(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2023年10月5日(05.10.2023)

WIPO | PCT

(10) 国際公開番号
WO 2023/190136 A1

(51) 国際特許分類:
G16B 40/00 (2019.01)　*C12Q 1/6869* (2018.01)
C12M 1/00 (2006.01)　*G01N 33/50* (2006.01)
C12M 1/34 (2006.01)　*G06N 20/00* (2019.01)
C12N 15/11 (2006.01)　*G16B 30/00* (2019.01)
C12Q 1/6844 (2018.01)

(21) 国際出願番号：　PCT/JP2023/011772

(22) 国際出願日：　2023年3月24日(24.03.2023)

(25) 国際出願の言語：　日本語

(26) 国際公開の言語：　日本語

(30) 優先権データ：
特願 2022-056626　2022年3月30日(30.03.2022)　JP

(71) 出願人: 富士フイルム株式会社 (**FUJIFILM CORPORATION**) [JP/JP]; 〒1068620 東京都港区西麻布２丁目２６番３０号 Tokyo (JP).

(72) 発明者: シン ジャンマジェイ (**SINGH, Janmajay**); 〒2588538 神奈川県足柄上郡開成町宮台７９８番地 富士フイルム株式会社内 Kanagawa (JP).

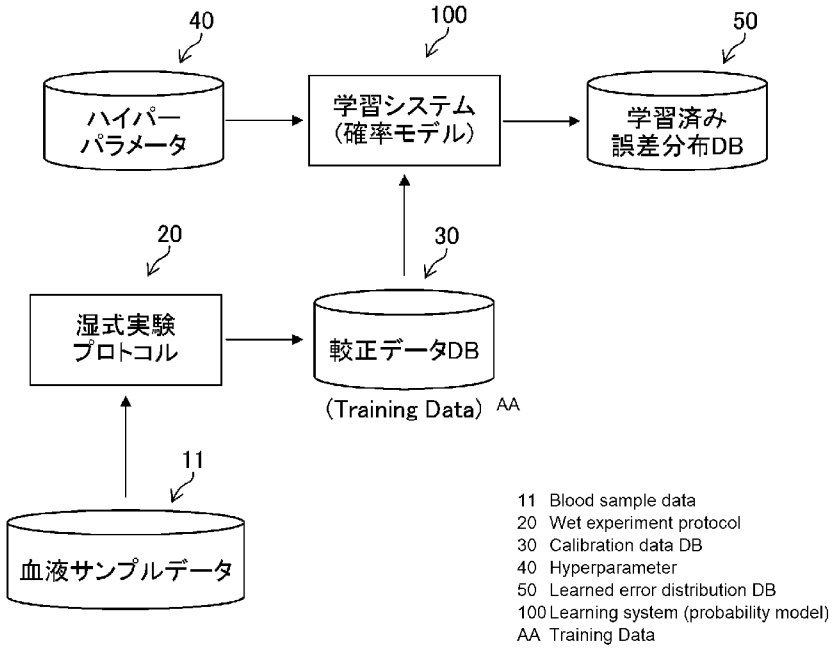
(74) 代理人: 松浦 憲三 (**MATSUURA, Kenzo**); 〒1600023 東京都新宿区西新宿一丁目８番１号 新宿ビルディング５階 新都心国際特許事務所 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH,

WO 2023/190136 A1

(54) **Title:** LEARNING SYSTEM, DECISION SYSTEM, AND PREDICTION SYSTEM, AND LEARNING METHOD, DECISION METHOD, AND PREDICTION METHOD

(54) 発明の名称: 学習システム、決定システム、及び予測システム、並びに学習方法、決定方法、及び予測方法



(57) **Abstract:** One embodiment of the present invention provides a learning system, a decision system, and a prediction system, and a learning method, a decision method, and a prediction method. In DNA methylation measurements, there exist a problem of incomplete bisulfite conversion (problem 1), a problem of the occurrence of bias when several different

［続葉有］

CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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.

(84) 指定国（表示のない限り、全ての種類の広域保護が可能）：ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

— 国際調査報告（条約第21条(3)）

biomarker sequences/genes are amplified together (problem 2), and a problem in that the degree of over-amplification of the non-methylation signal depends on the gene sequence itself and the chemical substance used for the measurement (problem 3). One embodiment of the present invention provides a system and a corresponding method for learning measurement error characteristics in the presence of the three problems and reflecting the learned error characteristics in biomarker selection criteria. Addressing the problem of measurement error characteristic evaluation for DNA methylation in the presence of a combination of problems 1-3 forms a major novelty of the present invention.

(57) 要約：本発明の一形態は、学習システム、決定システム、及び予測システム、並びに学習方法、決定方法、及び予測方法を提供する。ＤＮＡのメチル化測定においては、バイサルファイト変換の不完全性の問題（問題１）と、いくつかの異なるバイオマーカー配列／遺伝子が一緒に増幅された場合にバイアスが生じる問題（問題２）と、非メチル化シグナルの過剰増幅の程度が、遺伝子配列そのものと、測定に用いられる化学物質とに依存するという問題（問題３）が存在する。本発明の一態様では、３つの問題が存在する中で測定誤差特性を学習し、学習した誤差特性をバイオマーカー選択基準に反映させるシステム及びそのシステムに対応する方法を提供する。問題１～３の組合せが存在する下でのＤＮＡのメチル化に対する測定誤差特性評価の問題に取り組むことは、本発明の主要な新規性を形成する。

明　細　書

発明の名称 ：

学習システム、決定システム、及び予測システム、並びに学習方法、決定方法、及び予測方法

技術分野

[0001] 本発明は、バイオマーカーの値を測定する技術に関する。

背景技術

[0002] ＤＮＡ（deoxyribonucleic acid）では、「メチル化」と呼ばれる現象が起きることが知られている。メチル化とは、メチル分子がシトシンに化学的に結合することによる修飾をいう。このシトシン（Ｃ：cytosine）は、グアニン（Ｇ：guanine）、アデニン（Ａ：adenine）、チミン（Ｔ：thymine）と共に、ＤＮＡを構成する４つの必須核酸塩基を構成する。核酸塩基の任意の配列は「ヌクレオチド配列」と呼ばれ、タンパク質などの重要な情報をコードするヌクレオチド配列は「ゲノム配列」または「遺伝子」と呼ばれる。

[0003] ヒトでは、ＤＮＡストランド上でシトシンがグアニンに続く場所（「ＣｐＧサイト」と呼ばれる）では、メチル化がとりわけ一般的である。メチル化状態は遺伝子の活性化または抑制化に影響し、ある種の遺伝子のＣｐＧサイトのメチル化状態は、多くの疾患の重要なバイオマーカーを形成する。通常、疾患診断の定量的モデルを作製するために、幾つかのバイオマーカー候補配列の組合せから得られたデータが用いられる。このため、バイオマーカーのＤＮＡメチル化を測定することが重要になる。

[0004] ＤＮＡの測定過程ではデータに誤りが加わり、どんな推測/予測の信頼性にも影響を与える。バイオマーカーの選択を最適化するための従来の研究は、測定プロセスにおいてほんのわずかな誤りを想定し、利用可能なデータの予測値のみに焦点を当てている。このような方法の例としては、（Artificial Intelligence分類器の性能のような）定量モデルからの出力信号に頼って、分類のための特徴としてバイオマーカー配列を使用するかどうかを決定する

特徴選択アルゴリズムが知られている。

[0005] このような従来の技術に関し、例えば特許文献１では、代表的なバイオマーカーデーターからバイオマーカーセットを選択して評価することが記載されている。また、非特許文献１には、ＰＣＲバイアス（ＰＣＲ：polymerase chain reaction）の測定及び緩和について記載されている。

先行技術文献

特許文献

[0006] 特許文献1：特表２０１７－５２３４３７号公報

非特許文献

[0007] 非特許文献1：“Measuring and Mitigating PCR Bias in Microbiome Data”、 Justin D. Silverman他、［２０２２年３月２２日検索］、インターネット（https://www.biorxiv.org/content/10.1101/604025v1）

発明の概要

発明が解決しようとする課題

[0008] 次のセクションでは、バイオマーカー配列（シークエンス）の測定誤差特性を学習しようとした先行研究について詳細に議論する。これらの先行研究とそれらに関連する問題について論じ、それぞれの段階の詳細な説明を行う。

[0009] ［ＤＮＡのメチル化測定］

メチル化測定の概要を図１に示す。メチル化の測定では、血液サンプル１０がバイサルファイト変換され、ＰＣＲ装置で遺伝子/シグナルが増幅され、次世代シーケンサー等で測定される。これら一連の測定手順は、湿式実験プロトコル２０（wet experiment protocol）を構成する。

[0010] ［ＳＴＥＰ１：バイサルファイト変換］

Ｃｍ（メチル化シトシン）とＣｕ（非メチル化シトシン）を区別するために、バイサルファイト変換（Bisulfite conversion）の追加ステップが使用される。バイサルファイト変換では、Ｃｕはウラシル（Ｕ：uracil）に変換

され、ＣｍはＣｍのままである。変換されたサンプルがシーケンス化されると、ＣｍはＣ（シトシン）として読み出され、一方、ウラシルはチミンとして読み出される。これにより、シトシンのメチル化状態を区別することが可能になる。

[0011] ［問題１：バイサルファイト変換における問題］

この手順の理想的な結果は、Ｃｕが１００％ウラシルに変換され、Ｃｍは全くウラシルに変換されないこと（変換が０％であり、ＣｍがＣｍのまま）である。しかし、化学反応の性質上、変換の成功（または不成功）の程度は確率論的であり、定量的な研究は困難である。このような、バイサルファイト変換の不完全性を、以下「問題１」という。

