の発現データを格納する画像データを生成する画像生成部であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有する、画像生成部と
該遺伝因子集団の配列データと、該遺伝因子集団の発現データと、該画像データを格納するデータ格納部と
を備える、システム。
［項目Ｂ１６－１］
個体の遺伝因子の配列情報および発現情報から該個体の形質情報を予測するモデルを作成するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を項目Ｂ１～Ｂ１５のいずれか１項に記載の方法によって画像化し、画像データを提供する工程と、
該複数の個体の形質情報を提供する工程と、
該画像データおよび該形質情報から、深層学習により、形質と相関する画像中の特徴表現を抽出する工程と
を含む、プログラム。
［項目Ｂ１６－２］
個体の遺伝因子の配列情報および発現情報から該個体の形質情報を予測するモデルを作成するための方法をコンピュータに実行させるプログラムを格納した記録媒体であって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を前記項目のいずれかに記載の方法によって画像化し、画像データを提供する工程と、
該複数の個体の形質情報を提供する工程と、
該画像データおよび該形質情報から、深層学習により、形質と相関する画像中の特徴表現を抽出する工程と
を含む、記録媒体。
［項目Ｂ１６－３］
個体の遺伝因子の配列情報および発現情報から該個体の形質情報を予測するモデルを作成するための方法を実行するシステムであって、該システムは、
複数の個体の遺伝因子の配列情報および発現情報を前記項目のいずれかに記載の方法によって画像化し、画像データを提供する画像生成部と、
該複数の個体の形質情報と、該画像データを格納するデータ格納部と、
該画像データおよび該形質情報から、深層学習により、形質と相関する画像中の特徴表現を抽出する学習部と
を備える、システム。

［項目Ｃ１］
複数の遺伝因子を含む遺伝因子集団の配列情報および複数の遺伝因子を含む遺伝因子集団の発現情報を表す画像データのデータ構造であって、
該画像データは、該複数の遺伝因子に対応付けられた複数の領域を有し、
遺伝因子の配列中の各位置が、該遺伝因子に対応付けられた該領域内の位置に対応付けられており、
該遺伝因子の配列中の各位置における置換、欠失および／または挿入の情報が、該位置に対応する位置における色情報として格納され、
該遺伝因子の発現データが、該領域中のある領域における色情報として、および／または該領域中のある色を有する領域の面積の情報として格納されている、データ構造。
［項目Ｃ２］
前記遺伝因子の配列中の各位置におけるエピジェネティクス修飾の情報が、該位置に対応する位置における色情報としてさらに格納される、前記項目に記載のデータ構造。
［項目Ｃ３］
前記複数の遺伝因子におけるｍｉＲＮＡの配列中の各位置におけるメチル化が、該位置に対応する位置における色情報として格納される、前記項目のいずれかに記載のデータ構造。
［項目Ｃ４］
前記画像データが、行および列を有するマトリックスであり、前記各位置が、行および列の組み合わせとして格納される、前記項目のいずれかに記載のデータ構造。
［項目Ｃ５］
配列情報および発現情報を表す画像データのデータ構造であって、該画像データは、行および列を有するマトリックスであり、該画像データ中の各位置が、行および列の組み合わせとして格納され、
該配列情報は、ゲノム上の領域のＤＮＡ配列を含み、該ゲノム上の領域は、遺伝子、エクソン、イントロン、非発現領域、および／またはｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡをコードする領域を含み、
該発現情報は、ｍＲＮＡ、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、および／または長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡからなる群から選択される転写ユニットの発現量、スプライシング、転写開始点、および／またはエピジェネティック修飾の情報を含み、
該画像データは、各ゲノム上の領域および／または転写ユニットに対応付けられた複数の領域を有し、
該ゲノム上の領域に対応付けられた領域は、該ゲノム上の領域の長さに依存した数の列および一定数の行からなり、
該ゲノム上の領域の配列中の各位置が、該ゲノム上の領域に対応付けられた該領域内の奇数列における位置に対応付けられており、
該ゲノム上の領域の配列中の各位置における置換、欠失および／または挿入の情報が、該位置に対応する奇数列における位置における色情報として格納され、該色情報は、変異が存在しないことを示す色情報、Ａに置換されていることを示す色情報、Ｔに置換されていることを示す色情報、Ｇに置換されていることを示す色情報、Ｃに置換されていることを示す色情報、欠失していることを示す色情報、または該位置に隣接して挿入が存在することを示す色情報であり、
挿入される配列の情報が、挿入が存在することを示す色情報を有する位置に隣接する偶数列における位置を始点として、挿入される配列を示す色情報が格納され、
該ゲノム上の領域の配列中の各位置におけるエピジェネティック修飾の情報が、該位置に対応する奇数列における位置における色情報として格納され、該色情報は、エピジェネティック修飾が存在しないことを示す色情報、ＤＮＡメチル化されていることを示す色情報、ヒストンメチル化されていることを示す色情報、ヒストンアセチル化されていることを示す色情報、ヒストンユビキチン化されていることを示す色情報、またはヒストンリン

酸化されていることを示す色情報を含み、
あるゲノム上の領域から転写される転写ユニットについて、該転写ユニットの発現量が、該ゲノム上の領域に対応する画像中の領域における色の濃淡として、および／または該領域中のある色を有する領域の面積の情報として格納され、
遺伝子であるゲノム上の領域について、該遺伝子に対応するｍＲＮＡの発現量が、該領域中のある領域における色の濃淡として、および／または該領域中のある色を有する領域の面積の情報として格納されている、データ構造。
［項目Ｄ１］
画像と、該画像に対応する情報との関連を予測するモデルを作成するための方法であって、
複数の画像および該複数の画像に対応する複数の情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データを統合し、該画像と、該画像に対応する情報との関連を予測するモデルを生成する工程と
を含む、方法。
［項目Ｄ２］
前記統合する工程が、ＧＰＵを搭載したＣＰＵマシンを用い、メモリ搭載量を含めたＧＰＵスペックおよびＣＰＵスペックを検出することを含む、前記項目に記載の方法。
［項目Ｄ３］
前記統合する工程が、ＨＤＤ上でのＲｅａｄ－Ｗｒｉｔｅファイルの利用、ＣＰＵメモリを最大限利用できるような非線形最適化処理アルゴリズムを最適化することを含む、前記項目のいずれかに記載の方法。
［項目Ｄ４］
前記非線形最適化処理アルゴリズムが、必要なデータを随時メモリに移して計算し、計算結果をＨＤＤに戻すことによって、データサイズに非依存的に計算可能なアルゴリズムである、前記項目のいずれかに記載の方法。
［項目Ｄ５］
前記非線形最適化処理が、全判別パラメータを最適化することを含む、前記項目のいずれかに記載の方法。
［項目Ｄ６］
前記複数の分割学習データを得る工程において、各分割学習データの判別能力を検証し、判別力のある分割学習データを選択して統合に供することを特徴とする、前記項目のいずれかに記載の方法。
［項目Ｄ１－１］
画像と、該画像に対応する情報との関連を予測するモデルを作成するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の画像および該複数の画像に対応する複数の情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データを統合し、該画像と、該画像に対応する情報との関連を予測するモデルを生成する工程と
を含む、プログラム。
［項目Ｄ１－２］
画像と、該画像に対応する情報との関連を予測するモデルを作成するための方法をコンピュータに実行させるプログラムを格納した記録媒体であって、該方法は、
複数の画像および該複数の画像に対応する複数の情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データを統合し、該画像と、該画像に対応する情報との関連を予測す

るモデルを生成する工程と
を含む、記録媒体。
［項目Ｄ１－２］
画像と、該画像に対応する情報との関連を予測するモデルを作成するシステムであって、該システムは、
複数の画像および該複数の画像に対応する複数の情報のセットを提供するデータ格納部と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得るデータ学習部と、
該複数の分割学習データを統合し、該画像と、該画像に対応する情報との関連を予測するモデルを生成するモデル生成部と
を備える、システム。
［項目Ｅ１］
個体の形質情報を予測するためのシステムであって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを格納する格納部であって、該遺伝情報は、遺伝因子の配列情報および発現情報を含む、格納部と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を、該複数の個体の遺伝情報を画像化して学習するように構成されている、学習部と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する、計算部と
を備え、
ここで、該学習部が、該複数の個体の遺伝情報を画像化して生成した画像を分割して、画像の各領域と形質情報との関連を学習し、各領域から形質情報の判別能力を有するモデルを生成可能な領域を選択して、画像の各領域から形質情報を予測するモデルを生成するように構成されている、システム。
［項目Ｅ２］
個体の遺伝因子の配列情報および発現情報を含む遺伝情報と、該個体の形質情報との関連を予測するモデルを作成するための方法であって、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データから、形質情報の判別能力を有する分割学習データを選択し、画像の各領域から形質情報を予測するモデルを生成する工程と
を含む、方法。
［項目Ｅ３］
個体の遺伝因子の配列情報および発現情報を含む遺伝情報と、該個体の形質情報との関連を予測するモデルを作成するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データから、形質情報の判別能力を有する分割学習データを選択し、画像の各領域から形質情報を予測するモデルを生成する工程と
を含む、プログラム。
［項目Ｆ１］
個体の形質情報を予測するためのシステムであって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを格納する格納部であって、該遺伝情報は、遺伝因子の配列情報および発現情報を含む、格納部と、

該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を、該複数の個体の遺伝情報を画像化して学習するように構成されている、学習部と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝情報から、該個体の形質情報を予測する、計算部と
を備え、
ここで、該学習部が、該複数の個体の遺伝情報を画像化して生成した画像を分割して、画像の各領域と形質情報との関連を学習し、各領域から形質情報の判別能力を有するモデルを生成可能な領域を選択し、各領域において、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない領域における遺伝子から、形質情報と相関する変異を有する遺伝子を特定するように構成され、
該計算部は、該形質情報と相関する変異を有する遺伝子の情報に基づいて該個体の形質情報を予測するように構成されている、システム。
［項目Ｆ１－１］
前記発現情報に基づいて形質情報が予測可能かの判定が、
前記画像の各領域に含まれる遺伝子の各発現量を基に前記複数の個体についてクラスタリング分析を行うことと、
前記複数の個体を形質情報に従って群に分割することと、
該群と、クラスタリング分析によって分割されたクラスターとの同一性を算出することと、
該同一性が所定の閾値（例えば、８０～９０％）を超える場合に、発現情報に基づいて形質情報が予測可能であると判定することと
によって行われる、前記項目に記載のシステム。
［項目Ｆ１－２］
前記学習部が、発現情報に基づいて形質情報が予測可能かを判定した後に、発現情報に基づいて形質情報が予測可能である領域をさらに分割し、各分割領域について、発現情報に基づいて形質情報が予測可能かをさらに判定するように構成され、遺伝子発現量情報のみで判別できる領域から、形質情報と相関する変異を有する遺伝子を特定するように構成されている、前記項目のいずれかに記載のシステム。
［項目Ｆ１－３］
前記発現情報に基づいて形質情報が予測可能でない領域における遺伝子からの形質情報と相関する変異を有する遺伝子の特定が、該領域をさらに分割し、発現情報に基づいて形質情報が予測可能でない領域を絞りこむことをさらに含む、前記項目のいずれかに記載のシステム。
［項目Ｆ２］
形質に関与する遺伝子の変異を同定するための方法であって、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、方法。
［項目Ｆ２－１］
前記発現情報に基づいて形質情報が予測可能かの判定が、
前記画像の各領域に含まれる遺伝子の各発現量を基に前記複数の個体についてクラスタリング分析を行うことと、

前記複数の個体を形質情報に従って群に分割することと、
該群と、クラスタリング分析によって分割されたクラスターとの同一性を算出することと、
該同一性が所定の閾値（例えば、８０～９０％）を超える場合に、発現情報に基づいて形質情報が予測可能であると判定することと
によって行われる、前記項目に記載の方法。
［項目Ｆ２－２］
発現情報に基づいて形質情報が予測可能かを判定した後に、発現情報に基づいて形質情報が予測可能である領域をさらに分割し、各分割領域について、発現情報に基づいて形質情報が予測可能かをさらに判定し、遺伝子発現量情報のみで判別できる領域から、形質情報と相関する変異を有する遺伝子を特定することをさらに含む、前記項目のいずれかに記載の方法。
［項目Ｆ２－３］
前記発現情報に基づいて形質情報が予測可能でない領域における遺伝子からの形質情報と相関する変異を有する遺伝子の特定が、該領域をさらに分割し、発現情報に基づいて形質情報が予測可能でない領域を絞りこむことをさらに含む、前記項目のいずれかに記載の方法。
［項目Ｆ３］
形質に関与する遺伝子の変異を同定するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、プログラム。
［項目Ｆ３－１］
前記発現情報に基づいて形質情報が予測可能かの判定が、
前記画像の各領域に含まれる遺伝子の各発現量を基に前記複数の個体についてクラスタリング分析を行うことと、
前記複数の個体を形質情報に従って群に分割することと、
該群と、クラスタリング分析によって分割されたクラスターとの同一性を算出することと、
該同一性が所定の閾値（例えば、８０～９０％）を超える場合に、発現情報に基づいて形質情報が予測可能であると判定することと
によって行われる、前記項目に記載のプログラム。
［項目Ｆ３－２］
前記方法が、発現情報に基づいて形質情報が予測可能かを判定した後に、発現情報に基づいて形質情報が予測可能である領域をさらに分割し、各分割領域について、発現情報に基づいて形質情報が予測可能かをさらに判定し、遺伝子発現量情報のみで判別できる領域から、形質情報と相関する変異を有する遺伝子を特定することをさらに含む、前記項目のいずれかに記載のプログラム。
［項目Ｆ３－３］
前記発現情報に基づいて形質情報が予測可能でない領域における遺伝子からの形質情報と相関する変異を有する遺伝子の特定が、該領域をさらに分割し、発現情報に基づいて形

質情報が予測可能でない領域を絞りこむことをさらに含む、前記項目のいずれかに記載のプログラム。
【発明の効果】
【０００７】
本開示により、個体の遺伝子情報のデータから、個体の形質情報を予測する手段が提供され、例えば、医療、農業、畜産、食品、環境、薬学（創薬、育薬分野）の分野など、生物が関連する任意の技術分野において有用である。特に医療の分野において、疾患の生じる可能性や、適切な治療、または予測される応答などについての情報を提供することが可能となる。加えて、本開示に係る機械学習方法は、画像を用いる任意の機械学習において、巨大なデータを扱うことを可能にし得る。
【図面の簡単な説明】
【０００８】
【図１】図１は、本開示のシステムの例示的な模式図である。
【図２】図２は、本開示のシステムがクラウド／サーバを用いるなど、物理的に分離した
【図３】図３は、ＤＮＡ／ＲＮＡデータの機械学習を行う工程の例示的な模式図である。
【図４】図４は、ＤＮＡ／ＲＮＡデータの画像化を行う工程の例示的な模式図である。
【図５】図５は、ＤＮＡ／ＲＮＡデータの画像化の際の配置最適化の例示的な模式図である。
【図６】図６は、配置最適化のための遺伝子間の相関解析の例示的な模式図である。
【図７】図７は、分割した画像の学習における、Ｄｅｅｐ　Ｌｅａｒｎｉｎｇ処理の例示的な模式図である。
【図８】図８は、ＧＰＵ分割学習とＣＰＵの非線形最適化の例示的な模式図である。
【図９】図９は、生成したモデルの各Ｅｐｏｃｈ数における正答率を示すグラフである。構築した判別モデルでは、非学習画像を用いた細胞株に対しても１００％の精度で判別することが可能であった。
【図１０】ＤＮＡ変異データとＲＮＡ発現量データの両方を画像化したもの、ＤＮＡ変異データのみの情報を同様に画像化したもの、およびＲＮＡ発現量データのみの情報を同様に画像化したもののそれぞれを機械学習して生成したモデルのそれぞれについての、各Ｅｐｏｃｈ数での学習時に用いた画像での判別可能性と、学習時に未使用の画像での判別可能性とを示すグラフである。
【図１１】図１１は、画像の分割学習を示す模式図である。
【図１２】図１２は、５ＦＵ感受性を学習させた際の領域収束性の違いを示す図である。
【発明を実施するための形態】
【０００９】
以下、本開示を最良の形態を示しながら説明する。本明細書の全体にわたり、単数形の表現は、特に言及しない限り、その複数形の概念をも含むことが理解されるべきである。従って、単数形の冠詞（例えば、英語の場合は「ａ」、「ａｎ」、「ｔｈｅ」など）は、特に言及しない限り、その複数形の概念をも含むことが理解されるべきである。また、本明細書において使用される用語は、特に言及しない限り、当該分野で通常用いられる意味で用いられることが理解されるべきである。したがって、他に定義されない限り、本明細書中で使用される全ての専門用語および科学技術用語は、本開示の属する分野の当業者によって一般的に理解されるのと同じ意味を有する。矛盾する場合、本明細書（定義を含めて）が優先する。
【００１０】
以下に本明細書において特に使用される用語の定義および／または基本的技術内容を適宜説明する。
【００１１】
（定義等）
本明細書において、「全判別パラメータ」とは、分割学習後に統合した画像全体を判別するための判別式におけるパラメータを指す。個別学習での判別分析式では、分割された

画像上の部分データに重みを加えて判別しているため、それぞれ分割した画像間同士では、全く独立した判別式を採用しており、それぞれの相関はない。したがって、最終的な非線形最適化では、各部分学習において求められたパラメータによる判別式を元に、それらを統合した（分割前の画像全体に対する）新しい判別式を作成する。そのために、各部分学習のパラメータを初期値として、ＣＰＵを用いて全体を最適化する処理を実施する。

【００１２】

本明細書において、「Ｏｎ　ｔｈｅ　ｆｌｙ」な処理とは、必要なデータを随時メモリに移して計算し、計算結果をＨＤＤに戻し、それを繰り返す処理を指す。「Ｏｎ　ｔｈｅ　ｆｌｙ」のイメージとしては、メモリを机横の本棚に、ＨＤＤを図書館に例えることができる。机で処理をする際には、データである本が横の本棚にあると処理が早い。一般的には必要な本を一気に本棚に持ってくる。しかしながら、本棚の大きさには限界があるため、必要なデータ（本）を必要なときに随時メモリ（本棚）に移しては計算してＨＤＤ（図書館）に戻し、移しては計算して戻しを繰り返すことによって、大量の本を扱うことができる。本開示の最適化処理において「Ｏｎ　ｔｈｅ　ｆｌｙ」な処理を採用する例としては、最適化処理の最中に、メモリ通信時間はかかるが、（計算時間を犠牲にしても）どんなに大きな学習データでも計算可能なアルゴリズムを採用するということが挙げられる。