[0012] ［ＳＴＥＰ２：ＰＣＲ増幅］

この段階は、測定のシグナル増幅段階と理解することができる。標準的には（つまり、メチル化のためではなく、バイサルファイト変換をしなければ）、それぞれの「信号」は興味のある遺伝子または配列である。生のデータでは、このような配列の数は非常に少ないので、派生した信号は弱い。そのため、元の配列を何度もコピーすることで、配列数を増やし、シグナルを増幅することが考えられる。例えば、ＰＣＲ前の遺伝子１のシグナル強度をG1_preと呼び、ＰＣＲ後のシグナル強度をG1_postと呼ぶことにする。なお、実際には、多くの遺伝子/シグナルを同時に増幅することに焦点を当てる。したがって、遺伝子２に関し、G2_preとG2_postを遺伝子１と同様に定義する。

[0013] さて、まず上述のＳＴＥＰ１を行うと、たった１つの遺伝子でも2つのシグナルが得られる。例えば、遺伝子１は、ウラシルを含む別の配列に変換される、非メチル化としてＣｐＧを有するいくつかの配列を持つ。同様に、ＣｐＧがメチル化されている配列は変換されない。これは一般的であり、肝臓と胃のＤＮＡの混合物でみられる。そのような混合物では、肝臓に重要な遺伝子が肝細胞ではメチル化されていないが、胃細胞ではメチル化されている（したがって抑制されている）可能性がある。そこで、遺伝子１に関し、ＰＣＲ前信号の強さとＰＣＲ後信号の強さをG1_U_Pre およびG1_U_postとし（メ

チル化されていない場合）、G1_M_Pre およびG1_M_postとし（メチル化されている場合）、解読された配列をG1_M_Pre およびG1_M_post とする。

[0014] ［問題２：単一バイサルファイトプロトコルにおけるＰＣＲバイアス］

同じ遺伝子のシグナルを増幅しても、バイサルファイト変換は２つのシグナルタイプになる。したがって、G1_U_post/G1_U_pre = G1_M_post/G1_M_preは成り立たない。G1_U_pre = G1_M_preの場合であっても、増幅後は、G1_U_post/G1_U_pre ＞ G1_M_post/G1_M_preとなる（すなわち、非メチル化遺伝子が、メチル化遺伝子に対して過剰に増幅される）ことが知られている。しかし、このような非メチル化シグナルの過剰増幅の程度は、遺伝子配列そのものと、測定に用いられる化学物質とに依存する。この問題を、以下「問題２」と呼ぶ。

[0015] ［問題３：ＰＣＲ増幅における問題］

ＰＣＲの理想的な結果は、G1_post/G1_pre = G2_post/G2_preである。しかし実際には、ある種の遺伝子配列は他のものよりも測定しやすく、この等価性は成立しない。このような、いくつかの異なるバイオマーカー配列/遺伝子が一緒に増幅された場合に生じるバイアスを、多重化プロトコルにおける「ＰＣＲバイアス」と呼ぶ（以下、「問題３」という）。

[0016] ［従来の技術における対応］

上述した問題１～３に対する従来技術での対応を説明する。従来技術では、問題１に関し、定量的な研究にはしばしば極端な正確さが必要とされず、したがって、バイサルファイト変換の成功の程度を考慮していなかった。また、問題３に関し、これまでの微生物学の研究では、ＰＣＲの効果を掛け算的に考えていた。すなわち、従来の技術では、もし１回のＰＣＲサイクル後の遺伝子１のシグナル強度がｊであれば、２回のサイクル後のシグナル強度はj^2であり、ｘ回のサイクル後のシグナル強度は同様にj^xであると考えていた。この仮定を用いて、ＰＣＲは、多項ロジスティック－通常線形モデルを用いた対数線形過程としてモデル化された。「バッチ効果」（バッチごとにわずかに異なるバイアス特性を示す標本に対するＰＣＲ）などの他の共変

量も確率論的な方法で含まれた。モデルは、生成された較正データを「訓練」した後、ＰＣＲバイアスの補正に使用される。

[0017] また、問題２に関しては、単一プロトコル設定における測定誤差とバイアスの特徴付けはより簡単であるため、一部のＰＣＲデータでは、バイアスの度合いを見出すために線形回帰を行っている。線形回帰推定量を計算した後、この方程式を用いてこのようなバイアスを補正することができる。

[0018] ＤＮＡメチル化の正確な測定の重要性はすでに述べた。病気診断のような応用分野では、複数のバイオマーカー配列のデータを用い、定量モデルに入力することは珍しいことではない。複数のバイオマーカーのメトリクス値を同時に測定する測定プロセスを設計する際は、問題１、問題２、問題３が組み合わされてしまい、これらの全てが問題となる。このため、誤差の定量化と学習誤差の特性が非常に困難になる。本発明では、３つの問題が存在する中で測定誤差特性を学習し、学習した誤差特性をバイオマーカー選択基準に反映させるシステムを検討する。この問題の組合せが存在する下でのＤＮＡのメチル化に対する測定誤差特性評価の問題に取り組むことは、本発明の主要な新規性を形成する。

[0019] 本発明は、液体生命学（リキッドバイオプシー）のように、複数の遺伝子からのＤＮＡメチル化の同時的で非常に正確な測定が必要な場合に、特に重要となる。特に、癌のような疾患の正確な同定のために、ある種の癌細胞遺伝子は、健康な細胞における同じ遺伝子と比較して高いメチル化を示すことが知られている。問題２は、そのような場合、測定が、がんと正常なＤＮＡの混合から真のメチル化比を過小評価する（負のバイアス）ことを意味する。問題１と問題３は、過小評価の度合いをさらに悪化させる。

[0020] 本発明は上記事情に鑑みてなされたもので、その一形態は、バイオマーカー配列の測定誤差特性を学習する学習システム及び学習方法を提供する。また、本発明の一形態は、学習した誤差特性を反映して配列セットを決定する決定システム及び決定方法、並びに学習システムあるいは学習方法により得られたデータを用いて遺伝子配列の測定誤差特性を予測する予測システム及

び予測方法を提供する。

課題を解決するための手段

[0021] 本発明の第１の態様に係る学習システムは、測定プロトコル変数と、バイオマーカー配列の結果として生じる誤差特性との関係を学習する学習システムであって、プロセッサを備え、プロセッサは、重要性のある変数について適切なデータが入手できるように設計された較正データを入力し、確率モデルを用いて、重要性のある変数について各測定プロトコルにわたる誤差分布の特性を学習し、確率モデルは、バイサルファイト変換の誤差をモデル化するために、適切に選択された事前パラメータで初期化された第１のパラメータと、バイオマーカー配列の増幅の相互依存性をモデル化するために、適切に選択された事前パラメータで初期化された第２のパラメータと、ＰＣＲ全体のバイアスをモデル化するために、適切に選択された事前パラメータで初期化された第３のパラメータと、を含む。第１の態様に係る学習システムは、測定プロトコル変数とバイオマーカー配列の結果として生じる誤差特性の間の関係（テンプレート対プロダクト比と定義される）を学習するシステムである。

[0022] 第１の態様及び以下の各態様において、「重要性のある変数」とは、信号増幅性能に影響を与えることが実験室の専門家によって知られている変数であり、そのような変数について、ＰＣＲ装置が調整される。例えば、後述する図２に示すようなＰＣＲ温度やＰＣＲサイクル数は「重要性のある変数」の一例である。温度が高すぎると、ＤＮＡが分解され、標的遺伝子配列を複製するために必要な反応が起こらない。また、第１～第３のパラメータに関し、「適切に選択された事前パラメータ」として同じパラメータを用いてよい。また、「較正データ（キャリブレーションデータ）の入力」に関し、例えばＰＣＲ温度の場合、通常のＰＣＲで使用される温度の範囲で適切な表示が可能であることを要する。

[0023] 第２の態様に係る学習システムは第１の態様において、第２のパラメータは、バイサルファイト変換後の遺伝子のメチル化配列及び非メチル化配列の

カウントを別々に取得し、取得されたカウントを、メチル化配列及び非メチル化配列の各配列について、事前変数を別々に決定できる多項分布でモデル化したパラメータである。第２の態様は、上述した問題２に対応するための第２のパラメータの具体的態様を規定するもので、バイサルファイト変換の誤差をモデル化及び修正して、バイオマーカー配列のメチル化を正しく評価できるようにするものである。第２の態様では、経験的なデータ分析から、より優れた事前変数を選択することができる。なお、第２の態様において取得されたカウントは、塩基配列のＧＣ比（グアニンとシトシンの比）のような要因に基づいてモデル化することができる。

[0024] 第３の態様に係る学習システムは第１または第２の態様において、第３のパラメータは、ユニバーサルプライマーを用いて複数の配列を同時に増幅する場合に、多項分布で計算されたカウントの個々のカウントの合計がガウス分布に従う、という構成データ制約が課されたパラメータである。第３の態様は、上述した問題３に対応するための第２のパラメータの具体的態様を規定するもので、複数のバイオマーカーの数が多く、構成データ制約が計算可能になるようにモデリングパラメータを単純化する場合、複数の分散カウントにおける各カウントの合計はガウス分布に従う。また、複数の配列を同時に増幅する場合、個々のマーカーのカウント値が独立ではなく、合計値がほぼ一定になるような増幅の仕方をするため、上記のような多項分布によるモデリングが適している。さらに、バイサルファイト変換を伴うメチル化計測においては、各マーカーがメチル化、非メチル化の２状態があるため、マーカー数×２のカウント値に対するモデリングになる。

[0025] 本発明の第４の態様に係る決定システムは、プロセッサを備える決定システムであって、プロセッサは、多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し、第１から第３の態様のいずれか１つに係る学習システムから、学習された誤差特性、及び誤差特性に関連付けられたメタデータを入力し、あらかじめ決められた基準を用いて入力したヌクレオチド配列、測定プロトコル情報、学習された誤差

特性、及びメタデータを使用して、可能なバイオマーカー配列のセットのための第１のスコアを出力し、各セットについての第１のスコアの値を考慮してバイオマーカー配列セットを決定する。第４の態様に係る決定システムでは、多重パネルでバイオマーカー配列を使用するかどうかを決定するために、第１の態様に係るシステムからの出力を使用する。第１のスコアは測定精度に由来するスコアであり、測定誤差が小さいほど高い値となる「低誤差スコア」である。

[0026] 第５の態様に係る決定システムは第４の態様において、プロセッサは、決定すべきバイオマーカー配列ごとに、第２のスコアを入力し、バイオマーカー配列セットにおける各バイオマーカー配列についての第１のスコアを考慮して、第１のスコアと第２のスコアとのバランスを最適化することにより多重化パネルのベストなサブセットを選択する。第５の態様に係る決定システムでは、マルチプレックスパネルの最終目標を考慮することにより、バイオマーカー配列の、よりバランスの取れた選択を可能にするために、第４の態様を増強する。第２のスコアは、たとえば予測したい疾患との関連度が大きいほど高いスコア（関連度スコア）である。また、「第１のスコアと第２のスコアのバランス」は、例えば第１のスコアと第２のスコアの相加平均や相乗平均で規定される第３のスコアを算出し、その第３のスコアを最大化することにより、最適化することができる。

[0027] 本発明の第６の態様に係る予測システムは、遺伝子配列の測定誤差特性を予測する予測システムであって、プロセッサを備え、プロセッサは、多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し、第１から第３の態様のいずれか１つに係る学習システムから、学習された誤差特性、及び誤差特性に関連付けられたメタデータを入力し、２つの遺伝子配列間の類似性の尺度を計算するための測定基準を用いて、以前に較正データに含まれていたバイオマーカー配列と新たなバイオマーカー配列との類似度を計算し、計算した類似度を他の関連する入力及び学習された誤差特性と組み合わせて使用して、較正データに含まれていな

いバイオマーカー配列を測定する際の誤差特性を予測する。第６の態様に係る予測システムは、第１～第３の態様に係る学習システムを、較正データに含まれていなかったバイオマーカー配列に使用することを可能にする。

[0028] なお、第６の態様において「他の関連する入力」とは、例えばバイオマーカー配列に対応するメタデータを意味する。例えば、遺伝子タイプが「プロモーターもしくはエンハンサー」であり、ＣｐＧタイプが「アイランド、ショア、シェルフ」であり、ＣＧの豊富さが「高、低」であれば、あるバイオマーカー配列Ｇ１についてのこれらの情報の組み合わせ（メタデータの一例）は、「プロモーター、アイランド、低」というベクトルとして表すことができる。

[0029] 第７の態様に係る予測システムは第６の態様において、プロセッサは、予測された誤差特性を使用して、較正データに含まれていないバイオマーカー配列と最も類似する、較正データにおいて利用可能であったバイオマーカー配列を取得し、取得したバイオマーカー配列の情報を、第４または第５の態様に係る決定システムにおけるバイオマーカー配列セットの決定に反映する。第７の態様では、第４または第５の態様に係る決定システムを用いて、バイオマーカー配列セット選択において、較正データに含まれていないバイオマーカー配列を使用できるようにする。

[0030] 本発明の第８の態様に係る学習方法は、プロセッサを備え、測定プロトコル変数と、バイオマーカー配列の結果として生じる誤差特性との関係を学習する学習システムにより実行される学習方法であって、プロセッサが、重要性のある変数について適切なデータが入手できるように設計された較正データを入力し（較正データ入力ステップ）、確率モデルを用いて、重要性のある変数について各測定プロトコルにわたる誤差分布の特性を学習し（学習ステップ）、確率モデルは、バイサルファイト変換の誤差をモデル化するために、適切に選択された事前パラメータで初期化された第１のパラメータと、バイオマーカー配列の増幅の相互依存性をモデル化するために、適切に選択された事前パラメータで初期化された第２のパラメータと、ＰＣＲ全体のバ

イアスをモデル化するために、適切に選択された事前パラメータで初期化された第３のパラメータと、を含む。第８の態様は、上述した第１の態様に対応する学習方法を規定するものである。

[0031] 第９の態様に係る学習方法は第８の態様において、第２のパラメータは、バイサルファイト変換後の遺伝子のメチル化配列及び非メチル化配列のカウントを別々に取得し、取得されたカウントを、メチル化配列及び非メチル化配列の各配列について、事前変数を別々に決定できる多項分布でモデル化したパラメータである。第９の態様は、上述した第２の態様に対応する学習方法を規定するものである。

[0032] 第１０の態様に係る学習方法は第８または第９の態様において、第３のパラメータは、ユニバーサルプライマーを用いて複数の配列を同時に増幅する場合に、多項分布で計算されたカウントの個々のカウントの合計がガウス分布に従う、という構成データ制約が課されたパラメータである。第１０の態様は、上述した第３の態様に対応する学習方法を規定するものである。

[0033] 本発明の第１１の態様に係る決定方法は、プロセッサを備える決定システムにより実行される決定方法であって、プロセッサは、多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し（配列情報入力ステップ）、第８から第１０の態様のいずれか１つに係る学習方法の結果として得られる、学習された誤差特性、及び誤差特性に関連付けられたメタデータを入力し（学習結果入力ステップ）、あらかじめ決められた基準を用いて入力したヌクレオチド配列、測定プロトコル情報、学習された誤差特性、及びメタデータを使用して、可能なバイオマーカー配列のセットのための第１のスコアを出力し（スコア出力ステップ）、各セットについての第１のスコアの値を考慮してバイオマーカー配列セットを決定する（配列セット決定ステップ）。第１１の態様は、上述した第４の態様に対応する決定方法を規定するものである。

[0034] 第１２の態様に係る決定方法は第１１の態様において、プロセッサは、決定すべきバイオマーカー配列ごとに、第２のスコアを入力し（スコア入力ス

テップ）、バイオマーカー配列セットにおける各バイオマーカー配列についての第１のスコアを考慮して、第１のスコアと第２のスコアとのバランスを最適化することにより多重化パネルのベストなサブセットを選択する（サブセット選択ステップ）。第１２の態様は、上述した第５の態様に対応する決定方法を規定するものである。

[0035] 本発明の第１３の態様に係る予測方法は、プロセッサを備え、遺伝子配列の測定誤差特性を予測する予測システムにより実行される予測方法であって、プロセッサは、多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し（配列情報入力ステップ）、第８から第１０の態様のいずれか１つに係る学習方法により得られた、学習された誤差特性、及び誤差特性に関連付けられたメタデータを入力し（学習結果入力ステップ）、２つの遺伝子配列間の類似性の尺度を計算するための測定基準を用いて、以前に較正データに含まれていたバイオマーカー配列と新たなバイオマーカー配列との類似度を計算し（類似度計算ステップ）、計算した類似度を他の関連する入力及び学習された誤差特性と組み合わせて使用して、較正データに含まれていないバイオマーカー配列を測定する際の誤差特性を予測する（誤差特性予測ステップ）。第１３の態様は、上述した第６の態様に対応する予測方法を規定するものである。

[0036] 第１４の態様に係る予測方法は第１３の態様において、プロセッサは、予測された誤差特性を使用して、較正データに含まれていないバイオマーカー配列と最も類似する、較正データにおいて利用可能であったバイオマーカー配列を取得し（配列取得ステップ）、取得したバイオマーカー配列の情報を、第１１または第１２の態様に係る決定方法におけるバイオマーカー配列セットの決定に反映する（情報反映ステップ）。第１４の態様は、上述した第７の態様に対応する予測方法を規定するものである。

[0037] なお、上述した態様の学習方法、決定方法、及び予測方法をプロセッサに実行させるプログラム（学習プログラム、決定プログラム、予測プログラム）、及びそれらプログラムのコンピュータ読み取り可能なコードを記録した

非一時的記録媒体も、本発明の範囲に含まれる。

発明の効果

[0038] 以上説明したように、本発明に係る学習システム、決定システム、及び予測システム、並びに学習方法、決定方法、及び予測方法は、以下の効果を有する。

（１）複数の遺伝子配列を一緒に測定して多重化されたパネルを扱うことができる。

（２）バイサルファイト変換されたサンプルを処理することができる。

（３）配列パラメータとプロトコルパラメータを入力として使用して、測定誤差を予測することができる。

（４）配列を分析/分類の目的に使用するかどうかを決定することができる。

図面の簡単な説明

[0039] [図1]図１は、ＤＮＡのメチル化を測定する様子を示す図である。

[図2]図２は、較正データを作成する様子を示す図である。

[図3]図３は、学習システム、及び学習システムに関連するデータを示す図である。

[図4]図４は、学習システムの構成を示す図である。

[図5]図５は、確率モデルの実施形態を示す図である。

[図6]図６は、決定システムと予測システムとの関係を示す図である。

[図7]図７は、決定システムの構成を示す図である。

[図8]図８は、予測システムの構成を示す図である。

発明を実施するための形態

[0040] 以下、本発明の実施形態を説明する。説明においては、必要に応じて添付図面が参照される。なお、添付図面において、説明の便宜上一部の構成要素の記載を省略する場合がある。