【００１３】

本明細書において、「画像」とは、広義には、高次元空間に格納された任意のデータを指し、特に、狭義には平面（二次元空間）に格納されたデータを指す。狭義の画像としては、位置情報と、各位置の色（色調、明度および彩度）情報との組み合わせが挙げられる。「画像化」とは、一次元的に格納されたデータ（例えば、０および１の列）を、高次元に格納されたデータに変換することを指す。

【００１４】

本明細書において、「学習」とは、何らかのデータを用いて、入力に対する有用な出力を行うモデルを形成することを指す。また、入力とそれに対応すべき出力を学習データとして用いる場合、「教師あり学習」と称される。例えば、モデルは、ある遺伝情報を入力とした際に、その遺伝情報から推定される形質（例えば、薬剤耐性）を出力するものなどが挙げられる。

【００１５】

本明細書において、「形質情報」とは、生物または生物の一部（例えば、臓器（器官）、組織または細胞）の有する任意の特徴についての情報を指す。形質情報としては、疾患の特定（がんを例に挙げれば、がん種の特定、がんの悪性度等）や薬剤感受性（がんを例に挙げれば、抗がん剤耐性）等を挙げることができる。

【００１６】

本明細書において、「遺伝因子」とは、生物の活動において情報に基づいて何らかの機能を発揮する任意の因子を指す。例えば、ゲノムＤＮＡ上の遺伝子は、その配列の情報に基づいて、対応するｍＲＮＡに転写される点で、遺伝因子である。また、ｍＲＮＡは、その配列の情報に基づいて、対応するタンパク質等に翻訳される点で遺伝因子である。遺伝因子としては、タンパク質をコードしている遺伝子に加えて、ｍｉＲＮＡをコードするものや、調節領域や非発現領域などが包括的に包含される。したがって、本明細書では、「遺伝因子」としては、遺伝子、ｍＲＮＡ、の他、エクソン、イントロン、非発現領域、ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡ、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡが含まれることが理解される。

【００１７】

本明細書において、「遺伝情報」とは、生物または生物の一部（例えば、組織または細胞）の有する任意の遺伝因子の配列情報および／または発現情報を指す。

【００１８】

本明細書において、「リボ核酸（ＲＮＡ）」は、少なくとも１つのリボヌクレオチド残基を含む分子を意味する。「リボヌクレオチド」とは、β－Ｄ－リボ－フラノース部分の２’位においてヒドロキシル基を有するヌクレオチドを意味する。ＲＮＡには、例えば、メ

ッセンジャーＲＮＡ（ｍＲＮＡ）、トランスファーＲＮＡ（ｔＲＮＡ）、リボゾームＲＮＡ（ｒＲＮＡ）、ロングノンコーディングＲＮＡ（ｌｎｃＲＮＡ）、マイクロＲＮＡ（ｍｉＲＮＡ）が含まれる。

【００１９】

本明細書において、「デオキシリボ核酸（ＤＮＡ）」は、少なくとも１つのデオキシリボヌクレオチド残基を含む分子を意味する。「デオキシリボヌクレオチド」とは、リボヌクレオチドの２’位のヒドロキシル基が水素に置換されているヌクレオチドを意味する。

【００２０】

本明細書において、「メッセンジャーＲＮＡ（ｍＲＮＡ）」とは、ＤＮＡ鋳型を使用することによって作製され、ペプチドまたはポリペプチドをコードしている転写物に関連するＲＮＡを指す。典型的には、ｍＲＮＡは、５’－ＵＴＲ、タンパク質コード領域、および３’－ＵＴＲを含む。ｍＲＮＡの具体的な情報（配列など）は、例えば、ＮＣＢＩ（https://www.ncbi.nlm.nih.gov/）を参照することで利用可能である。

【００２１】

本明細書において、「マイクロＲＮＡ（ｍｉＲＮＡ）」とは、ゲノム上にコードされ、多段階的な生成過程を経て最終的に２０から２５塩基長の微小ＲＮＡとなる機能性核酸を指す。ｍｉＲＮＡの具体的な情報（配列など）は、例えば、ｍｉｒｂａｓｅ（http://mirbase.org）を参照することで利用可能である。

【００２２】

本明細書において、「ロングノンコーディングＲＮＡ（ｌｎｃＲＮＡ）」とは、タンパク質へ翻訳されずに機能する２００ｎｔ以上のＲＮＡを指す。ｌｎｃＲＮＡの具体的な情報（配列など）は、例えば、ＲＮＡｃｅｎｔｒａｌ（http://rnacentral.org/）を参照することで利用可能である。

【００２３】

本明細書において、「リボソームＲＮＡ（ｒＲＮＡ）」とは、リボソームを構成するＲＮＡを指す。ｒＲＮＡの具体的な情報（配列など）は、例えば、ＮＣＢＩ（https://www.ncbi.nlm.nih.gov/）を参照することで利用可能である。

【００２４】

本明細書において、「トランスファーＲＮＡ（ｔＲＮＡ）」とは、アミノアシルｔＲＮＡ合成酵素によりアミノアシル化されることが公知であるｔＲＮＡを指す。ｔＲＮＡの具体的な情報（配列など）は、例えば、ＮＣＢＩ（https://www.ncbi.nlm.nih.gov/）を参照することで利用可能である。

【００２５】

本明細書において、核酸の文脈において使用される「修飾」とは、核酸の構成単位またはその末端の一部または全部が他の原子団と置換されること、または官能基が付加されている状態を指す。ＲＮＡの修飾の集合は「ＲＮＡ　Ｍｏｄｏｍｉｃｓ」「ＲＮＡ　Ｍｏｄ」などとよぶことがあり、これらは、ＲＮＡがトランスクリプトであることから、エピトランスクリプトームと呼ばれることもあり、本明細書では同義で使用される。

【００２６】

本明細書において、核酸の文脈において使用される「メチル化」とは、任意の種類のヌクレオチドの任意の位置のメチル化を指すが、代表的には、アデニンのメチル化（例えば、６位；ｍ６Ａ、１位；ｍ１Ａ）、シトシンのメチル化（例えば、５位；ｍ５Ｃ、３位；ｍ３Ｃ）である。検出された修飾部位は、当該分野で公知の手法を用いて特定することができる。例えばｍ１Ａとｍ６Ａ、ｍ３Ｃとｍ５Ｃについては、それぞれ化学修飾により確定は可能である。例えば、スタンダードとなる合成ＲＮＡを利用して、化学修飾及びＭＡＬＤＩでの測定による挙動が正しいのかを確定することができる。

【００２７】

本明細書において「対象」とは、本開示の分析、診断または検出等の対象となる対象（例えば、ヒト等の生物または生物から取り出した細胞、血液、血清等）をいう。

【００２８】

本明細書において「バイオマーカー」は、ある対象の状態または作用の評価の指標となるものである。本明細書において特に断らない限り、「バイオマーカー」は「マーカー」と称することがある。

【００２９】

本明細書において「診断」とは、被験体における状態（例えば、疾患、障害）などに関連する種々のパラメータを同定し、そのような状態の現状または未来を判定することをいう。本開示の方法、装置、システムを用いることによって、体内の状態を調べることができ、そのような情報を用いて、被験体におけるがんの転移／原発性に関連する状態（例えば、対象が転移性のがんを有するかどうか、がんが原発性であるかどうか）、投与すべき処置または予防のための処方物または方法などの種々のパラメータを選定することができる。本明細書において、狭義には、「診断」は、現状を診断することをいうが、広義には「早期診断」、「予測診断」、「事前診断」等を含む。本開示の診断方法は、原則として、身体から出たものを利用することができ、医師などの医療従事者の手を離れて実施することができることから、産業上有用である。本明細書において、医師などの医療従事者の手を離れて実施することができることを明確にするために、特に「予測診断、事前診断もしくは診断」を「支援」すると称することがある。本開示の技術は、このような診断技術に応用可能である。

【００３０】

本明細書において「治療」とは、ある状態（例えば、疾患または障害）について、そのような状態になった場合に、そのような状態の悪化を防止、好ましくは、現状維持、より好ましくは、軽減、さらに好ましくは消退させることをいい、患者の状態、もしくは状態に伴う１つ以上の症状の、症状改善効果あるいは予防効果を発揮しうることを含む。事前に診断を行って適切な治療を行うことは「コンパニオン治療」といい、そのための診断薬を「コンパニオン診断薬」ということがある。本開示の技術を用いて、遺伝情報を、診断上有用な形質情報と関連付けることによって、このようなコンパニオン治療またはコンパニオン診断において有用であり得る。

【００３１】

本明細書において「予防」とは、正常でない状態（例えば、疾患または障害）とならないように処置することをいう。

【００３２】

本明細書において「予後」という用語は、がん等の疾患または障害などに起因する死亡または進行が起こる可能性を予測することを意味する。予後因子とは疾患または障害の自然経過に関する変数のことであり、これらは、いったん疾患または障害を発症した患者の再発率等に影響を及ぼす。予後の悪化に関連した臨床的指標には、例えば、本開示で使用される任意の細胞指標が含まれる。予後因子は、しばしば、患者を異なった病態をもつサブグループに分類するために用いられる。本開示の技術を用いて遺伝情報を、診断上有用な形質情報と関連付けることによって、対照の遺伝情報に基づいて予後因子を提供することを可能とし得る。

【００３３】

本明細書において「プログラム」は、当該分野で使用される通常の意味で用いられ、コンピュータが行うべき処理を順序立てて記述したものであり、法律上「物」として扱われるものである。すべてのコンピュータはプログラムに従って動作している。現代のコンピュータではプログラムはデータとして表現され、記録媒体または記憶装置に格納される。

【００３４】

本明細書において「記録媒体」は、本開示の方法を実行させるプログラムを格納した記録媒体であり、記録媒体は、プログラムを記録できる限り、どのようなものであってもよい。例えば、内部に格納され得るＲＯＭやＨＤＤ、磁気ディスク、ＵＳＢメモリ等のフラッシュメモリなどの外部記憶装置でありうるがこれらに限定されない。

【００３５】

本明細書において「システム」とは、本開示の方法またはプログラムを実行する構成を

いい、本来的には、目的を遂行するための体系や組織を意味し、複数の要素が体系的に構成され、相互に影響するものであり、コンピュータの分野では、ハードウェア、ソフトウェア、ＯＳ、ネットワークなどの、全体の構成をいう。

【００３６】

（予測システム）

本開示の１つの局面は、個体の形質情報を予測するためのシステムである。システムは、複数の個体の遺伝情報と複数の個体の形質情報とを格納する格納部と、複数の個体の遺伝情報と複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習するように構成されている学習部と、遺伝情報と形質情報との関連に基づき、個体の遺伝情報から個体の形質情報を予測する、計算部とを備え得る。１つの実施形態では、格納部に含まれる遺伝情報は、少なくとも２種類の情報を含み得る。必要に応じて、このシステムは計算部において予測された形質情報から、前記個体の診断および／または個体に対する治療または予防を分析する、分析部をさらに備えることができる。また、必要に応じて、このシステムは計算部において予測された形質情報を表示する、表示部をさらに備えることができる。

【００３７】

本開示はまた、上記システムを実現するプログラム、方法またはこれらを格納した記録媒体として提供されることもできる。

【００３８】

学習部は、複数の個体の遺伝情報を画像化して学習するように構成され得る。同時に、格納部においては、複数の個体の遺伝情報を画像化して格納しておいてもよい。別の実施形態では、学習に際してその都度画像化することも可能である。また、計算部において、個体の遺伝情報を、画像化し、その情報に基づいて個体の形質情報を予測してもよい。画像化は、本明細書の他の箇所に記載される特徴を有する方法またはシステムにより行うことができる。また、画像データは、本明細書の他の箇所に記載されるデータ形式を有するものであってよい。システムは、この他の構成要素を必要に応じて備え得る。例えば、システムは、計算部の出力を表示する表示部を備えてもよい。

【００３９】

１つの実施形態では、学習として、人工知能（ＡＩ）を用いた学習を行う。ＡＩ技術は、「画像」や音声などのデータの処理では特徴表現の抽出などを通じて高い性能を発揮できることが知られていますが、その他のデータでは未だ課題を有しているとされている。１つのポイントは、従来の細胞生物学的な検討で明らかにされてきたように、細胞の「形態」情報は大変重要であるが、この形態情報をゲノム情報に直結させるためには、従来法では、シークエンスする、またはシングル細胞解析などをするなどの方法で、ゲノムの数値データと画像とを人間の目で見て統計的な相関を取ることが必要となっていた。しかし今回の発明では、ゲノム情報を「画像化」することにより、ゲノム情報を画像どうしの土俵に上げることにより、画像間で比較することを可能にして、ＡＩの性能を最大限発揮されることが期待できる。

【００４０】

また、ヒトを対象とする場合、遺伝情報について、個人情報の観点をクリアすることは社会的に大変重要である。この点からも、ゲノム情報の画像化は、「個人情報のシールド」において、基本的な技術の１つになる可能性を秘めている。画像化において、変異情報を取り上げてデータベース化し、この場合にＳＮＰｓを許すように設定すれば、個人の識別に対するシールドになり得る。すなわち、変異情報のみでは、個人識別符号にならないのではないかと考えられる。

【００４１】

本開示で扱われる遺伝情報としては、遺伝因子の配列情報（例えば、変異情報）、発現情報、および／または修飾情報（例えば、メチル化情報）が挙げられる。学習に用いられるデータは、複数の個体のものが一般的に必要であるが、各個体について、全ての種類の遺伝情報が得られる必要はない。

【００４２】

遺伝情報としては、個体の配列情報として、親細胞から娘細胞に遺伝形質を伝搬するイベントに関わる因子であって、核内またはミトコンドリア内に存在し、ＲＮＡポリメラーゼの支配下にあって、タンパク質をコードするコーディング（ｃｏｄｉｎｇ）ＲＮＡまたはｍＲＮＡだけでなく、ノンコーディング（ｎｏｎ－ｃｏｄｉｎｇ）ＲＮＡとして数十塩基までの比較的短鎖のｍｉＲＮＡまたはｓｎｏＲＮＡまたはｓｉＲＮＡまたはｔＲＮＡまたはｒＲＮＡまたはｍｉｔＲＮＡ、さらにはより長鎖ｎｏｎ－ｃｏｄｉｎｇ ＲＮＡをコードするＤＮＡ配列を対象とすることができる。さらに、上記の発現産物の相補部分から離れた非発現領域のＤＮＡ配列を対象として、さらには、ＤＮＡ上のエピジェネティック修飾等も含めて対象とすることができる。個体の発現情報として、個体の遺伝因子（転写ユニット（ＲＮＡおよびｍｉＲＮＡ）の発現量、スプライシング、転写開始点、エピジェネティック修飾等）を含めて、ＲＮＡポリメラーゼの支配下にあって、タンパク質をコードするｃｏｄｉｎｇ ＲＮＡまたはｍＲＮＡだけでなく、ｎｏｎ－ｃｏｄｉｎｇ ＲＮＡとして数十塩基までの比較的短鎖のｍｉＲＮＡまたはｓｎｏＲＮＡまたはｓｉＲＮＡまたはｔＲＮＡまたはｒＲＮＡまたはｍｉｔＲＮＡ、さらにはより長鎖ｎｏｎ－ｃｏｄｉｎｇ ＲＮＡをコードするＤＮＡ配列を対象とすることができる。

【００４３】

本開示で扱われる形質情報としては、特に限定されるものではないが、例えば、個体がある疾患を発症する可能性があるかどうか、または個体がある薬剤に対して応答するかどうか、等が挙げられる。

【００４４】

格納部は、例えば、ＣＤ－Ｒ、ＤＶＤ、Ｂｌｕｅｒａｙ、ＵＳＢ、ＳＳＤ、ハードディスクなどの、システムに格納されるかあるいは離脱した記録媒体であってもよく、あるいは、サーバに格納されてもよく、クラウド上に適宜記録される形式でもよい。

【００４５】

学習部は、人工知能または機械学習を用いて、遺伝情報と形質情報との関連を学習するように構成され得る。本明細書において「機械学習」とは、明示的にプログラミングすることなく、コンピュータに学ぶ能力を与える技術をいう。機能単位が新しい知識・技能を獲得すること、又は既存の知識・技能を再構成することによって、自身の性能を向上させる過程である。経験から学ぶように計算機をプログラミングすることで、細部をプログラミングするのに必要になる手間の多くは減らせ、機械学習分野では、経験から自動的に改善を図れるようなコンピュータプログラムを構築する方法について議論している。データ分析・機械学習の役割としては、アルゴリズム分野と並んで知的処理の基盤になる要素技術であり、通常他の技術と連携して利用され、連携する分野の知識（ドメインスペシフィック（領域特有）知識；例えば、医学分野）が必要である。その応用範囲としては、予測（データを集め、これから起こることを予測する）、探索（集めたデータの中から、何か目立つ特徴を見つける）、検定・記述（データの中のいろいろな要素の関係を調べる）などの役割がある。機械学習は、実世界の目標の達成度を示す指標に基づくものであり、機械学習の利用者が、実世界での目標を把握していなければならない。そして、目的が達成されたときに、良くなるような指標を定式化する必要がある。機械学習は逆問題で、解が解けたかどうかが不明確な不良設定問題である。学習したルールの挙動は確定的ではなく確率（蓋然）的である。何らかの制御できない部分が残ることを前提とした運用上の工夫が必要であり、訓練時と運用時の性能指標をみながら、機械学習の利用者が、データや情報を実世界の目標に合わせて逐次的に取捨選択することも有用である。

【００４６】

機械学習としては、線形回帰、ロジスティック回帰、サポートベクターマシンなどが用いられ得、および交差検証（交差検定、交差確認ともいう。Ｃｒｏｓｓ Ｖａｌｉｄａｔｉｏｎ；ＣＶ）を行うことで、各モデルの判別精度を算出することができる。ランキングした後、１つずつ特徴量を増やして機械学習（線形回帰、ロジスティック回帰、サポートベクターマシンなど）と交差検証を行い、各モデルの判別精度を算出することができる。それにより、最も高い精度のモデルを選択することができる。本開示において、機械学習