[0041] ［較正データの作成］

本発明においては、図２に示すように、まず、血液サンプル１０から、湿式実験プロトコル２０（wet experiment protocol）により、ＰＣＲ温度やＰ

ＣＲサイクル数のような重要な測定プロトコル変数によって構成される較正データを作成することを必要とする。この較正データは、重要性のある変数について適切なデータが入手できるように設計されていることが好ましい。最終的には、配列の測定結果を、プロトコル情報と共に較正データＤＢ３０（ＤＢ：database）に保存する（以下では、データベースを「ＤＢ」と記載する場合がある）。なお、図２では、較正データ作成手順の一部を、明確化のため省略した。

[0042] 学習アルゴリズム（学習システム、学習方法）は、このようなプロトコル変数とその測定特性との間の関係を学習するために、この較正データを使用する。次いで、所与の測定プロトコル変数のセット（較正データには含まれていない）に対して、このシステム（予測システム、予測方法）は、所与のバイオマーカー配列の測定誤差特性を予測することができる。この予測を用いて、システム（決定システム、決定方法）はバイオマーカー配列が何らかの定量的研究に使用するのに適しているかどうかを決定することができる。最後に、較正データに存在しないバイオマーカー配列でさえ、本発明のシステム（決定システム、決定方法）は、測定誤差特性が知られている最も類似した配列を見つけ出し、それを用いて新しい配列について類似した決定を行うことができる。

[0043] 具体的には、本発明は、確率モデルの作用を詳述し、「テンプレート対プロダクト」比を推定することによって、バイオマーカー配列の測定誤差を特徴付ける。「テンプレート」はバイオマーカー配列の最初の量（ＰＣＲ増幅前の量）を指し、「プロダクト」はＰＣＲ増幅後の同じバイオマーカー配列の最終量（ＰＣＲ増幅後の量）を指す。

[0044] ［学習システムの構成］

図３は、本発明の一態様に係る学習システム１００及びこれに関連するデータ等を示している。多重バイサルファイトＰＣＲプロトコルのためのＤＮＡメチル化測定誤差特性を学習するための、このような学習システムの適用は、本発明の新規性を保証するための最小限の要件である。なお、後述する

ように、学習システム１００には、その結果（学習済み誤差分布ＤＢ５０）を利用する決定システム２００（決定システム）及び予測システム３００（予測システム）が付随していてもよい。

[0045] 図４は、学習システム１００の構成例を示す図である。図４に示すように、学習システム１００は、プロセッサ１１０（プロセッサ、コンピュータ）、確率モデル１２０（確率モデル）、記憶部１３０、ＲＯＭ１４０（ＲＯＭ：Read Only Memory）、ＲＡＭ１５０（ＲＡＭ：Random Access Memory）を備える。プロセッサ１１０は学習システム１００の各部が行う処理の統括制御を行うもので、較正データ入力部１１２及び学習部１１４を有する。

[0046] プロセッサ１１０は、図４に示す要素の他に、不図示の表示制御部や通信制御部、出力制御部等を含んでいてよい。

[0047] プロセッサ１１０は、例えば、ＣＰＵ（Central Processing Unit）、ＧＰＵ（Graphics Processing Unit）、ＦＰＧＡ（Field Programmable Gate Array）、ＰＬＤ（Programmable Logic Device）等の各種のプロセッサや電気回路で構成される。これらのプロセッサや電気回路がソフトウェア（プログラム）を実行する際は、実行するソフトウェアのコンピュータ（例えば、プロセッサを構成する各種のプロセッサや電気回路、及び／またはそれらの組み合わせ）で読み取り可能なコードをＲＯＭ１４０等の非一時的かつ有体の記録媒体に記憶しておき、コンピュータがそのソフトウェアを参照する。非一時的かつ有体の記録媒体に記憶しておくソフトウェアは、本発明に係る学習方法、予測方法、決定方法を実行するためのプログラム（学習プログラム、予測プログラム、決定プログラム）、及び実行に際して用いられるデータを含む。ＲＯＭ１４０ではなく各種の光磁気記録装置、半導体メモリ等の非一時的かつ有体の記録媒体にコードを記録してもよい。ソフトウェアを用いた処理の際には例えばＲＡＭ１５０が一時的記憶領域として用いられ、また例えば不図示のＥＥＰＲＯＭ（Electronically Erasable and Programmable Read Only Memory）やフラッシュメモリ等の非一時的かつ有体の記録媒体に記憶されたデータを参照することもできる。「非一時的かつ有体の記録媒体」

として記憶部１３０を用いてもよい。

[0048] 記憶部１３０はハードディスク、半導体メモリ等の各種記憶デバイス及びその制御部により構成され、上述した較正データや、学習方法の実行条件及び実行結果（学習済み誤差分布のデータ）等を記憶することができる。

[0049] 学習システム１００は、図４に示す要素の他に、不図示の表示装置（例えば、液晶モニタ）や操作装置（例えば、マウスやキーボード）を含んでいてよい。表示装置には、較正データや誤差分布のデータ等を表示することができ、また、ユーザは、操作部を介して、本発明に係る学習方法（学習プログラム）の実行に必要な操作を行うことができる。

[0050] 上述した図３は、血液サンプルデータ１１を示しているが、これは組織サンプルを含むあらゆる生物学的データである。血液サンプルデータ１１は、図１のように、上述したＳＴＥＰ　１，ＳＴＥＰ　２，それにＤＮＡ配列決定を加えた測定手順で測定するものであり、それ自体が、ＰＣＲのサイクル数など、その有効性に影響を与える変数（重要性のある変数）をいくつか持っている。このような変数のいくつかの値からデータを得る必要があるため、関連する変数が最初に識別され、これらの値の範囲で測定が行われる。例えば、もしＰＣＲサイクル数が唯一の重要な変数である場合、５，１０および１５　ＰＣＲサイクルの同じ血液サンプルのデータを生成することができる。これがいわゆる較正データである。

[0051] ［確率モデル］

学習システム１００は、較正データＤＢ３０に記憶された較正データ（訓練用データ）を利用して（較正データ入力ステップ）、確率モデルをトレーニングする（学習ステップ）。図５はそのような確率モデルの一例である確率モデル１２０を、ベイズ階層モデルを通して示している。本発明の重要な新規性は、(i)バイサルファイト変換誤差の事前情報（事前パラメータ；以下同じ）、(ii)バイサルファイト変換の共変量の事前情報、および(iii)バイオマーカー配列の増幅の相互依存性の事前情報を使用することにある。これらの事前情報（i）～（iii）は、本発明の第１～第３のパラメータに対応し、

従って上述の問題１～３に対応する。以上の３つの要素を総合すると、本発明は、上述した特許文献１や非特許文献１のような従来のモデルとは異なるものとなっている。

[0052] また、これら３つの要因により、上述した問題１＋問題２＋問題３が一体となった問題を解決することができる。学習システム１００は、最適化方法（最小化のための損失関数など）に従い、一連のハイパーパラメータ（ハイパーパラメータ４０）を通して調整される。このような調整は、システムの最終性能を確認し、それを最大化するハイパーパラメータを選択することによって行われる。

[0053] なお、上述した第１～第３のパラメータは確率モデル１２０の一部（したがって、学習システム１００の一部）であり、それらパラメータの値は訓練プロセス中に更新される。また、第１～第３のパラメータは学習システム１００の一部であるため、図３では表示されていない。

[0054] 一方、ハイパーパラメータを使用すると、確率モデル１２０のある側面を制御できる。ただし、ハイパーパラメータの値はユーザが設定するものであり、訓練プロセス中に値が更新されることはない。また、学習システム１００と決定システム２００（図６を参照）とでは、ハイパーパラメータが異なっている。

[0055] より具体的には、二項分布のバイオマーカーについて、バイサルファイト変換の誤差をモデル化することを選択することができる。そこで、バイサルファイト変換誤差の事前確率（事前パラメータの一例）を［０，１］の間の値として選ぶことができる。事前確率が０の場合、そのバイオマーカーの完全な変換（Ｃｕとウラシルとの１００％の変換と、Ｃｍの０％の変換）を想定し、事前確率が０より大きい場合はそのバイオマーカーの不完全な変換（Ｃｕの一部のみがウラシルに変換され、Ｃｍの一部もウラシルに変換される）を想定する。理想的には、事前変数は経験的データ分析から設定されるべきである。バイサルファイト共変量には、ナノグラムで測定したサンプル中に加えられた亜硫酸塩の量と初期ＤＮＡ量が含まれる。このようにして、事前

確率で初期化されたバイサルファイト変換誤差が、第１のパラメータである。

[0056] 同様に、ＰＣＲ誤差分布は多項分布でモデル化することができ、適切な事前確率を設定することができる。この段階では、ＰＣＲ後の配列カウント（配列の数）は、選択されたバイオマーカーの数をｘとした場合にN1、N2、...、Nxとして表すことができ、その配列カウントは多項分布としてモデル化することができる。ここで、“Ni”は、ｉ番目のバイオマーカーの配列カウントである。ＰＣＲ共変量には、ＰＣＲ温度やＰＣＲサイクル数のような、較正データ作成のために選択された要因が含まれることがある。

[0057] 本発明の新規性は、同じ配列から２つの異なるカウントの可能性を考える能力にある。１つは、バイサルファイト変換後のある配列の塩基化のカウント（メチル化配列のカウント）であり、もう１つは、その配列の非塩基化タイプのカウント（非メチル化配列のカウント）である。これにより、N1_M、N1_U、N2_M、N2_U などの可能性の数が２倍になることが考慮される。ここで“Ni_M”は、ｉ番目のバイオマーカーについてのメチル化配列のカウントを示し、“Ni_U”は、ｉ番目のバイオマーカーについての、非メチル化配列のカウントを示す。Nx_M とNx_U は、互いに自然な制約を課すため（一方の平均の回数が多いことは他方の回数が少ないことを意味する）、このような制約（相互依存性）を用いてモデル化問題を単純化することができる。このようにして、事前確率で初期化されたバイオマーカー配列の増幅の相互依存性が、第２のパラメータである。