は、任意のものを使用することができ、教師付き機械学習として、線形、ロジスティック、サポートベクターマシン（ＳＶＭ）などを利用することができる。

【００４７】

機械学習では論理的推論を行う。論理的推論にはおおまかに３種類あり、演繹（ｄｅｄｕｃｔｉｏｎ）、帰納（ｉｎｄｕｃｔｉｏｎ）、アブダクション（ａｂｄｕｃｔｉｏｎ）、類推（アナロジー）がある。演繹は、ソクラテスは人間、すべての人間は死ぬとの仮説があったときにソクラテスは死ぬとの結論を導き出すもので特殊な結論といえる。帰納は、ソクラテスは死ぬ、ソクラテスは人間との仮説があったときにすべての人間は死ぬとの結論を導き出すもので一般的な規則を導くものである。アブダクションは、ソクラテスは死ぬ、すべての人間は死ぬとの仮定があった時にソクラテスは人間であると導き出すものであり、仮説・説明にあたる。とはいえ、帰納にしてもどう一般化するかは前提によるため、客観的であるとは言えない可能性があることに留意する。類推は、対象Ａと対象Ｂがあり、対象Ａが４つの特徴を持ち、かつ対象Ｂがその特徴のうち共通して３つ持つ場合、対象Ｂは、残り一つの特徴を同様にもち、対象Ａと対象Ｂは同種か類似した近親性を持つと推論するような蓋然的な論理的思考法である。

【００４８】

不可能性には、不可能、非常に困難、未解決の３種類の基本原理がある。また、不可能性には、汎化誤差、ノーフリーランチ定理、醜いアヒルの子定理があり、真のモデルの観測は不可能なので検証できないという不良設定問題に留意する必要がある。

【００４９】

機械学習において、特徴（ｆｅａｔｕｒｅ）・属性（ａｔｔｒｉｂｕｔｅ）とは、予測対象をある側面で見たときに、どのような状態にあるのかを表すものである。特徴ベクトル・属性ベクトルとは、予測対象を記述する特徴（属性）をベクトルの形式にまとめたものである。

【００５０】

本明細書において、「モデル（ｍｏｄｅｌ）」または「仮説（ｈｙｐｏｔｈｅｓｉｓ）」とは、同義に用いられ、入力される予測対象から、予測結果への対象対応を記述する写像、もしくはそれらの候補集合で、数学的な関数か論理式を用いて表現する。機械学習での学習では、訓練データを参照して、モデル集合から真のモデルを最もよく近似すると思われるモデルが選択される。

【００５１】

モデルとしては、生成モデル、識別モデル、関数モデルなどが挙げられる。入力（予測対象）ｘと出力（予測結果）ｙとの写像関係の分類モデルを表現する方針の違いを示すものである。生成モデルは、入力ｘが与えられたときの出力ｙの条件付分布を表現する。識別モデルは、入力ｘと出力ｙの同時分布を表現する。識別モデルと生成モデルは写像関係が確率的である。関数モデルは、写像関係が確定的なもので、入力ｘと出力ｙの確定的な関数関係を表現する。識別モデルと生成モデルでは識別の方がやや高精度といわれることもあるが、ノーフリーランチ定理により基本的には優劣はない。

【００５２】

モデルの複雑さ：予測対象と予測結果の写像関係をより詳細で複雑に記述できるかどうかの度合い。モデル集合が複雑であるほど、一般により多くの訓練データが必要とされる。

【００５３】

写像関係を多項式で表した場合は、高次の多項式の方がより複雑な写像関係を表現できる。高次の多項式の方が、１次式より複雑なモデルといえる。

【００５４】

写像関係を決定木で表した場合、段数の大きな深い決定木の方がより複雑な写像関係を表現できる。したがって、段数の多い定木の方が、少ない決定木より複雑なモデルといえる。

【００５５】

入力と出力の対応関係を表現する方針による分類も可能であり、パラメトリックモデル

では、パラメータによって完全に分布や関数の形状が決定される、ノンパラメトリックモデルでは基本的にデータからその形状が決まり、パラメータが決めるのは滑らかさに限定される。

【００５６】

パラメータ：モデルの分布や関数の集合のうちの一つを指定するための入力で、他の入力と区別してＰｒ［ｙ｜ｘ；θ］やｙ＝ｆ（ｘ；θ）などとも表記される。

パラメトリックでは、訓練データ数と無関係に、ガウス分布の形状は平均・分散パラメータで決定され、ノンパラメトリックでは、ヒストグラムではビン数パラメータで滑らかさのみが決まり、パラメトリックより複雑であるとされる。

【００５７】

機械学習での学習では、訓練データを参照して、モデル集合から真のモデルを最もよく近似すると思われるモデルを選択する。どのような「近似」をするかで、いろいろな学習方法がある。代表的には、最尤推定があり、確率的なモデル集合の中から、訓練データが発生する確率が最も高いモデルを選択する学習の基準である。最尤推定で、真のモデルを最も近似するモデルが選択できる。ＫＬダイバージェンスは、尤度が大きくなると真の分布へのＫＬダイバージェンスは小さくなる。推定の種類は種々あり、推定した予測値やパラメータを求める形式の種類によって異なる。点推定は、最も確実性の高い値を一つだけ求めるもので、最尤推定やＭＡＰ推定など、分布や関数の最頻値を使うもので、最もよく利用される。他方、区間推定では、推定値が存在する範囲を求めるこの範囲に推定値が存在する確率が９５％といった形で統計分野でよく利用される。分布推定では、推定値が存在する分布を求める事前分布を導入した生成モデルと組み合わせてベイズ推定などで利用される。

【００５８】

機械学習では、過学習（過剰適合、ｏｖｅｒ－ｆｉｔｔｉｎｇ）が生じ得る。過学習では、訓練データに合わせ過ぎたモデルを選択したために、経験誤差（訓練データに対する予測誤差）は小さいが、汎化誤差（真のモデルからのデータに対する予測誤差）は大きくなり、本来の学習の目的を達成できていない状態になっている。汎化誤差は、バイアス（候補モデル集合に真のモデルは含まれないことで生じる誤差；単純なモデル集合ほど大きくなる）、バリアンス（訓練データが異なると、異なる予測モデルが選択されることで生じる誤差；複雑なモデル集合ほど大きくなる）、およびノイズ（モデル集合の選択に依存せず、本質的に減らせない真のモデルのばらつき）の三つに分割できる。バイアスとバリアンスは同時には小さくできないから、バイアスとバリアンスのバランスをとって全体の誤差を小さくする。

【００５９】

本明細書において「アンサンブル（アンサンブル学習、アンサンブル法などともいう）」とは、集団学習ともいい、比較的単純な学習モデルと計算量が妥当な学習則とを用い、与えられる例題の重みや初期値の違いなどによって多様な仮説を選び出しこれを組み合わせることによって最終的な仮説を構成し、複雑な学習モデルを学習するのと同等なことを行おうとするものである。本開示の学習において、アンサンブル学習を行ってもよい。

【００６０】

本明細書において「縮約」とは、特徴量という変数を少なくしたり、まとめることをいう。例えば、因子分析とは、複数の変数があったとき、その背後にそれらに影響する構成概念があるものと仮定し、少数の潜在的変数で複数の変数間の関係を説明することであり、小数の変数への変換、すなわち縮約の一形態をいう。この構成概念を説明する潜在的変数を因子という。因子分析は背後に共通した因子が想定できる変数を縮約し、新しい量的な変数を作り出す。

【００６１】

本明細書において、「判別関数」とは、判別するレベル数に連続する数値を割当て、判別するサンプルの並びに対応して作成された数列、すなわち関数である。例えば、判別レベルが２段階で、判別するサンプルをレベルに応じて並べた場合、その数列、すなわち、

判別関数は、例えば、シグモイド関数型を取ることで生成される。また、３段階以上の場合は、工程（階段）関数を用いることができる。モデル近似指数は、判別関数と判別するサンプルの判別レベルの対応を数値で表したものである。両者の差を使う場合は、変動幅を統制し、差分値の絶対値が小さいほど、近似性が高い。また、相関分析を行う場合は、相関係数（ｒ）が高いほど近似性が高い。また、回帰分析を行う場合は、$R^2$値が高いほど近似性が高いと判断される。

【００６２】

本明細書において「重みづけ係数」とは、本開示の計算において、重要である要素をより重要であると計算するように設定するための係数であり、近似係数を含む。例えば、関数をデータに近似させて係数を得られるが、それ自体は、近似の程度を示す記述量でしかなく、それを大小の基準などでランキングしたり、取捨選択したりする場合、特定の特徴量にモデル内における寄与の差を設けるので、重みづけ係数といえる。重みづけ係数は、判別関数の近似指数と同等の意味で用いられ、$R^2$値、相関係数、回帰係数、および残差平方和（判別関数と特徴量の差）等を挙げることができる。

【００６３】

本明細書において、「判別関数モデル」とは、形質などの判別の際に用いられる関数のモデルを言う。例えば、例えば、多層パーセプトロンやＣＮＮといったニューラルネットワークシステムを用いた機械学習による判別モデルを挙げることができるがこれらに限定されない。

【００６４】

学習部は、複数の個体の遺伝情報を分割して、部分遺伝情報と形質情報との関連を学習し、複数の部分遺伝情報と形質情報との関連を統合し、遺伝情報と形質情報との関連を学習するように構成され得る。このような遺伝情報の分割学習は、個体の遺伝情報の情報量の大きさに対処する上で有効であり得る。

【００６５】

本開示において、分析部は、計算部において予測された形質情報から、前記個体の診断および／または個体に対する治療または予防を分析する。形質情報は対象となる個体の情報であるから、他の情報（例えば、疾患情報データベースなど）を参照して、その個体について罹患しているまたはその可能性のある疾患や症状などを診断または診断の補助を行うことができる。診断結果に応じて、他の情報（例えば、疾患情報データベース、医薬品情報データベースなど）を参酌して、適切な治療方法や投薬情報を算出または示唆することができる。

【００６６】

本開示において、表示部は、計算部において予測された形質情報を表示する。表示部としては、ユーザーが形質予測結果を認知できるものであれば、どのようなものでもよく、テレビジョン、スマホやタブレットの画面、モニタ、音発生装置（例えば、スピーカ）等を用いてもよい。そのような表示は、計算部で予測された計算結果のうち、適宜選択した項目を表示することができる。そのような表示項目としては、患者のがんに最適な抗がん剤の提示、患者の疾患治療における最適な治療方針の提示が挙げられるがこれに限ったものではない。

【００６７】

本開示のシステム１０１の動作の内容について、例示のみを目的とし、図１を参照して説明する。システム１０１は、取得部１０７を有し、当該取得部１０７によって学習に用いるためのデータを取得し、格納部１０２に格納する。学習用データは、既存のデータベース１０８に存在するものを取得（ダウンロード）してもよく、個体の情報を測定する機器を備える測定部１０９から取得してもよい。

【００６８】

システム１０１は、必要に応じて、個体の遺伝情報を画像化する画像化部１０５を備え得る。画像化部が存在する実施形態において、取得した情報をそのまま格納部１０２に格納し、その後に、遺伝情報を画像化部１０５に送信して画像化し、それを再び格納しても

よい。あるいは、取得部１０７で取得された情報を画像化部に送信し、画像化した後に格納部に格納してもよい。システム１０１は必要に応じてこれらの動作を組み合わせて行い得る。すなわち、複数の個体のうちのそれぞれの個体に由来する情報について、必ずしも同一のプロセスによって格納するわけではない。

【００６９】

格納部に格納された複数の個体の遺伝情報および形質情報に基づき、学習部１０３において学習を行い、判別モデルを生成する。生成された判別モデルを用いて、対象の情報（例えば、遺伝情報）に基づいて、計算部１０４において対象の形質情報の予測を行う。予測された結果は、必要に応じて、表示部１０６に表示され得る。システム１０１の動作の間に、任意の時点でデータの保存が行われ得る。

【００７０】

（クラウド、ＩｏＴおよびＡＩを用いた実施形態）

本開示の形質予測技術は、１つのシステム１０１または装置として、すべてを含む形で提供され得る（図１を参照）。あるいは、形質予測装置として、個体の遺伝情報の入力の受け取りおよび結果の表示を主に行い、計算や判別モデルの計算は、サーバやクラウドで行う形態も想定され得る（図２を参照）。これらの一部または全部は、ＩｏＴ（Ｉｎｔｅｒｎｅｔ　ｏｆ　Ｔｈｉｎｇｓ）および／または人工知能（ＡＩ）を用いて実施され得る。あるいは、形質予測装置が判別モデルを格納し、その場で判別を行うが、判別モデルの計算などの主要な計算は、サーバやクラウドで行う形態である半スタンドアローン型の形態も想定され得る（図２）。病院等の一部の実施場所では、送受信が常にできると限らないことから、遮蔽した場合でも使えるモデルを想定したものである。学習部までを備える判別モデル生成用システムも、あるいは得られた判別モデルを保存し計算部において利用する予測システムも、本開示の実施形態として挙げられる（図２）。このようなクラウドサービスとしては、おおむね、「Ｓｏｆｔｗａｒｅ　ａｓ　ｓｅｒｖｉｃｅ　（ＳａａＳ）」が該当する。また、患者データを画像化するプログラムを配布する事で、病院等の実施場所において画像化したデータのみを転送してもらい、それを受信して解析する受託サービス等を提供することも可能である。

【００７１】

表示部は、ユーザーが形質予測結果を認知できるものであれば、どのようなものでもよく、入出力装置、表示装置、テレビジョン、モニタ、音発生装置（例えば、スピーカ）等を用いてもよい。

【００７２】

好ましい実施形態では、判別モデル改善を行う機能が備わっていてもよい。この機能は学習部にあってもよく、別個のモジュールとして備えられてもよい。この判別モデル改善機能は、例えば、オプション１（期間１年、年１～２回）、オプション２（期間１年、１、２ヶ月に１回）、オプション３（期間延長、年１～２回）、オプション４（期間延長＋１、２ヶ月に１回）などのオプションを備えていてもよい。

【００７３】

データ保存も必要に応じてなされ得る。データ保存は通常サーバ側に備えられるが（図２）、全装備型の場合はもとより、クラウド型の場合でも端末側にあってもよい（任意であるため、図では示していない）。クラウドでサービスを提供する場合、データ保存は、標準（例えば、クラウドに１０Ｇバイトまで）、オプション１（例えば、クラウドに１Ｔバイト増量）、オプション２（クラウドにパラメータ設定して分割保存）、オプション３（クラウドに判別モデル別に保存）のオプションを提供し得る。データを保存して、販売されたすべての装置からデータを吸い上げて格納部においてビッグデータを作り、判別モデルを継時的に更新したり、新たなモデルを構築して新たな判別モデルソフトウェアを提供することができる。保存部は、例えば、ＣＤ－Ｒ、ＤＶＤ、Ｂｌｕｅｒａｙ、ＵＳＢ、ＳＳＤ、ハードディスクなどの記録媒体であってもよく、サーバに格納されてもよく、クラウド上に適宜記録される形式でもよい。

【００７４】

また、データ解析オプションを有していてもよい。ここでは、患者のパターン分類（判別精度や特徴量のパターン変化に基づき、患者クラスターを探索する）などを提供することができる。すなわち、計算部１０４の計算方法のオプションとして想定され得る。

【００７５】

図３を参照し、遺伝情報としてＤＮＡデータおよびＲＮＡデータを用いる場合の本開示の判別モデル構築の例をさらに詳細に説明する。この説明は、例示を目的とするものであり、限定の意図を有するものではない。

【００７６】

まず、ＤＮＡのシークエンスデータを読み込む。そして、ＲＮＡ転写量およびエピジェネティック情報を読み込む。これは、システム１０１における取得部１０７を用いて行うことができる。次いで、これらのＤＮＡおよびＲＮＡデータの学習用画像化処理を行う。画像化方法は、本明細書の他の箇所において、図４を参照して詳述される画像化方法を採用することができる。

【００７７】

学習に際して、ＤＰＵマシンスペック（搭載ＧＰＵ数、キャッシュ等）を検出する。当該検出結果に基づいて、学習用画像を領域分割する。分割した画像を、各ノードにおいて学習する。分割学習の詳細は、本明細書の他の箇所において、図６を参照して詳述される分割学習方法を採用することができる。その後、分割学習データを統合する。データの統合にあたって、ＣＰＵマシンスペック（搭載ＣＰＵ数、メモリ等）の検出を行う。統合データを格納できるメモリが存在する場合、非線形最適化処理によって全判別パラメータを最適化し、判別モデルを構築する。統合データを格納できるメモリが存在しない場合、仮想メモリ領域の確保を行い、統合データを一時保存する。その後、Ｏｎ　ｔｈｅ　Ｆｌｙ処理による非線形最適化処理により、全判別パラメータを最適化する。その後分割最適化処理で最適化されたかを判別し、最適化されていない場合には、Ｏｎ　ｔｈｅ　Ｆｌｙ処理による非線形最適化処理を再度行い、再び判別を行う。最適化されたと判別された場合には、判別モデルの構築を終了する。

【００７８】

（画像化方法）

本開示の１つの局面は、遺伝情報を画像化する方法である。１つの態様では、画像化は、それぞれが位置情報および色情報を含む複数のピクセルを有する画像データを生成する工程を含むものとして捉えることができる。この画像データは、遺伝情報のデータを格納しているものであり得る。本開示の画像化方法は、複数の遺伝因子を含む遺伝因子集団の配列データと、複数の遺伝因子を含む遺伝因子集団の発現データとを画像化することを１つの特徴とし得る。このような画像化は、配列情報と、発現情報とを同時に学習することを可能にする点で有利であり得る。加えて、近年の深層学習では従来の機械学習法と比較して、画像の認識性能が格段に向上している事は周知の事実であり、様々な分野に応用されていることから、画像化されたデータであれば、現行の深層学習法を効率的に使用する事が可能となると考えられる。

【００７９】

本開示の１つの態様は、複数の遺伝因子を含む遺伝因子集団の配列データおよび複数の遺伝因子を含む遺伝因子集団の発現データを画像化する方法であって、遺伝因子集団の配列データおよび遺伝因子集団の発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有する、工程を含む、方法である。本開示のさらなる実施形態では、複数の遺伝因子のそれぞれが、画像データ中の領域に対応付けられており、画像データを生成する工程が、遺伝因子の発現量を、当該遺伝因子に対応する領域内の一定領域における色情報および／または当該領域中のある色を有する領域の面積の情報に変換する工程を含み得る。