[0058] 最後に、全体分布モデル（ＰＣＲ全体のバイアスを示すモデル）は、すべてのバイオマーカー、すなわち、N1+N2+....+、Nxの総数の配列を数量化することであり、バイオマーカーカウントを通して相互依存の制約（構成データ制約）を課すためにさえ使用され得る（例えば、N1が高すぎる場合、N3は低すぎる）。N1、N2等の各々（個々のカウント）は多項分布であるため、選択したバイオマーカーの数が多い（例えば、３０以上）条件下では、それらの合計（多項分布で計算されたカウントの個々のカウントの合計）は、中心極

限定理を満足するために、ガウス分布に従うと考えられる。配列タイプ間のこのような相互依存性（構成データ制約）は、すぐには明らかではないかもしれないが、ユニバーサルプライマーを用いて複数の配列（複数のバイオマーカー配列）を同時に増幅する場合には、このような相互依存性が存在することが知られている。このようにして、事前確率で初期化されたＰＣＲ全体のバイアスが、第３のパラメータである。

[0059] このようなユニバーサルプライマーの使用は、適切なアダプタ配列が標的バイオマーカー配列の両端に配備された後にのみ可能である。この段階で追加されたユニバーサルプライマーの有限な量は、バイオマーカー配列間の組成依存性を作り出し、純シグナル増幅に影響する。ＰＣＲ増幅中の多重化パネルに構成データ制約を課し、ユニバーサルプライマーを用いて相対的なバイオマーカー配列の豊富さをモデリングすることは、本発明の第二の新規性を形成する。

[0060] ［決定システム及び予測システムの位置づけ］

上述した学習システム１００には、図６に示すように、決定システム２００（決定システム）及び予測システム３００（予測システム）が付随していてもよい。これら決定システム及び予測システムを学習システム１００に付加することは、選択肢として推奨されるものである。決定システム２００及び予測システム３００を付加することで、例えば学習システム１００により学習された誤差特性を用いて、決定システム２００により候補バイオマーカーのベストなサブセットを見つけることができ（学習結果入力ステップやスコア入力ステップ、サブセット選択ステップ等を含む、本発明に係る決定方法の実行による）、これによりバイオマーカー配列の選択基準に情報を与えて（情報反映ステップ等を含む、本発明に係る予測方法の実行による；予測システム３００）、学習システム１００の効果的な活用を助けることができる。

[0061] 上述した学習システム１００は、最適化基準を統計的手段によって最大化または最小化することによって学習する確率モデル１２０を備えており、こ

のようにして学習することは、アルゴリズムを「トレーニング」することの意味を広くカバーしている。一方、決定システム２００は、学習システム１００がトレーニングを終了した後に機能する。決定システム２００自体には、最大化または最小化しようとする定義済みの最適化基準がないため、「トレーニング」されておらず、システムは学習されない。ただし、決定システム２００は、システムを「調整可能」にするハイパーパラメータを含んで構成されている。

[0062] ［決定システム及び予測システムの構成］

図７は、決定システム２００の構成を示す図である。同図に示すように、決定システム２００は、プロセッサ２１０（プロセッサ）と、ＲＯＭ２３０（非一時的かつ有体の記録媒体）と、ＲＡＭ２４０とを備える。プロセッサ２１０は、配列情報入力部２１２と、学習結果入力部２１４と、スコア出力部２１６と、配列セット決定部２１８と、を備える。決定システム２００は、これら要素の他に、図示せぬ表示制御部や表示装置、記憶装置、操作部等を有していてよい。

[0063] 図８は、予測システム３００の構成を示す図である。同図に示すように、予測システム３００は、プロセッサ３１０（プロセッサ）と、ＲＯＭ３３０（非一時的かつ有体の記録媒体）と、ＲＡＭ３４０とを備える。プロセッサ３１０は、配列情報入力部３１２と、学習結果入力部３１４と、類似度計算部３１６と、誤差特性予測部３１８と、配列情報反映部３２０と、を備える。予測システム３００は、これら要素の他に、図示せぬ表示制御部や表示装置、記憶装置、操作部等を有していてよい。

[0064] 決定システム２００及び予測システム３００のこれらの要素は、学習システム１００と同様に、例えば、ＣＰＵ、ＧＰＵ、ＦＰＧＡ、ＰＬＤ等の各種のプロセッサや電気回路で構成される。これらのプロセッサや電気回路がソフトウェア（プログラム）を実行する際は、実行するソフトウェアのコンピュータで読み取り可能なコードをＲＯＭ２３０やＲＯＭ３３０等の非一時的かつ有体の記録媒体に記憶しておき、コンピュータがそのソフトウェアを参

照する。非一時的かつ有体の記録媒体に記憶しておくソフトウェアは、本発明に係る予測方法、決定方法を実行するためのプログラム（予測プログラム、決定プログラム）、及び実行に際して用いられるデータを含む。ＲＯＭ２３０やＲＯＭ３３０ではなく、各種の光磁気記録装置、半導体メモリ等の非一時的かつ有体の記録媒体にコードを記録してもよい。ソフトウェアを用いた処理の際には例えばＲＡＭ２４０，ＲＡＭ３４０が一時的記憶領域として用いられ、また例えば不図示のＥＥＰＲＯＭやフラッシュメモリ等の非一時的かつ有体の記録媒体に記憶されたデータを参照することもできる。

[0065] ［スコアに基づくバイオマーカー配列セットの決定］

以下では、「スコアの最適化プロセス（最適化手法）」の２つの大まかな分類である、バイナリーベースのアプローチと、組み合わせベースのアプローチについて説明する。

[0066] バイナリーベースの最適化基準では、決定システム２００の配列情報入力部２１２（プロセッサ）が関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し（配列情報入力ステップ）、学習結果入力部２１４（プロセッサ）が、学習システム１００から、学習された誤差特性、及び誤差特性に関連付けられたメタデータを入力する（学習結果入力ステップ）。スコア出力部２１６（プロセッサ）は、学習した測定誤差特性を独立に考慮し、結果として生じる測定誤差グラフの傾きに基づいて、例えば｛＋１，０，－１｝のスコア（測定誤差スコア；第１のスコアの一例）を各バイオマーカー配列に割当てることができる（スコア出力ステップ）。配列セット決定部２１８は、各バイオマーカーの順序からスコア（第１のスコア）を合計し、その組み合わせ（バイオマーカー配列セット）を使うかどうかを決定することができる（配列セット決定ステップ）。

[0067] これを、よりロバストに実装すると、特徴選択アルゴリズムと同様の方法で、組み合わせベースの最適化基準を設計することができる。特徴選択アルゴリズムは、定量モデルの出力に依存し、定量モデルの性能を最適化するためにそれらの基準を更新する。この従来の特徴選択アルゴリズムの見方は、

測定エラー特性から生じるスコア（第１のスコア）を考慮し、与えられたバイオマーカー配列セットからのサブセット選択に最良の情報を与えるために、信号と同じスコア（第１のスコア）を使用するために修正することにより、本発明で用いる組み合わせベースの最適化基準を設計することができる。組み合わせベースの最適化基準の場合も、バイナリーベースの最適化基準の場合と同様に、決定システム２００の各要素を用いてバイオマーカー配列セットを決定することができる（配列情報入力ステップ～配列セット決定ステップの実行）。

[0068] なお、上述したバイナリーベースの最適化基準では各バイオマーカー配列に独立にスコアを割り当てるのに対して、組み合わせベースの最適化基準の場合は、バイオマーカー配列の組み合わせに対してスコアを与える。このため、測定誤差の小ささを各マーカー配列で独立に扱ってよい場合はバイナリーベースの最適化基準が適しており、相互依存性が特に大きい場合は組み合わせベースの最適化基準が適している。相互依存性とは、例えば、「バイオマーカー配列１はバイオマーカー配列２と同時に測定する場合は測定誤差が小さいが、バイオマーカー配列３と当時に測定する場合は測定誤差が大きい」という場合である。

[0069] 本発明では、上述した測定誤差スコア（第１のスコア）だけでなく「予測したい疾患との関連度が大きいほど高いスコア」（関連度スコア；第２のスコアの一例）を考慮し、これらスコアのバランスを最適化することによりバイオマーカー配列セットを決定することもできる。このような関連度スコア（第２のスコア）を併せて用いる場合、上述した測定誤差スコア（第１のスコア）と関連度スコア（第２のスコア）とのバランスを最適化する（例えば、測定誤差スコアと関連度スコアとの相加平均や相乗平均を最大化する）ことにより、バイオマーカー配列セットを決定することができる。この場合、関連度スコアも、測定誤差スコアの場合と同様にバイオマーカー配列ごとに独立に割り当てることもできるし、バイオマーカー配列の組み合わせに対して与えることもできる。例えば、マーカー１，２，３がいずれも疾患と関連

している場合に、マーカー１，２の相関が小さくマーカー１，３の相関が大きい場合は、マーカー１，２の組み合わせの方が疾患予測に有効であり、関連度スコアが高くなる。

[0070] 如何なる最適化基準（バイナリーベース、特徴選択アルゴリズム、あるいは組み合わせベースのような）が最良であるかは、応用分野、ユーザ、および時間の制約に依存し、それらの条件に合わせて適宜選択することができる。この出力は、システムが共同で考えるバイオマーカー配列のセットのものであり、多重化ＰＣＲ配列決定のための与えられた測定誤差のプロトコルには、最小限の誤差がある。実施態様は、本発明の第５，第１２の態様を考慮するために、決定システムの実施に基づいて（バランスの取れた配列選択を考慮するか否かにかかわらず）変更することができる。