【００８０】

１つの実施形態においては、発現量に関するデータを画像化する際には、特定数の段階にグルーピングすることができる。実際の遺伝子発現量には遺伝子毎に大きな差が生じて

おり、その発現分布の標準偏差も大きく異なる。従って、発現量データのまま学習させると画像化の必要色が多くなり、遺伝子間での同一値の発現量変化も意味が異なる為、多数（例えば、１０００超）のサンプルのデータから、標準偏差が一定（例えば、１）になるように発現量のスケーリングを行うことができる。さらに、このように変化させた発現量値をグループ化によって粗視化してもよく、これは、機械学習の際に容量削減と学習効率化に有益であり得る。

【００８１】

また、粗視化の際には、グループ化の粗視化単位スケールが細かすぎても粗視化の意味が失われるため、読み込み時のデータで最も標準偏差が小さかった遺伝子（実際に標準偏差が１以下）に対して、単位スケールを徐々に小さく変化させていき、正規分布近似が有効的と判断される範囲で最終的な単位スケールを決定することができる。発現量は、約１２０～約１８０段階、約１３０～約１６０段階、または約１５０段階のグループへとスケーリングしてよい。更に画像としてモノクロ画像を用いてもよい。モノクロ画像の場合、各位置での色情報は、明度の値のみとなり、その段階は特に限定されないが、例えば、明度２５６段階のモノクロ画像を用いることができる。これにより、効率的な容量圧縮を図ることができる。また、ピクセル領域として非常に小さな情報となるＭｕｔａｔｉｏｎ、Ｄｅｌｅｔｉｏｎ、Ｉｎｓｅｒｔｉｏｎの情報を、発現量で用いた差別化（例えば、明度１５０段階での差別化）よりも明度の低い色で表現することによって目立たせ、Ａ、Ｔ、Ｇ、Ｃの塩基もより鮮明に差別化できるよう明度が１０段階異なるもので表現してもよい。この必要明度の段階設定は、本開示の画像化方法に関して、データの圧縮と学習効率化の両面で最適な設定であり、従来技術とは大きく異なる点と考えられる。

【００８２】

また、本開示の１つの実施形態において、画像化は、遺伝子の発現量や変異情報を二次元画像領域の位置と色の明度差を用いて表現する事を目的とし、これにより数値データの場合（約９．６［ＧＢ］）に比べて、ＪＰＧやＰＮＧ等の圧縮画像形式に変換する事によって、情報量を減らすことなく２４分の１（約４００［ＭＢ］）程度まで容量を圧縮する事ができると考えられる。この画像化では、データ容量の圧縮だけではなく、数値データを二次元の位置情報、もしくは色彩情報に変換することで従来法への応用を可能にしたことがも強みと考えられる。

【００８３】

遺伝因子集団の配列データは、親細胞から娘細胞に遺伝形質を伝搬するイベントに関わる因子の配列データを含み得る。このような因子は、例えば、ＤＮＡの配列であり、タンパク質をコードする遺伝子、エキソンの配列、イントロン配列、調節領域配列などが挙げられる。遺伝因子集団の発現データは、当世代のみの情報伝達に関わる因子の発現データを含み得る。このような因子は、例えば、ＲＮＡの発現データであり、ｍＲＮＡ、ｍｉＲＮＡ、ｓｉＲＮＡ、ｌｎＲＮＡの発現量などが挙げられる。

【００８４】

画像化される配列データと発現データは、同一の個体の遺伝因子のものであり得る。

【００８５】

遺伝因子集団の配列データは、ゲノムＤＮＡ上の一定領域の配列を含んでよい。例えば、遺伝因子集団の配列データは、ゲノムＤＮＡ上の遺伝子の配列、ゲノムＤＮＡ上の遺伝子のエクソン配列、および／またはゲノムＤＮＡ上のｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡをコードするＤＮＡ配列を含み得る。

【００８６】

配列情報を画像化する場合には、ある遺伝因子の配列における変異の位置および型の情報を、当該遺伝因子に対応する領域内の位置および色情報に変換することによって行ってよい。すなわち、配列情報の全てを逐一画像に反映させるのではなく、変異を有する部分の情報のみを画像に反映させてよい。これにより、情報量の削減を図ることが可能である。

【００８７】

また、配列上の修飾情報を画像に反映させることが可能である。これは、ある遺伝因子

の配列における修飾の情報を、当該遺伝因子に対応する領域内の位置および色情報に変換する工程によって行ってよい。

【００８８】

発現データは、転写ユニットの発現データを含んでよく、例えば、ｍＲＮＡの発現データ、ｍＲＮＡの発現量、スプライシング、転写開始点、および／またはエピジェネティック修飾のデータを含み得る。遺伝因子集団の発現データは、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、および／または長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡの発現データを含み得る。遺伝因子集団の発現データは、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、および／または長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡの発現量、スプライシング、転写開始点、および／またはエピジェネティック修飾のデータを含み得る。

【００８９】

複数の遺伝因子のそれぞれを、画像データ中の領域に対応付け、遺伝因子の発現量を、当該遺伝因子に対応する領域内の一定領域における色情報および／または該領域中のある色を有する領域の面積の情報に変換することが可能である。

【００９０】

また、遺伝因子がエクソンを含む場合、当該エクソンに対応する転写物またはその一部の発現量を、当該エクソンに対応する領域内の一定領域における色情報および／または当該領域中のある色を有する領域の面積の情報に変換することにより、遺伝因子のスプライシングおよび／または転写開始点を、画像データに格納することができる。

【００９１】

遺伝因子が１または複数の遺伝子を含む場合、１または複数の遺伝子のそれぞれを、画像データ中の領域に対応付け、ある遺伝子のゲノム配列における変異の位置および型の情報を、当該遺伝子に対応する領域内の位置および色情報に変換する工程と、当該遺伝子から転写されるｍＲＮＡの発現量を、当該遺伝子に対応する領域内の一定領域における色情報および／または当該領域中のある色を有する領域の面積の情報に変換する工程とによって、遺伝子の配列および発現情報を画像データに格納することができる。

【００９２】

遺伝因子が１または複数のｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡをコードするＤＮＡ配列を含む場合、１または複数のＤＮＡ配列のそれぞれを、画像データ中の領域に対応付け、あるｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡをコードするＤＮＡ配列のゲノム配列における変異および／またはエピジェネティック修飾の位置および型の情報を、当該遺伝子に対応する領域内の位置および色情報に変換する工程と、当該ＤＮＡ配列から転写されるｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡの発現量、スプライシング、転写開始点、エピジェネティック修飾の情報を、当該遺伝子に対応する領域内の位置および色情報に変換する工程とによって、ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡの配列および発現情報を画像データに格納することができる。

【００９３】

遺伝因子が１または複数の非発現領域のＤＮＡ配列および１または複数の転写ユニットを含む場合、１または複数のＤＮＡ配列および転写ユニットのそれぞれを、画像データ中の領域に対応付け、あるＤＮＡ配列のゲノム配列における変異および／またはエピジェネティック修飾の位置および型の情報を、当該遺伝子に対応する領域内の位置および色情報に変換する工程と、転写ユニットの発現情報を、当該転写ユニットに対応する領域内の一定領域における位置および色情報に変換する工程とによって、非発現領域の配列およびそれに関連する発現情報を画像データに格納することができる。

【００９４】

遺伝因子が１または複数のゲノム上のＤＮＡ領域および転写ユニットを含む場合、１または複数のＤＮＡ領域および転写ユニットのそれぞれを、画像データ中の領域に対応付け、あるＤＮＡ領域のゲノム配列におけるエピジェネティック修飾の位置および型の情報を、当該ＤＮＡ領域に対応する領域内の位置および色情報に変換する工程と、転写ユニットの発現情報を、当該転写ユニットに対応する領域内の一定領域における位置および色情報

に変換する工程とによって、配列およびそれに関連する発現情報を画像データに格納することができる。

【００９５】

本開示の画像化においては、配列情報として、親細胞から娘細胞に遺伝形質を伝搬するイベントに関わる因子であって、核内またはミトコンドリア内に存在し、ＲＮＡポリメラーゼの支配下にあって、タンパク質をコードするｃｏｄｉｎｇ　ＲＮＡまたはｍＲＮＡだけでなく、ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡとして数十塩基までの比較的短鎖のｍｉＲＮＡまたはｓｎｏＲＮＡまたはｓｉＲＮＡまたはｔＲＮＡまたはｒＲＮＡまたはｍｉｔＲＮＡ、さらにはより長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡをコードするＤＮＡ配列を対象とすることができる。さらに、上記の発現産物の相補部分から離れた非発現領域のＤＮＡ配列を対象として、さらには、ＤＮＡ上のエピジェネティック修飾等も含めて対象とすることができる。

【００９６】

発現情報として、遺伝因子（転写ユニット（ＲＮＡおよびｍｉＲＮＡ）の発現量、スプライシング、転写開始点、エピジェネティック修飾等）を含めて、ＲＮＡポリメラーゼの支配下にあって、タンパク質をコードするｃｏｄｉｎｇ　ＲＮＡまたはｍＲＮＡだけでなく、ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡとして数十塩基までの比較的短鎖のｍｉＲＮＡまたはｓｎｏＲＮＡまたはｓｉＲＮＡまたはｔＲＮＡまたはｒＲＮＡまたはｍｉｔＲＮＡ、さらにはより長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡをコードするＤＮＡ配列を対象とすることができる。

【００９７】

これにより、配列に関する包括的な情報と、発現に関する包括的な情報が一枚の画像にまとめられ、機能が同定されていないような領域の変異についても、抗がん剤感受性のような形質と関連付けられる可能性がある。

【００９８】

例えば、ゲノム遺伝子配列とともに、様々なＲＮＡ発現量を発現情報として画像化することで、ある遺伝子の配列情報とその遺伝子の発現量を１つの領域にまとめ、ある遺伝子の配列情報等とその遺伝子の発現量等を同時に処理することができる。

【００９９】

ｍＲＮＡを対象としてみると、他にも、遺伝子の塩基置換として、体細胞変異、胚細胞変異、遺伝子多系、さらにはＡ、Ｔ、Ｇ、Ｃ以外のマイナー塩基への変化（例えば、ナノポアシークエンサーによって測定）を画像に反映させ得る。遺伝子の発現として、発現ユニットとしての遺伝子全体の平均的な発現量だけでなく、スプライシング（この中に、ａｌｔｅｒｎａｔｉｖｅ、ｓｐｌｉｃｅ－ｏｕｔなどがある）、転写開始点の組織・細胞による変化（例えば、ＲＩＫＥＮ　ＦＡＮＴＯＭを用いてこのようなシークエンス情報を得ることができる）を反映させてもよく、エピゲノム、エピトランスクリプトーム修飾として、メチル化Ｃ５、Ａ１、Ａ５、リン酸化なども反映させることができる。

【０１００】

非発現領域については、ＲＮＡへの転写イベントには、ほぼ例外なくクロマチンの開閉が関わるため、免疫沈降－シークエンス法などで、ゲノム全体をプロファイルする、または免疫沈降－ＰＣＲ法で、ターゲットを絞って解析することができる。例えば、ヒストンＨ３第四リジン（Ｈ３Ｋ４）のトリメチルｍｅ３（メチル基が３つ）やジメチルｍｅ２（メチル基が２つ）の修飾は、この付近のクロマチンを開き、その付近への転写因子のリクルートを促進し、転写を活性化する方向に働く。また、Ｈ３Ｋ９のメチル化（ｍｅ３、ｍｅ２）は、クロマチンを閉じて転写を抑制する方向に働く。これらを、免疫沈降－シークエンス法、または免疫沈降－ＰＣＲ法で解析することにより、転写をマップすることができる。このような情報を含めることで、遺伝子と遺伝子との間の領域の転写活性をみることができると考えられる。

【０１０１】

本開示の他の形態では、個体の遺伝因子の配列情報および発現情報から当該個体の形質

情報を予測するモデルを作成するための方法が提供され得る。方法は、複数の個体の遺伝子の配列情報および発現情報を本明細書の他の箇所に記載される方法によって画像化し、画像データを提供する工程と、複数の個体の形質情報を提供する工程と、画像データおよび形質情報から、深層学習により、形質と相関する画像中の特徴表現を抽出する工程とを含み得る。

【０１０２】

画像化のプロセスは、図４を参照してさらに詳細に説明され得るが、この説明は限定を目的としない。画像化処理に際して、遺伝子発現量のスケーリング処理を行う。次いで、各遺伝子領域に応じたメモリを確保する。そして、各遺伝子のデータマトリックスを作成する。そして、スケーリング値に応じてグループ化し、グループ番号をマトリックスの奇数列に代入する。

【０１０３】

Ｍｕｔａｔｉｏｎ（配列置換）の有無を判別し、存在する場合には、変異情報を奇数列の対応位置に代入する。Ｄｅｌｅｔｉｏｎの有無を判別し、存在する場合には、欠損情報を奇数列の対応位置に代入する。Ｉｎｓｅｒｔｉｏｎの有無を判別し、存在する場合には、挿入情報を偶数列の対応位置に代入する。そして、未処理が無ければ、各マトリックスの配置の最適化を行い、画像化処理を行う。配置の最適化については、後述の手順にしたがって行うことができる。画像を書き出し、処理を終了する。

【０１０４】

（配置最適化）

本開示の一部の局面は、画像化において、遺伝因子の配置の最適化を行うことに関する。画像上での遺伝因子の配置は、特に限定されず、例えば、データベースの記載順や、何らかの番号に従って並べてもよい。しかしながら、遺伝子配置を最適化することによって、画像を用いた機械学習効率のさらなる改善が期待できる。したがって、本開示の一部の局面に係る遺伝因子の配置の最適化は、このような改善を目的として応用され得る。とりわけ、外部相関寄与の多い遺伝因子を中心に配置し、相関の重みの大きい順に遺伝因子を周囲に配置していけば、画像を用いた機械学習効率を改善できると考えられる。

【０１０５】

したがって、本開示のこの局面において、遺伝情報を画像化する方法であって、遺伝情報は、複数の遺伝因子を含む遺伝因子集団の配列データおよび／または発現データを含み、当該方法は、遺伝因子集団の配列データおよび／または発現データを格納する画像データを生成する工程であって、該画像データは、それぞれが位置情報および色情報を含む複数のピクセルを有し、当該工程は、当該複数の遺伝因子のそれぞれを、画像データ中の領域に対応付けることを含み、各遺伝因子に対応する領域は、各遺伝因子の相関重みが強いものが近接するように配置されることを特徴とする、工程を含む、方法が提供される。

【０１０６】

画像データを生成する工程は、遺伝因子について必要な画像データ中の領域の面積を算出することを含み得る。必要な領域の面積は、一例として、遺伝因子の配列情報の大きさ（配列長）にしたがって算出してもよい。

【０１０７】

遺伝因子の相関重みは、遺伝因子間の相関解析から強い相関を有する遺伝因子の組み合わせを抽出すること、各遺伝因子についての強い相関遺伝因子を抽出すること、抽出された遺伝因子を用いた変数選択重回帰を行うこと、および／または変数選択重回帰の結果から相関重みを算出することによって算出され得る。

【０１０８】

配置の最適化に関しては、限定を意図するものではないが、図５を参照してさらに詳細に説明する。配置の最適化に際して、遺伝子相関解析を行う（図６参照）。そして、強い相関を持つ遺伝子の組み合わせを抽出する。抽出された遺伝子組み合わせで他の遺伝子との相関が多い順にランキングする。各遺伝子毎に自身の遺伝子と強い相関のある遺伝子を抽出する。前処理した各遺伝子毎に抽出された遺伝子を用いた重回帰（必要変数の選択）

を行う。注目遺伝子からの相関係数βｊｉと対象遺伝子から見た係数βｉｊを抽出し、二乗平均を算出する。ランキングされた遺伝子のトップを中心遺伝子とする。そして、中心遺伝子の必要領域を計算する。中心遺伝子と高相関な遺伝子の必要領域を計算する。次に高相関な遺伝子の必要領域を計算する。遺伝子間相関の二乗平均値を領域間引力係数とし、必要領域に重なりが生じないように最適化する。全遺伝子の配置が完了したかを判別し、完了していない場合には、上記処理を繰り返す。全遺伝子の配置が完了したところで配置最適化処理を終了する。

【０１０９】

遺伝子の相関解析は、図６を参照して、より詳細に例示する。複数の個体（例えば、１０１８の細胞株）の発現データを読み込む。そして、遺伝子相関解析を行う。１対１の相関解析を、ピアソン相関係数：

【数１】

$$\rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\left((\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)\right)^{1/2}}$$

またはスピアマン相関係数：

【数２】

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{n^3 - n}$$

を用いて行う。その後、強い相関遺伝子の組み合わせを抽出する。そして、各遺伝子から見た相関遺伝子を抽出する。この処理で抽出した遺伝子を用いた変数選択重回帰を行う。そして、重回帰の結果から、相関の重みβｊｉとｐ－ｖａｌｕｅを抽出する。相関の重みβｊｉは、

【数３】

$$y_j = \sum_i \beta_{ji} x_i + \varepsilon_j$$

を満たす値として算出され得る。強い相関遺伝子の組み合わせの抽出結果から、最も相関の多い遺伝子を抽出する。そして、この処理で得られた遺伝子を中心に相関重みを抽出する。そして、中心遺伝子と強い相関遺伝子を抽出し、必要領域を計算する。その後、次に強い遺伝子と前出遺伝子との重みを考慮し配置する。全遺伝子を配置したかを判別し、完了していない場合には、上記処理を繰り返す。全遺伝子の配置が完了したところで配置最適化処理を終了する。

【０１１０】

遺伝因子配置は、ＭｉｎＳｕｍ型問題（配置間距離の最小化問題）として最適化することができる。都市内施設配置問題として定式化されているものもあるが、本開示の遺伝因子の配置の最適化は、（１）有効範囲領域（今回の場合は遺伝因子の面積）の末端は接して配置されること、および（２）施設間距離（今回の場合は中心間距離）は必ずしも利用者・重要度（今回の場合は重みと有意性）に比例させるわけではないことによって、施設配置問題とは異なっている。