[0071] なお、決定システム２００は、学習システム１００で得られた誤差分布（図６では、学習済み誤差分布データベース５０）に依存しており、それ自体では、元の較正データに含まれないバイオマーカー配列のスコアを計算することができない。このような、較正データに含まれていなかったバイオマーカー配列の測定誤差特性の予測については、図６に示すように、また以下に説明するように、本発明の第６，第７の態様に係る予測システム３００（及び、本発明の第１３，第１４の態様に係る予測方法）が必要である。この予測システム３００は、決定システム２００について上述したのと同様に本発明に係る学習システム１００（学習システム）への追加であり、このような新しいバイオマーカー配列の使用事例、存在、重要性に依存する。

[0072] ［較正データに含まれていないバイオマーカー配列の測定誤差特性の予測］

関心配列データベース６０に含まれる関心バイオマーカー配列が較正データに含まれていないバイオマーカー配列であることが判明した場合（図６の判断「訓練データに含まれている配列か？」でＹＥＳの場合）に、その関心バイオマーカー配列の測定誤差特性を予測する手法について説明する。この場合、まず予測システム３００に入力を渡す。具体的には、予測システム３

００の配列情報入力部３１２（プロセッサ）は、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し（配列情報入力ステップ）、また学習結果入力部３１４（プロセッサ）は、学習システム１００から、例えば学習済み誤差特性、及び誤差特性に関連付けられたメタデータを入力する（学習結果入力ステップ）。ここで、「メタデータ」は、例えば遺伝子のタイプ（プロモーターかエンハンサーか）、遺伝子の領域（転写開始サイト等）であるが、これらには限定されない。

[0073] そして、類似度計算部３１６（プロセッサ）は、レーベンシュタイン距離やＧＣ含量（GC-content；ＤＮＡ分子中の窒素塩基のうち、グアニンとシトシンの割合）のような、２つの遺伝子配列間の類似度の測定基準（類似性の尺度）を用いて、以前に較正データに含まれていたバイオマーカー配列（較正データにおいて利用可能であったバイオマーカー配列）と新たなバイオマーカー配列（関心バイオマーカー配列）との類似度を計算する（類似度計算ステップ）。類似度計算部３１６は、学習済み誤差分布データベース５０に存在するバイオマーカー配列から、関心バイオマーカー配列と「最も類似する」バイオマーカー配列を見つける（類似度計算ステップ）。予測システム３００は、検出された「最も類似する配列」の情報を用いて、その「最も近い配列」に対応する学習済み誤差特性を（学習済み誤差分布データベース５０から）取得することができ（誤差特性予測ステップ、配列取得ステップ）、これにより本発明の第６，第１３の態様を完全に実装することができる。配列情報反映部３２０は、この情報を、本発明の第４，第５の態様を実施する決定システム２００（及び、本発明の第１１，第１２の態様に係る決定方法）と併用して、決定システム２００におけるバイオマーカー配列セットの決定に反映する（情報反映ステップ）こともできる。

[0074] ［実施例］

遺伝子配列の候補セットは、まず、ＧＣ含量のような測定関連要因の十分な変化を示すと考えられている。例えば、配列ＧＣ内容のみを重要と仮定し、「高」の３遺伝子配列および「低」のＧＣ含量の３遺伝子配列を同時測定

のために決定することができる。次に、一連の重要な測定プロトコル関連変数を特定し、範囲を考慮する。これに続いて、すべての値とすべての変数の考えられる全範囲について、湿式実験手順を実行する。例えば、｛５，１０，１５｝のＰＣＲサイクルを考慮し、メチル化比率を｛５％，１０％｝と考えるならば、同一の生物標本のアリコット（aliquot）から６個の遺伝子の独立測定を３＊２＝６個行う。続いて、学習システムで用いられる前述の確率モデルを訓練し、決定システムのハイパーパラメータを調整（チューニング）した。さて、がん診断の場合のように、より多くの遺伝子バイオマーカーを探しながら、その遺伝子が測定特性の良否を評価するために決定モデルを用いることができる。

[0075] １００の遺伝子配列測定で訓練されたArtificial Intelligenceガン分類モデルが７０％の感度で行われることを考えると、パフォーマンスが低い理由の一部は、一部の遺伝子で高い測定ノイズである可能性がある。上記のシステムを用いて１００の遺伝子を再考し、測定困難なものを取り除くと、測定誤差を回避することにより、Artificial Intelligenceの性能は８０％に上昇する可能性があり、より良い頑健性を持つ。

[0076] 以上説明した実施形態は、以下の効果を有する。

（１）複数の遺伝子配列を一緒に測定して多重化されたパネルを扱うことができる。

（２）バイサルファイト変換されたサンプルを処理することができる。

（３）配列パラメータとプロトコルパラメータを入力として使用して、測定誤差を予測することができる。

（４）配列を分析/分類の目的に使用するかどうかを決定することができる。

（５）適切に学習された誤差特性、適切に選択されたバイオマーカー配列セット（及びそのサブセット）、精度良く予測されたバイオマーカー配列セットの測定誤差により、バイオマーカー配列を用いた分析や診断（例えば、上述したＡＩによるがんの分類）等を精度良く行うことができる。

[0077] 以上で本発明の実施形態について説明してきたが、本発明は上述した態様

に限定されず、種々の変形が可能である。

符号の説明

[0078] １０　血液サンプル

１１　血液サンプルデータ

２０　湿式実験プロトコル

３０　較正データＤＢ

４０　ハイパーパラメータ

５０　学習済み誤差分布データベース

６０　関心配列データベース

１００　学習システム

１１０　プロセッサ

１１２　較正データ入力部

１１４　学習部

１２０　確率モデル

１３０　記憶部

２００　決定システム

２１０　プロセッサ

２１２　配列情報入力部

２１４　学習結果入力部

２１６　スコア出力部

２１８　配列セット決定部

３００　予測システム

３１０　プロセッサ

３１２　配列情報入力部

３１４　学習結果入力部

３１６　類似度計算部

３１８　誤差特性予測部

３２０　配列情報反映部

請求の範囲

[請求項1] 測定プロトコル変数と、バイオマーカー配列の結果として生じる誤差特性との関係を学習する学習システムであって、プロセッサを備え、

前記プロセッサは、

重要性のある変数について適切なデータが入手できるように設計された較正データを入力し、

確率モデルを用いて、前記重要性のある変数について各測定プロトコルにわたる誤差分布の特性を学習し、

前記確率モデルは、

バイサルファイト変換の誤差をモデル化するために、適切に選択された事前パラメータで初期化された第１のパラメータと、

バイオマーカー配列の増幅の相互依存性をモデル化するために、適切に選択された事前パラメータで初期化された第２のパラメータと、

ＰＣＲ全体のバイアスをモデル化するために、適切に選択された事前パラメータで初期化された第３のパラメータと、

を含む学習システム。

[請求項2] 前記第２のパラメータは、バイサルファイト変換後の遺伝子のメチル化配列及び非メチル化配列のカウントを別々に取得し、取得されたカウントを、前記メチル化配列及び前記非メチル化配列の各配列について、事前変数を別々に決定できる多項分布でモデル化したパラメータである請求項１に記載の学習システム。

[請求項3] 前記第３のパラメータは、ユニバーサルプライマーを用いて複数の配列を同時に増幅する場合に、多項分布で計算されたカウントの個々のカウントの合計がガウス分布に従う、という構成データ制約が課されたパラメータである、請求項１または２に記載の学習システム。

[請求項4] プロセッサを備える決定システムであって、

前記プロセッサは、

多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し、

請求項１から３のいずれか１項に記載の学習システムから、前記学習された誤差特性、及び前記誤差特性に関連付けられたメタデータを入力し、

あらかじめ決められた基準を用いて前記入力した前記ヌクレオチド配列、前記測定プロトコル情報、前記学習された誤差特性、及び前記メタデータを使用して、可能なバイオマーカー配列のセットのための第１のスコアを出力し、

各セットについての前記第１のスコアの値を考慮してバイオマーカー配列セットを決定する、

決定システム。

[請求項5] 前記プロセッサは、

決定すべきバイオマーカー配列ごとに、第２のスコアを入力し、

前記バイオマーカー配列セットにおける各バイオマーカー配列についての前記第１のスコアを考慮して、前記第１のスコアと前記第２のスコアとのバランスを最適化することにより前記多重化パネルのベストなサブセットを選択する、

請求項４に記載の決定システム。

[請求項6] 遺伝子配列の測定誤差特性を予測する予測システムであって、プロセッサを備え、

前記プロセッサは、

多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し、

請求項１から３のいずれか１項に記載の学習システムから、前記学習された誤差特性、及び前記誤差特性に関連付けられたメタデータを入力し、

２つの遺伝子配列間の類似性の尺度を計算するための測定基準を用

いて、以前に較正データに含まれていたバイオマーカー配列と新たなバイオマーカー配列との類似度を計算し、

前記計算した類似度を他の関連する入力及び前記学習された誤差特性と組み合わせて使用して、前記較正データに含まれていないバイオマーカー配列を測定する際の誤差特性を予測する、

予測システム。

[請求項7] 前記プロセッサは、

前記予測された誤差特性を使用して、前記較正データに含まれていないバイオマーカー配列と最も類似する、前記較正データにおいて利用可能であったバイオマーカー配列を取得し、

前記取得したバイオマーカー配列の情報を、請求項４または５に記載の決定システムにおけるバイオマーカー配列セットの決定に反映する、

請求項６に記載の予測システム。

[請求項8] プロセッサを備え、測定プロトコル変数と、バイオマーカー配列の結果として生じる誤差特性との関係を学習する学習システムにより実行される学習方法であって、