【０１１１】

（データ構造）

本開示の別の局面において、画像データの、特定のデータ構造に関する。本開示の実施形態において、例えば、複数の遺伝因子を含む遺伝因子集団の配列情報および複数の遺伝因子を含む遺伝因子集団の発現情報を表す画像データのデータ構造であって、画像データは、複数の遺伝因子に対応付けられた複数の領域を有し、遺伝因子の配列中の各位置が、遺伝因子に対応付けられた該領域内の位置に対応付けられており、遺伝因子の配列中の各

位置における置換、欠失および／または挿入の情報が、位置に対応する位置における色情報として格納され、遺伝因子の発現データが、該領域中のある領域における色情報として、および／または該領域中のある色を有する領域の面積の情報として格納されている、データ構造が提供される。

【０１１２】

遺伝因子の配列中の各位置におけるエピジェネティクス修飾の情報も、当該位置に対応する位置における色情報としてさらに格納され得る。例えば、複数の遺伝因子におけるｍｉＲＮＡの配列中の各位置におけるメチル化が、当該位置に対応する位置における色情報として格納され得る。画像データは、行および列を有するマトリックスであってよい。そして、各位置は、行および列の組み合わせとして格納され得る。

【０１１３】

配列情報は、ゲノム上の領域のＤＮＡ配列を含み得る。ゲノム上の領域としては、例えば、遺伝子、エクソン、イントロン、非発現領域、および／またはｎｏｎ－ｃｏｄｉｎｇＲＮＡをコードする領域が挙げられる。

【０１１４】

発現情報としては、ｍＲＮＡ、ｍｉＲＮＡ、ｓｎｏＲＮＡ、ｓｉＲＮＡ、ｔＲＮＡ、ｒＲＮＡ、ｍｉｔＲＮＡ、および／または長鎖ｎｏｎ－ｃｏｄｉｎｇ　ＲＮＡからなる群から選択される転写ユニットの発現量、スプライシング、転写開始点、および／またはエピジェネティック修飾の情報を含み得る。

【０１１５】

画像データは、各ゲノム上の領域および／または転写ユニットに対応付けられた複数の領域を有し得る。ゲノム上の領域に対応付けられた領域は、当該ゲノム上の領域の長さに依存した数の列および一定数の行からなるものであり得る。ゲノム上の領域の配列中の各位置は、ゲノム上の領域に対応付けられた領域内の奇数列における位置に対応付けられ得る。ゲノム上の領域の配列中の各位置における置換、欠失および／または挿入の情報は、当該位置に対応する奇数列における位置における色情報として格納され得る。色情報は、変異が存在しないことを示す色情報、Ａに置換されていることを示す色情報、Ｔに置換されていることを示す色情報、Ｇに置換されていることを示す色情報、Ｃに置換されていることを示す色情報、欠失していることを示す色情報、または当該位置に隣接して挿入が存在することを示す色情報であり得る。挿入される配列の情報は、挿入が存在することを示す色情報を有する位置に隣接する偶数列における位置を始点として、挿入される配列を示す色情報として格納されてよい。

【０１１６】

ゲノム上の領域の配列中の各位置におけるエピジェネティック修飾の情報は、当該位置に対応する奇数列における位置における色情報として格納され得る。当該色情報は、エピジェネティック修飾が存在しないことを示す色情報、ＤＮＡメチル化されていることを示す色情報、ヒストンメチル化されていることを示す色情報、ヒストンアセチル化されていることを示す色情報、ヒストンユビキチン化されていることを示す色情報、またはヒストンリン酸化されていることを示す色情報などを含み得る。

【０１１７】

あるゲノム上の領域から転写される転写ユニットについて、当該転写ユニットの発現量が、当該ゲノム上の領域に対応する画像中の領域における色の濃淡として、および／または当該領域中のある色を有する領域の面積の情報として格納され得る。

【０１１８】

また、遺伝子であるゲノム上の領域について、当該遺伝子に対応するｍＲＮＡの発現量が、当該領域中のある領域における色の濃淡として、および／または当該領域中のある色を有する領域の面積の情報として格納され得る。

【０１１９】

上述の画像化方法および画像データは、個体の遺伝情報を包括的に扱う上で有用であり、例えば、医療、農業、畜産、食品、環境、薬学（創薬、育薬の分野）の分野など、生物

が関係する任意の技術分野において有用である。
【０１２０】
（分割学習）
本開示の別の局面において、画像と、当該画像に対応する情報との関連を予測するモデルを作成するための方法が提供される。方法は、画像を分割して学習することを１つの特徴とし得る。方法は、複数の画像および該複数の画像に対応する複数の情報のセットを提供する工程と、複数の画像を分割し、複数の画像の部分と、当該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、複数の分割学習データを統合し、画像と、画像に対応する情報との関連を予測するモデルを生成する工程とを含み得る。
【０１２１】
統合する工程は、ＧＰＵを搭載したＣＰＵマシンを用い、メモリ搭載量を含めたＧＰＵスペックおよびＣＰＵスペックを検出することを含み得る。統合する工程は、ＨＤＤ上でのＲｅａｄ－Ｗｒｉｔｅファイルの利用、ＣＰＵメモリを最大限利用できるような非線形最適化処理アルゴリズムを最適化することを含み得る。
【０１２２】
非線形最適化処理アルゴリズムは、必要なデータを随時メモリに移して計算し、計算結果をＨＤＤに戻すことによって、データサイズに非依存的に計算可能なアルゴリズムであり得る（Ｏｎ　ｔｈｅ　Ｆｌｙなメモリ処理）。非線形最適化処理は、全判別パラメータを最適化することを含み得る。
【０１２３】
分割画像の学習について、限定を意図するものではないが、図７を参照してさらに例示的に詳述する。機械学習は、Ｄｅｅｐ　Ｌｅａｒｎｉｎｇ処理によって行うことができる。機械学習に際して、学習データ、教師データ、検証データを分割する。乱数処理による判別パターン係数の決定と、全判別パターンの計算を行う。出力される誤差を計算する。全体の誤差が最小になるように判別パターン係数（重み）を最適化する。追加学習の有無を判別する。追加学習が必要である場合は、上記の処理を繰り返す。追加学習が必要ない場合は機械学習を終了する。
【０１２４】
分割学習データの統合を含めた学習の流れについて、限定を意図するものではないが、図８を参照してさらに例示的に詳述する。学習用の画像データを読み込む。搭載ＧＰＵ数を検出し、分割数を決定する。学習データの画像を分割する。ＧＰＵ処理部において、ＧＰＵ単位で画像部位を別に学習する。学習におけるそれぞれのノードは、物理的に分離されていてもよく、一体となっていてもよい。分割学習データの統合を行う。搭載ＣＰＵ数とメモリ確保可能領域を検出する。十分なメモリが搭載されている場合、非線形最適化を行い処理を終了する。十分なメモリが搭載されていない場合、計算に必要なデータをＨＤＤに一時保存し、メモリ搭載可能な分だけを読み込む。メモリ格納部分の非線形最適化を行う。最適かを判別し、最適でない場合は処理を繰り返す。最適であると判別された場合には処理を終了する。
【０１２５】
上述の分割学習の方法は、比較的容量の大きいデータ（例えば、画像データ）を用いた機械学習における効率を向上させる。例えば、生物の情報を画像化したものの学習のほか、物理学・天文学のようなデータ量が多い分野での学習、物体認識、および文字認識等における学習において有用である。
【０１２６】
分割学習において、各分割学習データの判別能力を検証してもよい。画像について言えば、画像を分割した各領域ごとに、形質情報などの目的変数との相関を検証してよい。判別能力および／または相関の検証は、各領域と目的変数との関係を機械学習に供し、Ｅｐｏｃｈ数を増加させていく際に、予測能力が収束するかを判定することによって行い得る。各分割学習データの中から、判別力のある分割学習データを選択してその後統合し、全体の学習の効率化を図り得る。あるいは、各分割学習データの中から、判別力のある分割

学習データを選択し、それ自体を予測モデルとして使用し得る。
【０１２７】
分割の程度について、全体のサイズに鑑みて調整することができる。遺伝子変異情報および発現情報を画像化した画像を用いる場合には、例えば、１領域あたりに約１００～約２００程度の遺伝子の情報が格納されるようなサイズに分割することができる。
【０１２８】
システムとしては、個体の形質情報を予測するためのシステムであって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを格納する格納部であって、該遺伝情報は、遺伝因子の配列情報および発現情報を含む、格納部と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を、該複数の個体の遺伝情報を画像化して学習するように構成されている、学習部と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、該個体の形質情報を予測する、計算部と
を備え、
ここで、該学習部が、該複数の個体の遺伝情報を画像化して生成した画像を分割して、画像の各領域と形質情報との関連を学習し、各領域から形質情報の判別能力を有するモデルを生成可能な領域を選択して、画像の各領域から形質情報を予測するモデルを生成するように構成されている、システムとして提供され得る。
【０１２９】
方法としては、個体の遺伝因子の配列情報および発現情報を含む遺伝情報と、該個体の形質情報との関連を予測するモデルを作成するための方法であって、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データから、形質情報の判別能力を有する分割学習データを選択し、画像の各領域から形質情報を予測するモデルを生成する工程と
を含む、方法として提供され得る。
【０１３０】
本開示はまた、個体の遺伝因子の配列情報および発現情報を含む遺伝情報と、該個体の形質情報との関連を予測するモデルを作成するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
該複数の分割学習データから、形質情報の判別能力を有する分割学習データを選択し、画像の各領域から形質情報を予測するモデルを生成する工程と
を含む、プログラムを提供する。
【０１３１】
画像を遺伝因子の配列情報および発現情報を含む遺伝情報から生成している場合には、形質情報の判別能力を有する分割学習データを得られる画像の部分を選択し、形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択してもよい。これにより、形質と相関する遺伝子またはその変異を同定する方法として使用し得る。発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子には、形質情報と相関する変異を有する遺伝子が特定され得、このような遺伝子またはその変異は、機能的に形質と相関している可能性がある。特定された遺伝子は、個体の形質情報の予測に使用可能であると考えられる。特定された遺伝子は、それ自体が個体の形質情報を予測するモデルとなり得、また、必要に応じて、個体の形質情報を予測するモデルに組み込んで

使用され得る。
【０１３２】
ある領域について、発現情報に基づいて形質情報が予測可能かの判定は、例えば、ある領域に含まれる遺伝子の各個体での発現量をクラスタリング分析することによって行い得る。クラスタリング分析の他、任意の回帰分析または機械学習の手法を用いて判定してもよい。
【０１３３】
システムとしては、個体の形質情報を予測するためのシステムであって、
複数の個体の遺伝情報と、該複数の個体の形質情報とを格納する格納部であって、該遺伝情報は、遺伝因子の配列情報および発現情報を含む、格納部と、
該複数の個体の遺伝情報と、該複数の個体の形質情報とから、遺伝情報と形質情報との関連を、該複数の個体の遺伝情報を画像化して学習するように構成されている、学習部と、
該遺伝情報と形質情報との関連に基づき、個体の遺伝情報から、該個体の形質情報を予測する、計算部と
を備え、
ここで、該学習部が、該複数の個体の遺伝情報を画像化して生成した画像を分割して、画像の各領域と形質情報との関連を学習し、各領域から形質情報の判別能力を有するモデルを生成可能な領域を選択し、各領域において、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない領域における遺伝子から、形質情報と相関する変異を有する遺伝子を特定するように構成され、
該計算部は、該形質情報と相関する変異を有する遺伝子の情報に基づいて該個体の形質情報を予測するように構成されている、システムとして提供され得る。
【０１３４】
方法としては、形質に関与する遺伝子の変異を同定するための方法であって、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、
該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、方法として提供され得る。
【０１３５】
収束性があり、遺伝子発現量のみで分離できる場合でも、その特定領域の画像をさらに細かく分割することで、判別に重要となる遺伝子を抽出できる。分割画像領域においても収束性があり、遺伝子発現量情報のみで判別できる領域が、判別に重要な遺伝子情報である。従って、分割を繰り返すことで遺伝子情報を抽出することができる。
【０１３６】
収束性があるにもかかわらず、遺伝子発現量のみでは分離できない場合でも、
その特定領域の画像をさらに細かく分割することで判別に重要となる遺伝子変異情報を抽出することが可能である。そこでも収束性があるにもかかわらず、遺伝子発現量情報のみでは分割できない領域を絞り込み、絞り込んだ領域に含まれる遺伝子変異情報を抽出する。
【０１３７】
本開示はまた、形質に関与する遺伝子の変異を同定するための方法をコンピュータに実行させるプログラムであって、該方法は、
複数の個体の遺伝因子の配列情報および発現情報を画像化した複数の画像および該複数の画像に対応する複数の形質情報のセットを提供する工程と、

該複数の画像を分割し、該複数の画像の部分と、該画像に対応する情報との関連を学習し、複数の分割学習データを得る工程と、
形質情報の判別能力を有する分割学習データを得られる画像の部分を選択する工程と、
該形質情報の判別能力を有する分割学習データを得られる画像の部分から、発現情報に基づいて形質情報が予測可能かを判定し、発現情報に基づいて形質情報が予測可能でない部分を選択する工程と、
該発現情報に基づいて形質情報が予測可能でない部分に含まれる遺伝子から、形質情報と相関する変異を有する遺伝子を特定する工程と
を含む、プログラムも提供する。

【０１３８】

（他の実施の形態）

以上、本開示の１つまたは複数の態様に係る形質予測方法について、実施の形態に基づいて説明したが、本開示は、この実施の形態に限定されるものではない。本開示の趣旨を逸脱しない限り、当業者が思いつく各種変形を本実施の形態に施したものや、異なる実施の形態における構成要素を組み合わせて構築される形態も、本開示の１つまたは複数の態様の範囲内に含まれてもよい。

【０１３９】

形質予測方法は、プログラムによって実行され得る。すなわち、個体の形質情報を予測するための方法をコンピュータに実行させるプログラムであって、当該方法は、複数の個体の遺伝情報と、当該複数の個体の形質情報とを提供する情報提供工程であって、当該遺伝情報は、少なくとも２種類の情報を含む、工程と、当該複数の個体の遺伝情報と、当該複数の個体の形質情報とから、遺伝情報と形質情報との関連を学習させる学習工程と、当該遺伝情報と形質情報との関連に基づき、個体の遺伝子情報から、当該個体の形質情報を予測する予測工程とを含む、プログラムが提供され得る。プログラムにおいて、前記予測された形質情報を表示する表示工程がさらに含まれ得る。このようなプログラムを格納した記録媒体もまた提供され得る。

【０１４０】

システムは、本明細書に記載される方法をコンピュータに実行させるためのプログラムを備えていてよく、例えば、そのようなプログラムを格納した記録媒体を備え得る。また、プログラムによって指示される命令を実行するための計算装置（例えば、コンピュータ）を備えていてよい。計算装置は、物理的に一体としていても、あるいは、物理的に分離した複数の構成要素からなっていてもよい。これらの計算装置の内部において、本開示における画像化部１０５、学習部１０３、計算部１０４および取得部１０７等に対応する機能が必要に応じて備えられ得る。

【０１４１】

本開示のシステムは、複数の構成部を１個のチップ上に集積して製造された超多機能ＬＳＩとして実現することができ、具体的には、マイクロプロセッサ、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）などを含んで構成されるコンピュータシステムであり得る。ＲＯＭには、コンピュータプログラムが記憶されている。前記マイクロプロセッサが、コンピュータプログラムに従って動作することにより、システムＬＳＩは、その機能を達成する。

【０１４２】

なお、ここでは、システムＬＳＩとしたが、集積度の違いにより、ＩＣ、ＬＳＩ、スーパーＬＳＩ、ウルトラＬＳＩと呼称されることもある。また、集積回路化の手法はＬＳＩに限るものではなく、専用回路または汎用プロセッサで実現してもよい。ＬＳＩ製造後に、プログラムすることが可能なＦＰＧＡ（Ｆｉｅｌｄ Ｐｒｏｇｒａｍｍａｂｌｅ Ｇａｔｅ Ａｒｒａｙ）、あるいはＬＳＩ内部の回路セルの接続や設定を再構成可能なリコンフィギュラブル・プロセッサを利用してもよい。

【０１４３】

さらには、半導体技術の進歩または派生する別技術によりＬＳＩに置き換わる集積回路

化の技術が登場すれば、当然、その技術を用いて機能ブロックの集積化を行ってもよい。バイオ技術の適用等が可能性としてありえる。

【０１４４】

また、本開示の一態様は、このような画像化分析、診断、治療、予防予測装置だけではなく、検査分析・診断・治療予測装置に含まれる特徴的な構成部をステップとする検査分析・診断・治療予測方法であってもよい。また、本開示の一態様は、検査分析・診断・治療予測方法に含まれる特徴的な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であってもよい。また、本開示の一態様は、そのようなコンピュータプログラムが記録された、コンピュータ読み取り可能な非一時的な記録媒体であってもよい。

【０１４５】

なお、上記各実施の形態において、各構成要素は、専用のハードウェアで構成されるか、各構成要素に適したソフトウェアプログラムを実行することによって実現されてもよい。各構成要素は、ＣＰＵまたはプロセッサなどのプログラム実行部が、ハードディスクまたは半導体メモリなどの記録媒体に記録されたソフトウェアプログラムを読み出して実行することによって実現されてもよい。ここで、上記各実施の形態の痛み推定装置などを実現するソフトウェアは、本明細書において上述したプログラムであり得る。

【０１４６】

本明細書において「または」は、文章中に列挙されている事項の「少なくとも１つ以上」を採用できるときに使用される。「もしくは」も同様である。本明細書において「２つの値の範囲内」と明記した場合、その範囲には２つの値自体も含む。

【０１４７】

本明細書において引用された、科学文献、特許、特許出願などの参考文献は、その全体が、各々具体的に記載されたのと同じ程度に本明細書において参考として援用される。

【０１４８】

以上、本開示を、理解の容易のために好ましい実施形態を示して説明してきた。以下に、実施例に基づいて本開示を説明するが、上述の説明および以下の実施例は、例示の目的のみに提供され、本開示を限定する目的で提供したのではない。従って、本開示の範囲は、本明細書に具体的に記載された実施形態にも実施例にも限定されず、特許請求の範囲によってのみ限定される。

【実施例】

【０１４９】

以下に実施例を示す。

（実施例１）ＤＮＡとＲＮＡとを用いたＡＩによる解析

本実施例においては、以下：

（１）データ取得（トランスクリプトームデータ、ゲノム配列データ、変異データ、ゲノムエピジェネティクスデータ、ｍｉＲＮＡ発現データ、ＲＮＡメチル化データ）；

（２）画像化；

（３）画像をＧＰＵとＣＰＵの両方を搭載したマシンで学習；

（４）別画像を用いて抗がん剤への感受性予測

の工程によるＡＩ解析を実証する。

【０１５０】

（３）の学習工程は、プログラム上では、ＧＰＵ数、ＧＰＵ搭載メモリおよび、ＣＰＵ数、ＣＰＵ用メモリを検出し、画像の分割学習と予測統合に関して実施できるようにする。

【０１５１】

（実施例１－１）前処理について

（データ取得）

以下に示す細胞株についての網羅的解析データを取得した：

【表１－１】

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 201T | 22RV1 | 23132-87 | 42-MG-BA | 451Lu | 5637 | 639-V | 647-V | 697 | 769-P |
| 786-0 | 8-MG- | 8305C | 8505C | A101D | A172 | A204 | A2058 | A253 | A2780 |
| A3-KAW | A375 | A388 | A4-Fuk | A427 | A431 | A498 | A549 | A673 | A704 |
| ABC-1 | ACHN | ACN | AGS | ALL-PO | ALL-SIL | AM-38 | AMO-1 | AN3-CA | ARH-77 |
| ASH-3 | ATN-1 | AU565 | AsPC-1 | BALL-1 | BB30-HNC | BB49-HNC | BB65-RCC | BC-1 | BC-2 |
| BC-3 | BCPAP | BE-13 | BE2- | BEN | BFTC- | BFTC- | BHT-101 | BHY | BICR10 |
| BICR22 | BICR31 | BICR78 | BL-41 | BL-70 | BOKU | BPH-1 | BT-20 | BT-474 | BT-483 |
| BT-549 | BV-173 | Becker | BxPC-3 | C-33-A | C-4-I | C2BBe1 | C32 | C3A | CA46 |
| CADO-ES1 | CAKI-1 | CAL-120 | CAL-12T | CAL-148 | CAL-27 | CAL-29 | CAL-33 | CAL-39 | CAL-51 |
| CAL-54 | CAL-62 | CAL-72 | CAL-78 | CAL-85-1 | CAMA-1 | CAPAN-1 | CAS-1 | CCF-STTG1 | CCK-81 |
| CCRF-CEM | CESS | CFPAC-1 | CGTH-W-1 | CHL-1 | CHP-126 | CHP-134 | CHP-212 | CHSA0011 | CHSA0108 |
| CHSA8926 | CL-11 | CL-34 | CL-40 | CMK | CML-T1 | COLO-205 | COLO-320-HSR | COLO-668 | COLO-678 |
| COLO-679 | COLO-680N | COLO-684 | COLO-741 | COLO-783 | COLO-792 | COLO-800 | COLO-824 | COLO-829 | COR-L105 |
| COR-L23 | COR-L279 | COR-L303 | COR-L311 | COR-L32 | COR-L321 | COR-L88 | COR-L95 | CP50-MEL-B | CP66-MEL |
| CP67-MEL | CPC-N | CRO-AP2 | CRO-AP3 | CS1 | CTB-1 | CTV-1 | CW-2 | Ca-Ski | Ca9-22 |
| CaR-1 | Calu-1 | Calu-3 | Calu-6 | Caov-3 | Caov-4 | Capan-2 | ChaGo-K-1 | D-247MG | D-263MG |
| D-283MED | D-336MG | D-392MG | D-423MG | D-502MG | D-542MG | D-566MG | DAN-G | DB | DBTRG-05MG |
| DEL | DG-75 | DIFI | DJM-1 | DK-MG | DMS-114 | DMS-153 | DMS-273 | DMS-53 | DMS-79 |
| DND-41 | DOHH-2 | DOK | DOV13 | DSH1 | DU-145 | DU-4475 | DV-90 | Daoy | Daudi |
| Detroit562 | DoTc2-4510 | EB-3 | EB2 | EBC-1 | EC-GI-10 | ECC10 | ECC12 | ECC4 | EFE-184 |
| EFM-19 | EFM-192A | EFO-21 | EFO-27 | EGI-1 | EHEB | EJM | EKVX | EM-2 | EMC-BAC-1 |
| EMC-BAC-2 | EN | EPLC-272H | ES-2 | ES1 | ES3 | ES4 | ES5 | ES6 | ES7 |
| ES8 | ESO26 | ESO51 | ESS-1 | ETK-1 | EVSA-T | EW-1 | EW-11 | EW-12 | EW-13 |
| EW-16 | EW-18 | EW-22 | EW-24 | EW-3 | EW-7 | EW7476 | EoL-1- | FADU | FLO-1 |
| FTC-133 | FU-OV-1 | FU97 | Farage | G-292-Clone-A141B1 | G-361 | G-401 | G-402 | G-MEL | GA-10 |
| GAK | GAMG | GB-1 | GCIY | GCT | GDM-1 | GI-1 | GI-ME-N | GMS-10 | GOTO |
| GP5d | GR-ST | GRANT A-519 | GT3TKB | H-EMC-SS | H2369 | H2373 | H2461 | H2591 | H2595 |
| H2596 | H2722 | H2731 | H2795 | H2803 | H2804 | H2810 | H2818 | H2869 | H290 |
| H3118 | H3255 | H4 | H513 | H9 | HA7- | HAL-01 | HARA | HC-1 | HCC-15 |
| HCC-33 | HCC-366 | HCC-44 | HCC-56 | HCC-78 | HCC-827 | HCC114 | HCC118 | HCC139 | HCC141 |
| HCC142 | HCC150 | HCC156 | HCC159 | HCC180 | HCC193 | HCC195 | HCC202 | HCC215 | HCC221 |
| HCC299 | HCC38 | HCC70 | HCE-4 | HCT-116 | HCT-15 | HD-MY- | HDLM-2 | HDQ-P1 | HEC-1 |
| HEL | HGC-27 | HH | HL-60 | HLE | HMV-II | HN | HO-1-N- | HO-1-u-1 | HOP-62 |
| HOP-92 | HOS | HPAC | HPAF-II | HSC-2 | HSC-3 | HSC-39 | HSC-4 | HT | HT-1080 |
| HT-115 | HT-1197 | HT-1376 | HT-144 | HT-29 | HT-3 | HT55 | HTC-C3 | HUH-6-clone5 | HUTU-80 |
| HeLa | Hep3B2-1-7 | Hey | Hs-445 | Hs-578-T | Hs-633T | Hs-683 | Hs-766T | Hs-939-T | Hs-940-T |
| Hs746T | HuCCT1 | HuH-7 | HuO- | HuO9 | HuP-T3 | HuP-T4 | IA-LM | IGR-1 | IGR-37 |
| IGROV-1 | IHH-4 | IM-9 | IM-95 | IMR-5 | IOSE-364(-) | IOSE-397 | IOSE-523(-) | IOSE-75-16SV40 | IPC-298 |

【表１－２】

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| IST-MEL1 | IST-MES1 | IST-SL1 | IST-SL2 | Ishikawa (Heraklio)02ER- | J-RT3-T3-5 | J82 | JAR | JEG-3 | JEKO-1 |
| JHH-1 | JHH-2 | JHH-4 | JHH-6 | JHH-7 | JHOS-2 | JHOS-3 | JHOS-4 | JHU-011 | JHU-013 |
| JHU-019 | JHU-022 | JHU-028 | JHU-029 | JIMT-1 | JJN-3 | JM1 | JSC-1 | JURL-MK1 | JVM-2 |
| JVM-3 | JiyoyeP-2003 | Jurkat | K-562 | K052 | K1 | K19 | K2 | K4 | K5 |
| K8 | KALS-1 | KARPAS-1106P | KARPAS-231 | KARPAS-299 | KARPAS-422 | KARPAS-45 | KARPAS-620 | KASUMI-1 | KATOIII |
| KCL-22 | KE-37 | KELLY | KG-1 | KG-1-C | KGN | KINGS-1 | KLE | KM-H2 | KM12 |
| KMH-2 | KMOE-2 | KMRC-1 | KMRC-20 | KMS-11 | KMS-12-BM | KNS-42 | KNS-62 | KNS-81-FD | KON |
| KOPN-8 | KOSC-2 | KP-1N | KP-2 | KP-3 | KP-4 | KP-N-RT-BM-1 | KP-N-YN | KP-N-YS | KS-1 |
| KU-19-19 | KU812 | KURAMOCHI | KY821 | KYAE-1 | KYM-1 | KYSE-140 | KYSE-150 | KYSE-180 | KYSE-220 |
| KYSE-270 | KYSE-30 | KYSE-410 | KYSE-450 | KYSE-50 | KYSE-510 | KYSE-520 | KYSE-70 | Kasumi-3 | L-1236 |
| L-363 | L-428 | L-540 | LAMA-84 | LAN-6 | LB1047-RCC | LB2241-RCC | LB2518-MEL | LB373-MEL-D | LB647-SCLC |
| LB771-HNC | LB831-BLC | LB996-RCC | LC-1-sq | LC-1F | LC-2-ad | LC4-1 | LCLC-103H | LCLC-97TM1 | LIM1215 |
| LK-2 | LN-18 | LN-229 | LN-405 | LNCaP-Clone-FGC | LNZTA3WT4 | LOU-NH91 | LOUCY | LOXIMVI | LP-1 |
| LS-1034 | LS-123 | LS-180 | LS-411N | LS-513 | LU-134- | LU-135 | LU-139 | LU-165 | LU-65 |
| LU-99A | LXF-289 | LoVo | M059J | M14 | MB157 | MC-1010 | MC-CAR | MC-IXC | MC116 |
| MCAS | MCC13 | MCC26 | MCF7 | MDA-MB-134- | MDA-MB-157 | MDA-MB-175- | MDA-MB-231 | MDA-MB-330 | MDA-MB-361 |
| MDA-MB-415 | MDA-MB-436 | MDA-MB-453 | MDA-MB-468 | MDST8 | ME-1 | ME-180 | MEC-1 | MEG-01 | MEL-HO |
| MEL-JUSO | MES-SA | MFE-280 | MFE-296 | MFE-319 | MFH-ino | MFM-223 | MG-63 | MHH-CALL-2 | MHH-ES-1 |
| MHH-NB-11 | MHH-PREB-1 | MIA-PaCa-2 | MKL-1-subclone-2 | MKL-2 | MKN1 | MKN28 | MKN45 | MKN7 | ML-1 |
| ML-2 | MLMA | MM1S | MMAC-SF | MN-60 | MOG-G-CCM | MOG-G-UVW | MOLM-13 | MOLM-16 | MOLP-8 |
| MOLT-13 | MOLT-16 | MOLT-4 | MONO-MAC-6 | MPP-89 | MRK-nu-1 | MS-1 | MS751 | MSTO-211H | MV-4-11 |
| MY-M12 | MZ1-PC | MZ2-MEL | MZ7-mel | Mewo | Mo-T | NALM-6 | NAMALWA | NB(TU)1-10 | NB1 |
| NB10 | NB12 | NB13 | NB14 | NB17 | NB4 | NB5 | NB6 | NB69 | NB7 |
| NBsusSR | NCC010 | NCC021 | NCI-H1048 | NCI-H1092 | NCI-H1105 | NCI-H1155 | NCI-H1184 | NCI-H128 | NCI-H1299 |
| NCI-H1304 | NCI-H1341 | NCI-H1355 | NCI-H1385 | NCI-H1395 | NCI-H1404 | NCI-H1417 | NCI-H1435 | NCI-H1436 | NCI-H1437 |
| NCI-H146 | NCI-H1522 | NCI-H1563 | NCI-H1568 | NCI-H1573 | NCI-H1581 | NCI-H1618 | NCI-H1623 | NCI-H1648 | NCI-H1650 |
| NCI-H1651 | NCI-H1666 | NCI-H1688 | NCI-H1693 | NCI-H1694 | NCI-H1703 | NCI-H1734 | NCI-H1755 | NCI-H1770 | NCI-H1781 |
| NCI-H1792 | NCI-H1793 | NCI-H1836 | NCI-H1838 | NCI-H1869 | NCI-H187 | NCI-H1876 | NCI-H1915 | NCI-H1926 | NCI-H1944 |
| NCI-H196 | NCI-H1963 | NCI-H1975 | NCI-H1993 | NCI-H2009 | NCI-H2023 | NCI-H2029 | NCI-H2030 | NCI-H2052 | NCI-H2066 |
| NCI-H2081 | NCI-H2085 | NCI-H2087 | NCI-H209 | NCI-H2107 | NCI-H211 | NCI-H2110 | NCI-H2122 | NCI-H2126 | NCI-H2135 |
| NCI-H2141 | NCI-H2170 | NCI-H2171 | NCI-H2172 | NCI-H2196 | NCI-H220 | NCI-H2227 | NCI-H2228 | NCI-H226 | NCI-H2286 |

【表１－３】

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| NCI-H2291 | NCI-H23 | NCI-H2342 | NCI-H2347 | NCI-H2405 | NCI-H2444 | NCI-H2452 | NCI-H250 | NCI-H28 | NCI-H292 |
| NCI-H3122 | NCI-H322M | NCI-H345 | NCI-H358 | NCI-H378 | NCI-H441 | NCI-H446 | NCI-H460 | NCI-H508 | NCI-H510A |
| NCI-H520 | NCI-H522 | NCI-H524 | NCI-H526 | NCI-H596 | NCI-H630 | NCI-H64 | NCI-H647 | NCI-H650 | NCI-H661 |
| NCI-H69 | NCI-H716 | NCI-H719 | NCI-H720 | NCI-H727 | NCI-H735 | NCI-H747 | NCI-H748 | NCI-H810 | NCI-H82 |
| NCI-H820 | NCI-H835 | NCI-H838 | NCI-H841 | NCI-H847 | NCI-H865 | NCI-H929 | NCI-N87 | NCI-SNU-1 | NCI-SNU-16 |
| NCI-SNU-5 | NEC8 | NH-12 | NK-92MI | NKM-1 | NMC-G1 | NOMO-1 | NOS-1 | NTERA-S-cl-D1 | NU-DUL-1 |
| NUGC-3 | NUGC-4 | NY | OACM5-1 | OACp4C | OAW-28 | OAW-42 | OC-314 | OCI-AML2 | OCI-AML3 |
| OCI-AML5 | OCI-LY-19 | OCI-LY7 | OCI-M1 | OCUB-M | OCUM-1 | OE19 | OE21 | OE33 | OMC-1 |
| ONS-76 | OPM-2 | OS-RC-2 | OSA-80 | OSC-19 | OSC-20 | OUMS- | OV-17R | OV-56 | OV-7 |
| OV-90 | OVCA42 | OVCA43 | OVCA43 | OVCAR- | OVCAR- | OVCAR- | OVCAR- | OVISE | OVK-18 |
| OVKATE | OVMIU | OVTOKO | P116 | P12-ICHIKAWA | P30-OHK | P31-FUJ | P32-ISH | P3HR-1 | PA-1 |
| PA-TU-8902 | PA-TU-8988T | PANC-02-03 | PANC-03-27 | PANC-04-03 | PANC-08-13 | PANC-10-05 | PC-14 | PC-3 | PC-3 [JPC- |
| PCI-15A | PCI-30 | PCI-38 | PCI-4B | PCI-6A | PE-CA-PJ15 | PEO1 | PF-382 | PFSK-1 | PL-21 |
| PL18 | PL4 | PLC-PRF-5 | PSN1 | PWR-1E | QGP-1 | QIMR-WIL | RC-K8 | RCC-AB | RCC-ER |
| RCC-FG2 | RCC-JF | RCC-JW | RCC-MF | RCC10RGB | RCH-ACV | RCM-1 | RD | RD-ES | REH |
| RERF-GC-1B | RERF-LC-KJ | RERF-LC-MS | RERF-LC-Sq1 | RF-48 | RH-1 | RH-18 | RH-41 | RKN | RKO |
| RL | RL95-2 | RMG-I | RO82-W-1 | ROS-50 | RPMI-2650 | RPMI-6666 | RPMI-7951 | RPMI-8226 | RPMI-8402 |
| RPMI-8866 | RS4-11 | RT-112 | RT4 | RVH-421 | RXF393 | Raji | Ramos-2G6- | S-117 | SAS |
| SAT | SBC-1 | SBC-3 | SBC-5 | SCC-15 | SCC-25 | SCC-3 | SCC-4 | SCC-9 | SCC90 |
| SCH | SCLC-21H | SCaBER | SF126 | SF268 | SF295 | SF539 | SH-4 | SHP-77 | SIG-M5 |
| SIMA | SISO | SJRH30 | SJSA-1 | SK-CO-1 | SK-ES-1 | SK-GT-2 | SK-GT-4 | SK-HEP-1 | SK-LMS-1 |
| SK-LU-1 | SK-MEL-1 | SK-MEL-2 | SK-MEL-24 | SK-MEL-28 | SK-MEL-3 | SK-MEL-30 | SK-MEL-31 | SK-MEL-5 | SK-MES-1 |
| SK-MG-1 | SK-MM-2 | SK-N-AS | SK-N-DZ | SK-N-FI | SK-N-SH | SK-NEP-1 | SK-OV-3 | SK-PN-DW | SK-UT-1 |
| SKG-IIIa | SKM-1 | SKN | SKN-3 | SLVL | SN12C | SNB75 | SNG-M | SNU-1040 | SNU-175 |
| SNU-182 | SNU-387 | SNU-398 | SNU-407 | SNU-423 | SNU-449 | SNU-475 | SNU-61 | SNU-81 | SNU-C1 |
| SNU-C2B | SNU-C5 | SR | ST486 | STS-0421 | SU-DHL-1 | SU-DHL-10 | SU-DHL-16 | SU-DHL-4 | SU-DHL-5 |
| SU-DHL-6 | SU-DHL-8 | SU8686 | SUIT-2 | SUP-B15 | SUP-B8 | SUP-HD1 | SUP-M2 | SUP-T1 | SW1088 |
| SW1116 | SW1271 | SW13 | SW1417 | SW1463 | SW156 | SW1573 | SW1710 | SW1783 | SW1990 |
| SW403 | SW48 | SW620 | SW626 | SW684 | SW756 | SW780 | SW837 | SW872 | SW900 |
| SW948 | SW954 | SW962 | SW982 | Saos-2 | Sarc9371 | Sci-1 | Set2 | SiHa | T-24 |
| T-T | T47D | T84 | T98G | TALL-1 | TASK1 | TC-71 | TC-YIK | TCCSUP | TE-1 |
| TE-10 | TE-11 | TE-12 | TE-15 | TE-4 | TE-441- | TE-5 | TE-6 | TE-8 | TE-9 |
| TF-1 | TGBC11TKB | TGBC1TKB | TGBC24TKB | TGW | THP-1 | TI-73 | TK | TK10 | TMK-1 |
| TOV-112D | TOV-21G | TT | TT2609-C02 | TUR | TYK-nu | Takigawa | Tera-1 | Toledo | U-118-MG |