前記プロセッサが、

重要性のある変数について適切なデータが入手できるように設計された較正データを入力し、

確率モデルを用いて、前記重要性のある変数について各測定プロトコルにわたる誤差分布の特性を学習し、

前記確率モデルは、

バイサルファイト変換の誤差をモデル化するために、適切に選択された事前パラメータで初期化された第１のパラメータと、

バイオマーカー配列の増幅の相互依存性をモデル化するために、適切に選択された事前パラメータで初期化された第２のパラメータと、

ＰＣＲ全体のバイアスをモデル化するために、適切に選択された事

前パラメータで初期化された第３のパラメータと、

を含む学習方法。

[請求項9] 前記第２のパラメータは、バイサルファイト変換後の遺伝子のメチル化配列及び非メチル化配列のカウントを別々に取得し、取得されたカウントを、前記メチル化配列及び前記非メチル化配列の各配列について、事前変数を別々に決定できる多項分布でモデル化したパラメータである、請求項８に記載の学習方法。

[請求項10] 前記第３のパラメータは、ユニバーサルプライマーを用いて複数の配列を同時に増幅する場合に、多項分布で計算されたカウントの個々のカウントの合計がガウス分布に従う、という構成データ制約が課されたパラメータである、請求項８または９に記載の学習方法。

[請求項11] プロセッサを備える決定システムにより実行される決定方法であって、

前記プロセッサは、

多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し、

請求項８から１０のいずれか１項に記載の学習方法の結果として得られる、前記学習された誤差特性、及び前記誤差特性に関連付けられたメタデータを入力し、

あらかじめ決められた基準を用いて前記入力した前記ヌクレオチド配列、前記測定プロトコル情報、前記学習された誤差特性、及び前記メタデータを使用して、可能なバイオマーカー配列のセットのための第１のスコアを出力し、

各セットについての前記第１のスコアの値を考慮してバイオマーカー配列セットを決定する、

決定方法。

[請求項12] 前記プロセッサは、

決定すべきバイオマーカー配列ごとに、第２のスコアを入力し、

前記バイオマーカー配列セットにおける各バイオマーカー配列についての前記第１のスコアを考慮して、前記第１のスコアと前記第２のスコアとのバランスを最適化することにより前記多重化パネルのベストなサブセットを選択する、

請求項１１に記載の決定方法。

[請求項13] プロセッサを備え、遺伝子配列の測定誤差特性を予測する予測システムにより実行される予測方法であって、

前記プロセッサは、

多重化パネルで使用する、関心バイオマーカー配列のヌクレオチド配列及び測定プロトコル情報を入力し、

請求項８から１０のいずれか１項に記載の学習方法により得られた、前記学習された誤差特性、及び前記誤差特性に関連付けられたメタデータを入力し、

２つの遺伝子配列間の類似性の尺度を計算するための測定基準を用いて、以前に較正データに含まれていたバイオマーカー配列と新たなバイオマーカー配列との類似度を計算し、

前記計算した類似度を他の関連する入力及び前記学習された誤差特性と組み合わせて使用して、前記較正データに含まれていないバイオマーカー配列を測定する際の誤差特性を予測する、

予測方法。

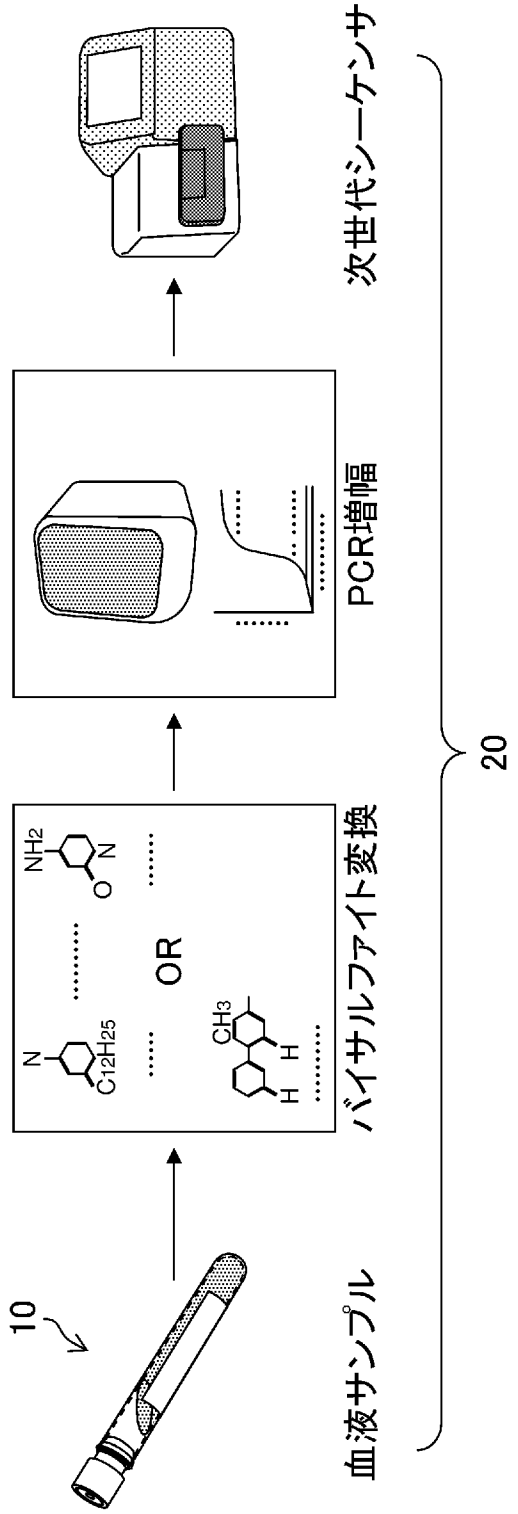
[請求項14] 前記プロセッサは、

前記予測された誤差特性を使用して、前記較正データに含まれていないバイオマーカー配列と最も類似する、前記較正データにおいて利用可能であったバイオマーカー配列を取得し、

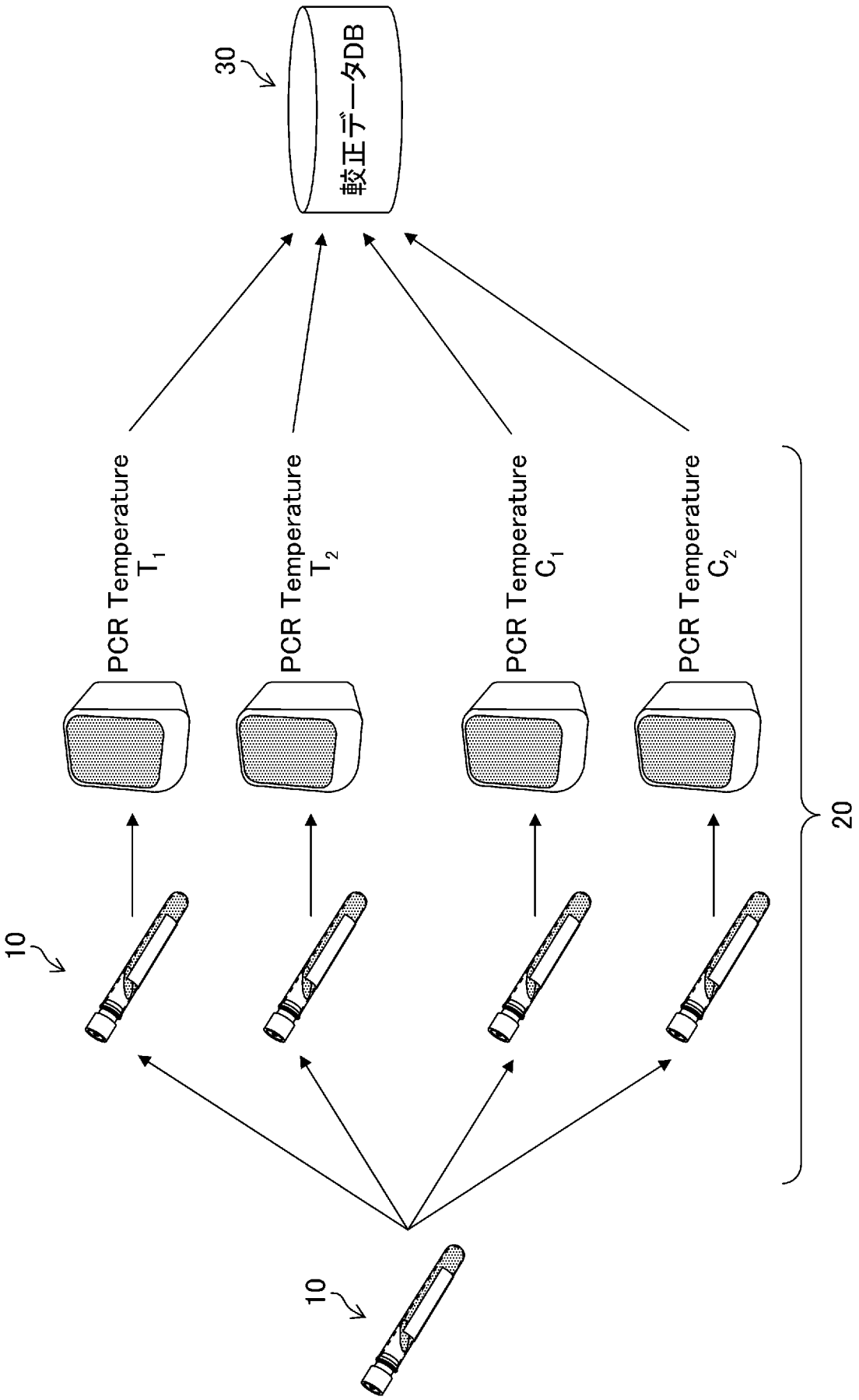
前記取得したバイオマーカー配列の情報を、請求項１１または１２に記載の決定方法におけるバイオマーカー配列セットの決定に反映する、