【表１－４】

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| U-2-OS | U-266 | U-698-M | U-87-MG | U-CH1 | U031 | U251 | UACC-257 | UACC-62 | UACC-812 |
| UACC-893 | UDSCC2 | UISO-MCC-1 | UM-UC-3 | UMC-11 | UWB1.289 | VA-ES-BJ | VAL | VCaP | VM-CUB-1 |
| VMRC-LCD | VMRC-MELG | VMRC-RCW | VMRC-RCZ | WIL2-NS | WM-115 | WM1158 | WM1552C | WM239A | WM278 |
| WM35 | WM793B | WM902B | WSU-DLCL2 | WSU-NHL | YAPC | YH-13 | YKG-1 | YMB-1-E | YT |
| ZR-75-30 | huH-1 | no-10 | no-11 | | | | | | |

【０１５２】
網羅的解析データは、Ｇｅｎｏｍｉｃｓ ｏｆ Ｄｒｕｇ Ｓｅｎｓｉｔｉｖｉｔｙ ｉｎ Ｃａｎｃｅｒ（ＧＤＳＣ； ｈｔｔｐｓ：／／ｗｗｗ．ｃａｎｃｅｒｒｘｇｅｎｅ．ｏｒｇ／）で統括されており、このサイトから取得した。データとしては、各細胞株における、トランスクリプトームデータ、ゲノム配列データ、変異データ、ゲノムエピジェネティクスデータ、ｍｉＲＮＡ発現データ、ＲＮＡメチル化データを取得した。発現データは、ＥＭＢＬ－ＥＢＩ ＡｒｒａｙＥｘｐｒｅｓｓ、Ｅ－ＭＴＡＢ－３６１０ Ｔｒａｎｓｃｒｉｐｔｉｏｎａｌ Ｐｒｏｆｉｌｉｎｇ ｏｆ １，０００ ｈｕｍａｎ ｃａｎｃｅｒ ｃｅｌｌ ｌｉｎｅｓ（ｈｔｔｐｓ：／／ｗｗｗ．ｅｂｉ．ａｃ．ｕｋ／ａｒｒａｙｅｘｐｒｅｓｓ／ｅｘｐｅｒｉｍｅｎｔｓ／Ｅ－ＭＴＡＢ－３６１０／）から、変異データと感受性データは、ＧＤＳＣから直接ダウンロードした（ｈｔｔｐｓ：／／ｗｗｗ．ｃａｎｃｅｒｒｘｇｅｎｅ．ｏｒｇ／ｄｏｗｎｌｏａｄｓ）。各細胞株について、５－ＦＵに対する耐性情報を取得した。

【０１５３】
（画像化に使用した機器）
画像化には、以下の機器を使用した。当業者には明らかなことであるが、これと同等の機器であれば、同様に利用することができることが理解される。

【０１５４】
使用した機器としては、Ｗｉｎｄｏｗｓ（登録商標） ７， Ｃｏｒｅ ｉ７－４８１０ＭＱ ２．８０ＧＨｚ、ｍａｃＯＳ Ｘ１０．１３．６ ３．５ＧＨｚ ６－Ｃｏｒｅ Ｉｎｔｅｌ Ｘｅｏｎ Ｅ５、およびＣｅｎｔＯＳ ６．４ Ｉｎｔｅｌ Ｘｅｏｎ Ｅ５－２６９７ ｖ２@２．７０ＧＨｚを併用した。但し、最新版のＲ、ｉｆｏｒｔを利用できる環境にすれば、画像化のコンピュータは特段限定されず、いずれか１ｃｏｒｅで十分に計算可能な計算量である。並列化は時間短縮にのみ影響する。ソフトウェアとしては、Ｒ、Ｆｏｒｔｒａｎによる自作プログラムを用いて処理を行った。

【０１５５】
（画像化の方法）
画像化を行うため、縦および横方向に配列した２次元数値マトリックスに対して発現ユニットを割り当てた。具体的には、Ｅｎｓｅｍｂｌｅに登録されている全遺伝子およびｍｉＲＮＡをそれぞれ発現ユニットとした。数値マトリックスの１つの要素に対し１ピクセルを割り当てる。縦に１２５ピクセル（行）、横に２ピクセル（列）の長方形の領域（２５０ピクセル単位）を１つの単位として、発現ユニットの長さに応じて、横に隣接する複数の当該単位領域を割り当てた。各ピクセルには、２５６段階の色［モノクロの場合は明度］（０～２５５）のいずれかが設定される。

【０１５６】
各発現ユニットについて、上記で取得したデータから、発現量を求めた。各遺伝子もしくはエクソンについて、トランスクリプトーム内に出現する頻度をカウントし、トランスクリプトームの総リード長によって標準化し、各エクソンの発現量とした。また、各ｍｉＲＮＡの発現量を、ｍｉＲＮＡシーケンシングデータ中での各ｍｉＲＮＡにマッピングされたリード数を、総リード長によって標準化し、各ｍｉＲＮＡの発現量とした。当該発現量を正規化し１５０段階にグループ化する。各発現ユニット中の２５０ピクセル単位の左側の列を、発現量に対応する１～１５０の濃度の色のいずれかに設定した。

【０１５７】
各発現ユニットについて、上記で取得したデータから、配列データを求めた。各エクソンおよびｍｉＲＮＡをコードしている部分配列についての参照配列と、上記で取得したゲノムデータから、各細胞株におけるゲノム中に変異が存在する箇所および変異の内容の情報を取得した。各変異の情報を、各発現ユニットに割り当てられた領域に反映させた。各領域における行のそれぞれのピクセルは、発現ユニット中の配列の位置に対応する。

【０１５８】
各遺伝子もしくはエクソンおよび各ｍｉＲＮＡをコードしているゲノム中の部分配列に

おいて、参照配列と比較した塩基の置換が存在する場合には、各置換位置に対応する２５０ピクセル単位の行の左側のピクセルを、変異後の塩基に応じて、アデニン（２００）、チミン（２１０）、グアニン（２２０）、またはシトシン（２３０）の色に設定した。

【０１５９】

各遺伝子もしくはエクソンおよび各ｍｉＲＮＡをコードしているゲノム中の部分配列において、参照配列と比較した塩基の欠失が存在する場合には、各置換位置に対応する２５０ピクセル単位の行の左側のピクセルに、２５０（欠失）の色を設定した。

【０１６０】

各エクソンおよび各ｍｉＲＮＡをコードしているゲノム中の部分配列において、参照配列と比較した塩基の挿入が存在する場合には、各挿入の開始位置に対応する２５０ピクセル単位の行の左側のピクセルに、１８０（挿入開始）の色を設定し、１８０の色のピクセルの右側のピクセルから開始して、ピクセルを１つずつ、挿入されている塩基配列に応じて、アデニン（２００）、チミン（２１０）、グアニン（２２０）、またはシトシン（２３０）の色に順次設定した。

【０１６１】

各遺伝子またはエクソンおよび各ｍｉＲＮＡをコードしているゲノム中の部分配列において、エピジェネティック修飾が検出されている場合には、各修飾位置に対応する２５０ピクセル単位の行の右側のピクセルに以下のとおり修飾の種類に応じて色を設定した。

【０１６２】

ＤＮＡメチル化：１８６、ヒストンアセチル化：１８８、ヒストンメチル化：１９０、ヒストンユビキチン化：１９２、ヒストンリン酸化：１９４、ヒストンＳＵＭＯ化：１９６。

【０１６３】

各ＲＮＡにおいてメチル化が検出されている場合には、各修飾位置に対応する２５０ピクセル単位の行の左側のピクセルにメチル化の色を以下の通り設定した：ｍＲＮＡのメチル化について、ｍ６Ａ：２３５、Ａｍ：２３６、Ｍ６Ａｍ：２３７、ｍ６２Ａｍ：２３８、Ｉ：２４０、ｍ５Ｃ：２４２、Ｃｍ：２４３、ｍ７Ｇ：２４５、Ｇｍ：２４６、ｍ２７Ｇ：２４８、ｍ２２７Ｇ：２４９、Ｕｍ：２５１、Ｍ３Ｕｍ：２５２。なお、色の追加（例えば、２５６カラー、１６Ｂｉｔカラー等への変更）によってｔＲＮＡ、ｍｒＲＮＡ等のメチル化も対応可能と考えられる。

【０１６４】

各細胞の各発現ユニットについて上記工程を行い、各細胞について、発現データおよび配列データをまとめた画像を生成した。

【０１６５】

（実施例１－２）分析について

（特徴抽出）

画像解析用のニューラルネットワークを用いた機械学習によって、判別パラメータを最適化する。その際に、部分画像から連続的な明度色彩のつながりから、特徴となる部分を抽出する事を行う。その後、判別パラメータ係数の最適化を実施する。それを用いた判別モデルを構築する。

【０１６６】

（分類）

実施した判別パラメータを用いた判別モデルに基づき、グループ分類を行う。

【０１６７】

（実施例２）アレイ上の配置の工夫

（相関解析）

登録されている全ての細胞株において、正規化した遺伝子発現情報を用いて、全ての遺伝子組について連動して変化する傾向の度合いの解析を実施する。その際に、ピアソンの相関係数とスピアマンの相関係数を共に算出し、その平均化数値を算出する。また、相関の強い組み合わせ上位（今回は１００個）で抽出される遺伝子名をカウントする。

【０１６８】
（重回帰）
相関解析でカウントされた遺伝子の多い順で、その遺伝子が他の遺伝子発現量（正規化された値）を用いて、どのような係数を付与する事で記述できるか（線形結合で記述できるか）の決定を行う。
【０１６９】
（最適化）
相関解析で抽出し、最もカウントされた遺伝子をアレイの中心に配置する。その後、対象とした遺伝子との相関組を抜き出し、ピアソンとスピアマンの相関係数の平均値を、配置すべき遺伝子領域（１２５行×○○列）間の相互作用係数とする。中心遺伝子からの初期配置を相互作用係数に反比例するように設定し、次に配置した遺伝子からも同様に配置を繰り返して初期配置を設置する。その後の最適化の時点では、遺伝子間領域間の相互作用は、平均化相互作用係数をばね定数的に考え、初期配置の横方向にのみ位置を最適化する。そのため、各部分行（１２５行単位）では遺伝子間でのズレは許していないが遺伝子の部分領域の上下の接する場所は、先のばね定数に応じた力によって左右にずれる事を許容する。その結果、最適な配置を探索するというアルゴリズムを採用する。
【０１７０】
（実施例３）計算の効率化
（マシンスペック検出）
今回の機械学習に用いるマシンは、Linux（登録商標）OSを想定してプログラムを作成する。その場合、
cat/proc/cpuinfo
と言うコマンドを用いると、CPUのスペックを知る事が出来る。
【０１７１】
同様にメモリは、
cat/proc/meminfo
GPUは、
lspci┆grepVGA
NVIDIAドライバがインストールされている場合は、
nvidia-smi
にてマシンスペックを検出することができる。
【０１７２】
（データの分割）
画像の機械学習はGPUによる学習を想定しているため、GPU搭載メモリを考えて、学習データ数と検証データ数がメモリに乗る容量を考慮して、データ分割を実行する。
【０１７３】
（データの統合）
分割学習によって生成される各モデルの係数パラメータをニューラルネットワークの次元に応じた行列に格納する。分割分のパラメータ行列を一つの行列に格納する。そこで、この前パラメータを初期値とした新規の予測モデルを構築する。
【０１７４】
（最適化）
統合した初期パラメータとした予測モデルの部分パラメータを変化させたときに予測効率に生じる変化率を観測し、非線形最適化を実施する事によって、最安定パラメータを探索する。このときの計算は、ＨＤＤを仮想メモリとしＯｎ　ｔｈｅ　ｆｌｙでメモリとのやり取りを行い、ＣＰＵを使って最適化を実施する。
【０１７５】
（実施例４）解析例
対象とする腫瘍細胞株について、網羅的なトランスクリプトームデータ、ゲノム配列データ、変異データを取得した。上記学習によって得られたモデルを適用し、当該腫瘍細胞

株の５－ＦＵ耐性について予測する。当該腫瘍細胞株の５－ＦＵ耐性情報を取得し、モデルの妥当性を検証する。

【０１７６】

（実施例４－１）抗がん剤感受性の解析例

実施例１の（データ取得）に記載されるように腫瘍細胞株について、網羅的なトランスクリプトームデータ、ゲノム配列データ、変異データを取得した。５－ＦＵに対する感受性が特に高い１０の細胞株（MV-4-11、NOMO-1、OCI-AML2、PSN1、RPMI-6666、SIG-M5、SLVL、SR、SUPおよびYT）と、５－ＦＵに対する感受性が特に低い１０の細胞株（CAS-1、FU-OV-1、HCC1143、NCI-H1693、NCI-H2291、OVKATE、Saos-2、SKG-IIIa、SW684およびSW111）とを含む２０の腫瘍細胞株を訓練データとして用いた。

【０１７７】

上記データについて、実施例１の（画像化の方法）に記載される手順に実施例２に記載される改変を加えて、画像化を行った。

【０１７８】

画像について、実施例１に記載される（特徴抽出）および（分類）の手順にしたがい、また、実施例３に記載される（データの分割）にしたがって、画像と、抗がん剤感受性との相関の機械学習を行った。すなわち、生成した画像を１６×１６に分割し、各領域ごとに、画像解析用のニューラルネットワークを用いた機械学習によって、判別パラメータを最適化し、各領域ごとにモデルを生成した。

【０１７９】

各領域での学習において求められたパラメータによる判別式を元に、それらを統合した（分割前の画像全体に対する）新しい判別式を作成する。そのために、各部分学習のパラメータを初期値として、ＣＰＵを用いて全体を最適化する処理を実施し、画像全体から抗がん剤感受性を予測するモデルを生成した。

【０１８０】

２０種全ての細胞株データを一通り学習するのを１Ｅｐｏｃｈとカウントし、学習を繰り返す度に、生成されたモデルによる予測の精度を検証した。学習に用いたのとは異なる細胞株から同様に生成した画像をもとに、当該細胞株の５－ＦＵ感受性の予測における正答率を調べた。Ｅｐｏｃｈ数と正答率との関係は、図９に示される。構築した判別モデルでは、非学習画像を用いた細胞株に対しても１００％の精度で判別することが可能であった（図９）。

【０１８１】

同様の検証を、ＣＤＤＰ（シスプラチン）感受性について実施したところ、こちらも１００％の精度での判別が可能であった。

【０１８２】

（実施例４－２）画像化に使用するデータ種による学習効率の変化

実施例４－１に記載される手法にしたがい、腫瘍細胞株の訓練データを取得した。実施例４－１に記載されるＤＮＡ変異データとＲＮＡ発現量データの両方を画像化したものに加えて、ＤＮＡ変異データのみの情報を同様に画像化したものと、ＲＮＡ発現量データのみの情報を同様に画像化したものを生成した。

【０１８３】

それぞれの画像を実施例４－１と同様に学習に供し、Ｅｐｏｃｈ毎に生成されたモデルの精度を検証した。モデルの精度は、学習時に用いた画像での判別可能性と、学習時に未使用の画像での判別可能性とを調べた。結果を図１０に示す。

【０１８４】

ＤＮＡ変異データのみでは、抗がん剤感受性を判別することができるモデルの生成は困難であると考えられる。発現量データのみを用いる場合には、学習を繰り返すことによって、判別可能なモデルが生成できると考えられる。しかしながら、両方のデータを用いた場合には、およそ１００Ｅｐｏｃｈ程度で精度が１００％（図１０のグラフ中の１．０）

に収束していると考えられ、より効率的に学習できることが理解される。また、発現量データのみを使用した場合と、両方のデータを使用した場合とで、正答率の標準偏差を比較すると、発現量データのみの場合に１００Ｅｐｏｃｈで到達した標準偏差の値に、両方のデータを用いた場合には５８Ｅｐｏｃｈで到達した。このことから、両方のデータを用いる場合には、平均約４割ほど同一精度に到達する学習回数を削減することができる。

【０１８５】

（実施例４－３）分割領域ごとの収束性の相違

実施例４－１に記載されるように、生成した画像を１６×１６に分割し、各領域ごとに、画像解析用のニューラルネットワークを用いた機械学習によって、判別パラメータを最適化し、各領域ごとにモデルを生成した。上記分割では、１領域ごとにおよそ１００～２００遺伝子の情報が格納されることとなる。領域ごとのモデルについて、Ｅｐｏｃｈごとの検証精度の収束性を検証した（図１１）。

【０１８６】

５ＦＵ感受性を学習させた際の領域収束性を検証したところ、大抵の領域が収束性がない（Ｅｐｏｃｈ数を増やしても正答率が１に収束しない）領域に該当するが、一部の領域について、収束性があるモデルが生成されることが観察された（図１２）。これらの領域にて生成されたモデルは、それ自体が抗がん剤感受性の予測に利用可能であると考えられる。また、これらの収束性のある領域に着目してデータを統合して学習し、画像全体から抗がん剤感受性を予測するモデルを生成することができると考えられる。

【０１８７】

さらに、収束傾向がある領域のそれぞれについて、発現量情報によって判別が可能かどうかを検証した。具体的には、収束傾向がある領域に含まれる遺伝子の各細胞株における発現量について、クラスタリング分析を行い、抗がん剤感受性と相関するかどうかを調べた。

【０１８８】

分割領域に含まれる遺伝子の各発現量を基にクラスタリング分析を実施した。
対象判別グループが２つであり、各グループがそれぞれ同数を有するため、類似性に従って並び替えた各個体を中央で分離し、それぞれ分離したグループ内での同一性の割合を算出した。其々の同一性の割合が１００％であれば、発現情報のみで完全に分離可能であることを示し、５０％であれば、ランダムに分割されており発現情報のみでは分割できないことを意味する。本実施例では１０個中１個～２個の異なり以下、つまり、８０～９０％以上の場合、発現量のみで判別可能と判定した。

【０１８９】

収束傾向がある領域の大部分は、発現量の情報のみで抗がん剤感受性を判別可能であったが、収束性がある領域のうちのわずかな領域については、発現量のみでは抗がん剤感受性の判別ができなかった。当該領域を記述する遺伝子は５ＦＵ感受性に関与する遺伝子変異を有している可能性がある。これにより、遺伝子変異から抗がん剤感受性を予測するモデルを生成し得ると考えられる。また、領域ごとの収束性の相違を利用することで、ある形質に関与する遺伝子の変異を同定する方法に応用可能であると考えられる。

【０１９０】

抗がん剤の有効性判定モデルの分割学習により、抗がん剤の有効性を左右する遺伝子領域を同定し得る。全ゲノム情報を用いて抗がん剤耐性に関与する遺伝子領域の同定を行うことより、これまで確認されなかった抗がん剤耐性と遺伝子の新たな相関関係が判明する可能性があり、これは、抗がん剤に対する新規コンパニオン診断法の開発へとつながるものである。

【０１９１】

なお、本実施例では、抗がん剤感受性についての予測モデルを検証したが、学習データとして他の形質を用いれば、抗がん剤感受性以外の形質についても同様に予測モデルを生成することができると考えられる。

【０１９２】

（実施例５）ＤＮＡ・ＲＮＡ発現以外のメチル化を含めた実施例

複数の腫瘍細胞株について、網羅的なトランスクリプトームデータ、ゲノム配列データ、変異データ、ＤＮＡ上のエピジェネティック修飾データ、ＲＮＡ上のエピジェネティック修飾データを取得した。これらの情報をまとめ、上述のとおりの画像化を行う。この画像を用いて、当該腫瘍細胞株の薬剤耐性情報と、遺伝子情報との関係を上記のとおり学習する。学習によって生成したモデルを適用し、対象とする細胞株の薬剤耐性を予測する。対象とする細胞株からは、網羅的なトランスクリプトームデータ、ゲノム配列データ、変異データ、ＤＮＡ上のエピジェネティック修飾データ、ＲＮＡ上のエピジェネティック修飾データの全てまたは一部を取得し、モデルを適用することができる。

【０１９３】

（実施例６）ヘルスケアサービスへのサービス提供

新薬をがん細胞に投与し、そこから得られたＤＮＡ／ＲＮＡ情報を、上記のシステムで学習し、解析することで薬剤の作用機序を予測する。この予測された作用機序を、例えば、製薬企業に提供し得る。

【０１９４】

上記のシステムで、抗がん剤の応答結果を予測し、抗がん剤治療の薬剤選択を支援する。この予測結果を、例えば、病院を対象として提供し得る。

【０１９５】

上記のシステムで、複数の被験体の遺伝情報と、発症した疾患との関係を学習する。これによって得られたモデルに基づいて、対象とする被験体の遺伝情報から、当該被験体が発症する可能性がある疾患についての情報を提供することができる。

【０１９６】

上記のシステムで、ある疾患を有する被験体の遺伝情報と、ある薬剤に対する当該被験体の応答との関係を学習する。これによって得られたモデルに基づいて、対象とする被験体に対して、有効と考えられる薬剤についての情報を提供することができる。

【０１９７】

遺伝情報を入力すると、当該遺伝情報を送信し、上記モデルの適用結果を受信し、所望の結果を表示するアプリケーションもまた提供され得る。アプリケーションは、遺伝情報を画像化することが可能であり得る。

【０１９８】

がん患者のシークエンスの画像化データから、その人に最適な抗がん剤を予測する医療支援システムを開発及び提供する。かかるシステムは、真の個別化医療の実現に貢献するものと考える。最適な抗がん剤の選択システムを構築し、医療機関や検査機関からの依頼により、検査受託および／またはクラウド上での診断補助サービスなどの提供を行う。データを蓄積することも想定される。抗がん剤以外の他の疾患の治療への応用や、製薬企業の新薬開発の際の効果または副作用などの予測、基礎研究におけるシークエンスデータの解析サービスなどを提供する。ゲノム情報の機械学習におけるプラットフォームを提供する。

【０１９９】

（注釈）

以上のように、本開示の好ましい実施形態を用いて本開示を例示してきたが、本開示は、請求の範囲によってのみその範囲が解釈されるべきであることが理解される。本明細書において引用した特許、特許出願および文献は、その内容自体が具体的に本明細書に記載されているのと同様にその内容が本明細書に対する参考として援用されるべきであることが理解される。

【０２００】

本出願は、日本国特許出願第２０１８－２４７９５９号（２０１８年１２月２８日出願）の優先権を主張し、当該出願の内容は、その全体が全ての目的について本明細書において参考として援用される。

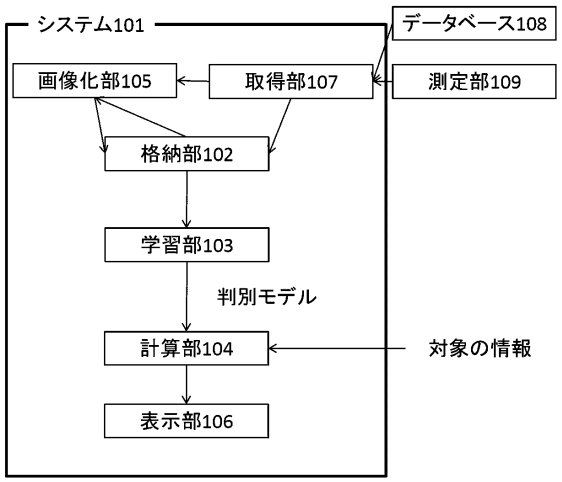
【産業上の利用可能性】

【０２０１】
本開示は、個体の形質の予測が有用である分野、とりわけ医療の分野において利用可能である。予め疾患の発症の傾向を予測する他、例えば、適切な処置の決定などに有用である。
【符号の説明】
【０２０２】
１０１：システム
１０２：格納部
１０３：学習部
１０４：計算部
１０５：画像化部
１０６：表示部
１０７：取得部
１０８：データベース
１０９：測定部

【図面】

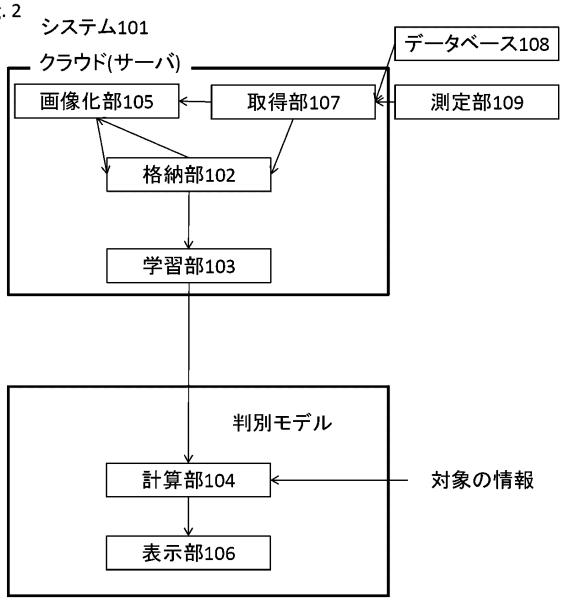
【図１】

Fig. 1



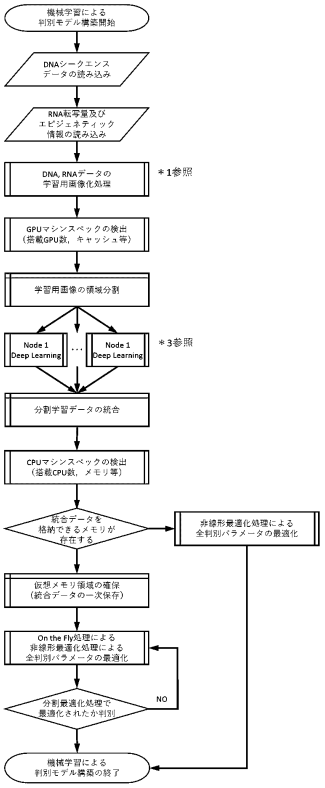

【図２】

Fig. 2



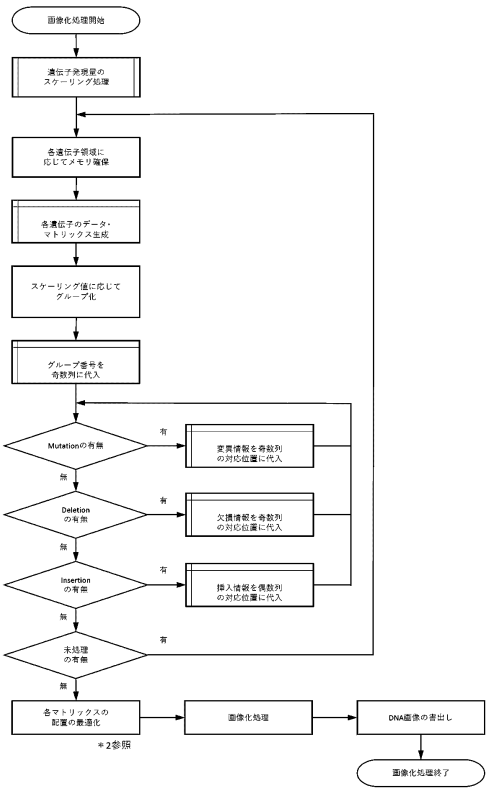

【図３】

DNA, RNA画像化データの機械学習用概要



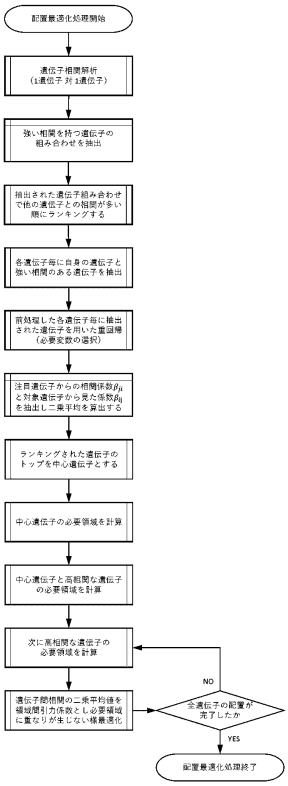

【図４】

＊1： DNA, RNAデータの学習用画像化処理

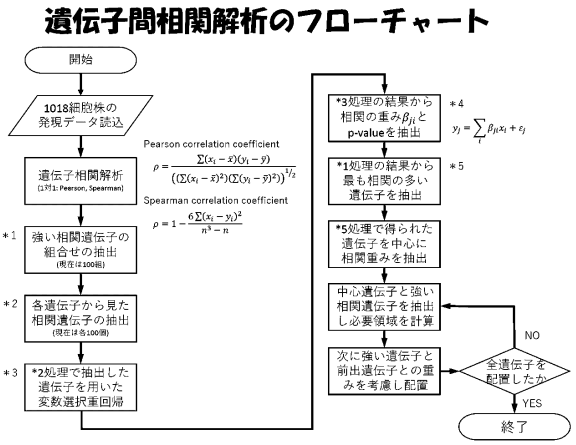


【図５】

＊2：　遺伝子データの配置最適化（学習効率促進）

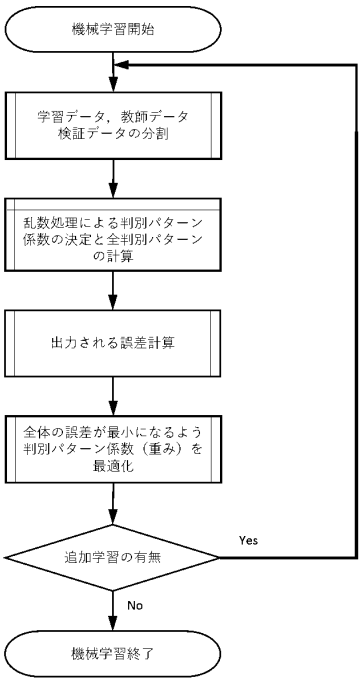



【図６】

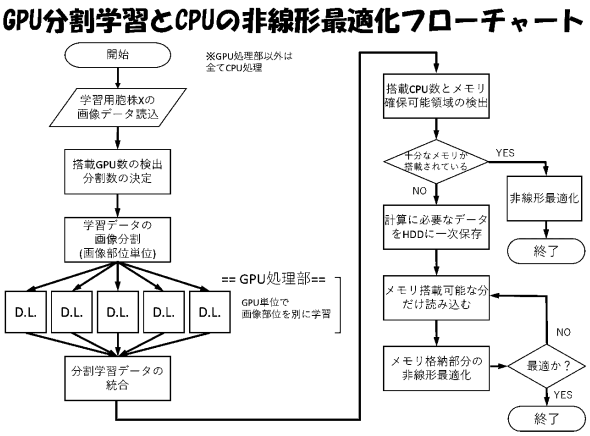



【図７】

＊3：　Deep Learning処理の概要

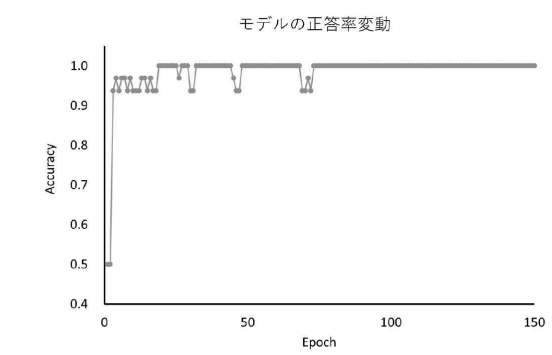



【図８】



【図９】

Fig. 9




【図１０】

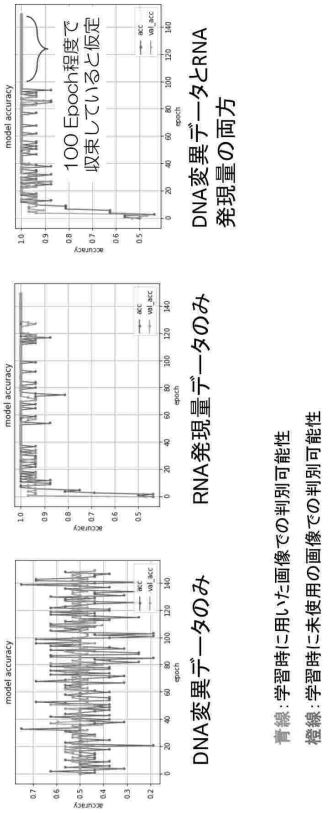



Fig. 10

【図１１】

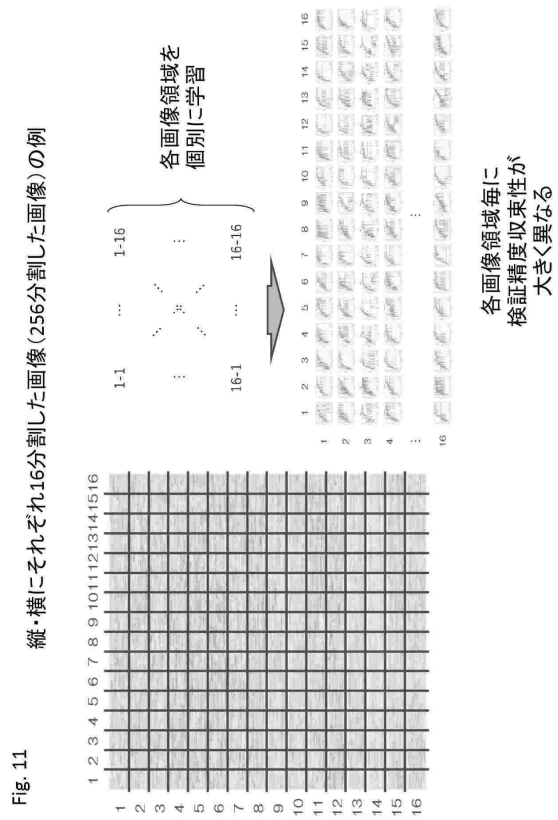



Fig. 11

【図１２】

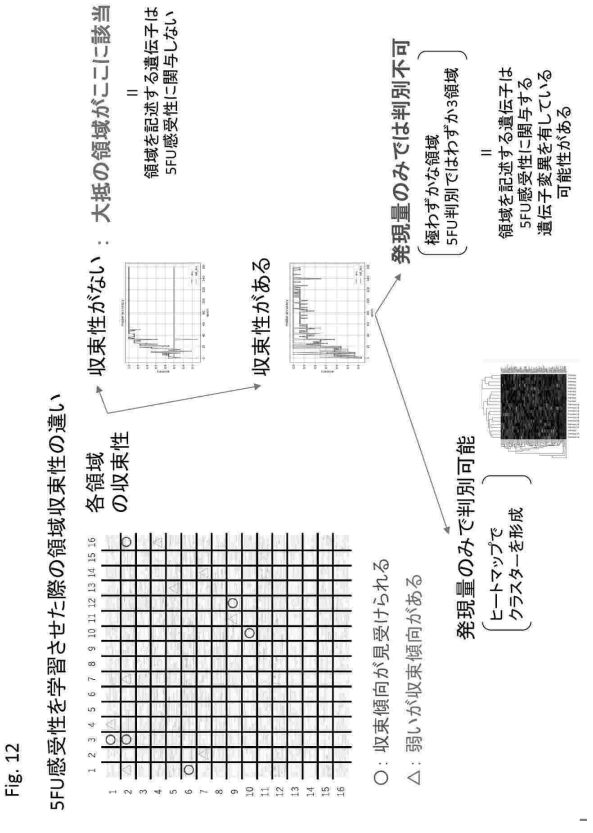



Fig. 12

フロントページの続き

(72)発明者　森 正樹
大阪府吹田市山田丘１番１号 国立大学法人大阪大学内
(72)発明者　浅井 歩
大阪府吹田市山田丘１番１号 国立大学法人大阪大学内
(72)発明者　小関 準
大阪府吹田市山田丘１番１号 国立大学法人大阪大学内
審査官　山崎 誠也
(56)参考文献　特表２０２１－５２１５３６（ＪＰ，Ａ）
特表２０２１－５３１０９８（ＪＰ，Ａ）
特開２０１６－０９９９０１（ＪＰ，Ａ）
特開２０１８－０９２４５３（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ　５／００－９９／００