請求項１３に記載の予測方法。

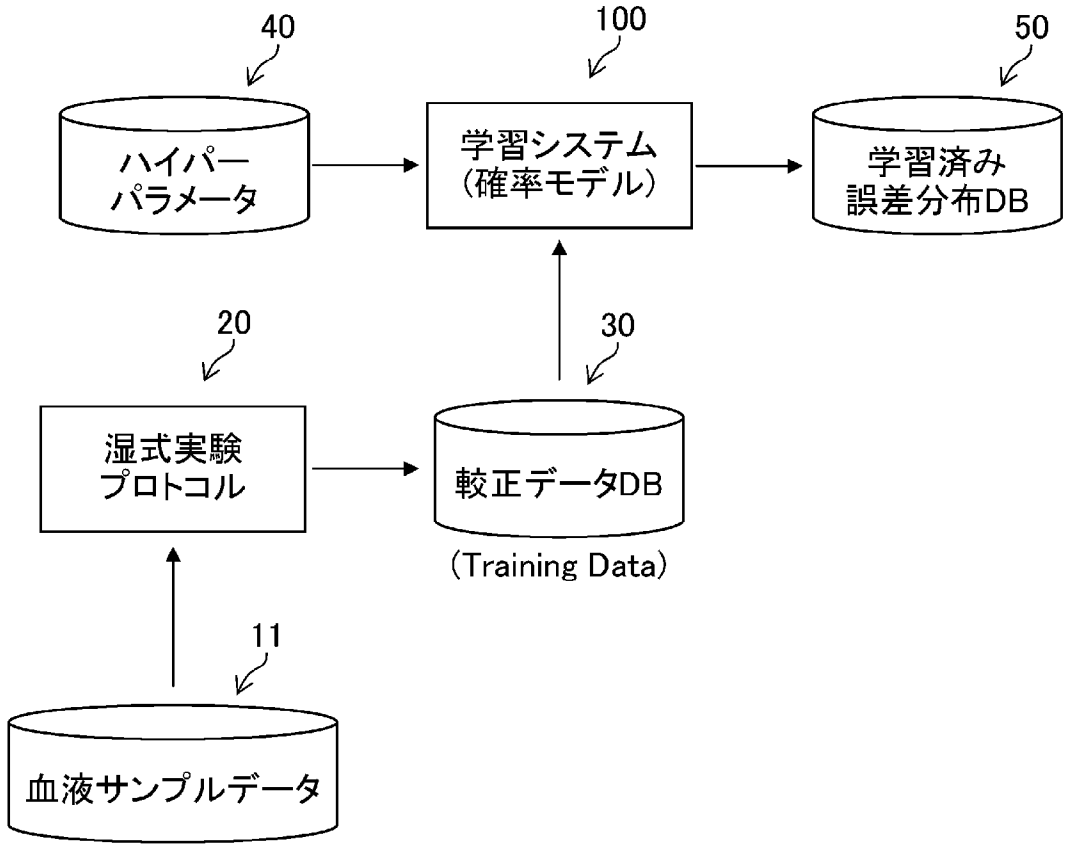
[図1]


次世代シーケンサ
PCR増幅
バイサルファイト変換
血液サンプル
10
20
OR
NH2
N
O
N
C12H25
CH3
H
H

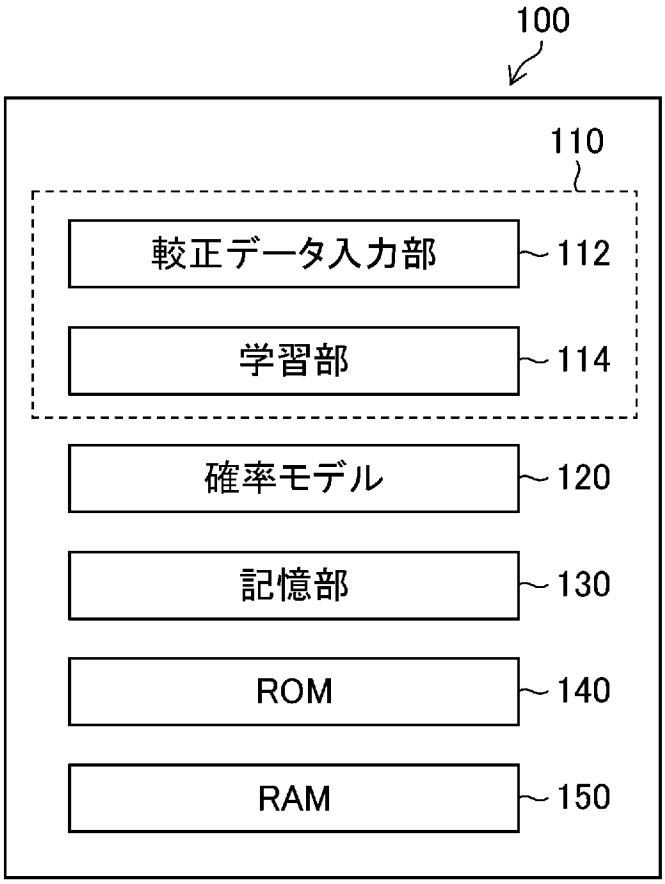
[図2]


30
較正データDB
PCR Temperature T_1
PCR Temperature T_2
PCR Temperature C_1
PCR Temperature C_2
20
10
10

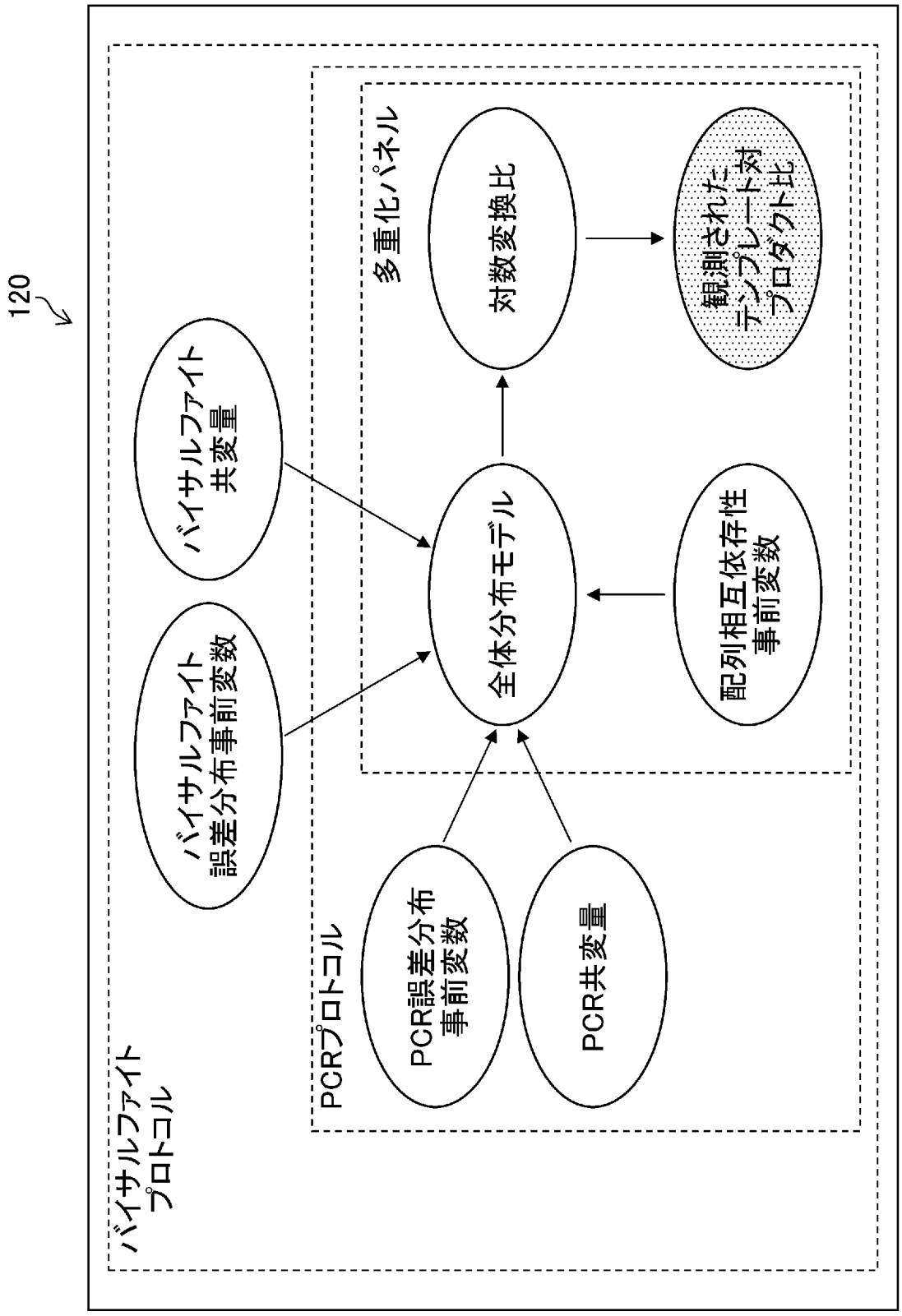
[図3]


40
100
50
ハイパー
パラメータ
学習システム
(確率モデル)
学習済み
誤差分布DB
20
30
湿式実験
プロトコル
較正データDB
(Training Data)
11
血液サンプルデータ

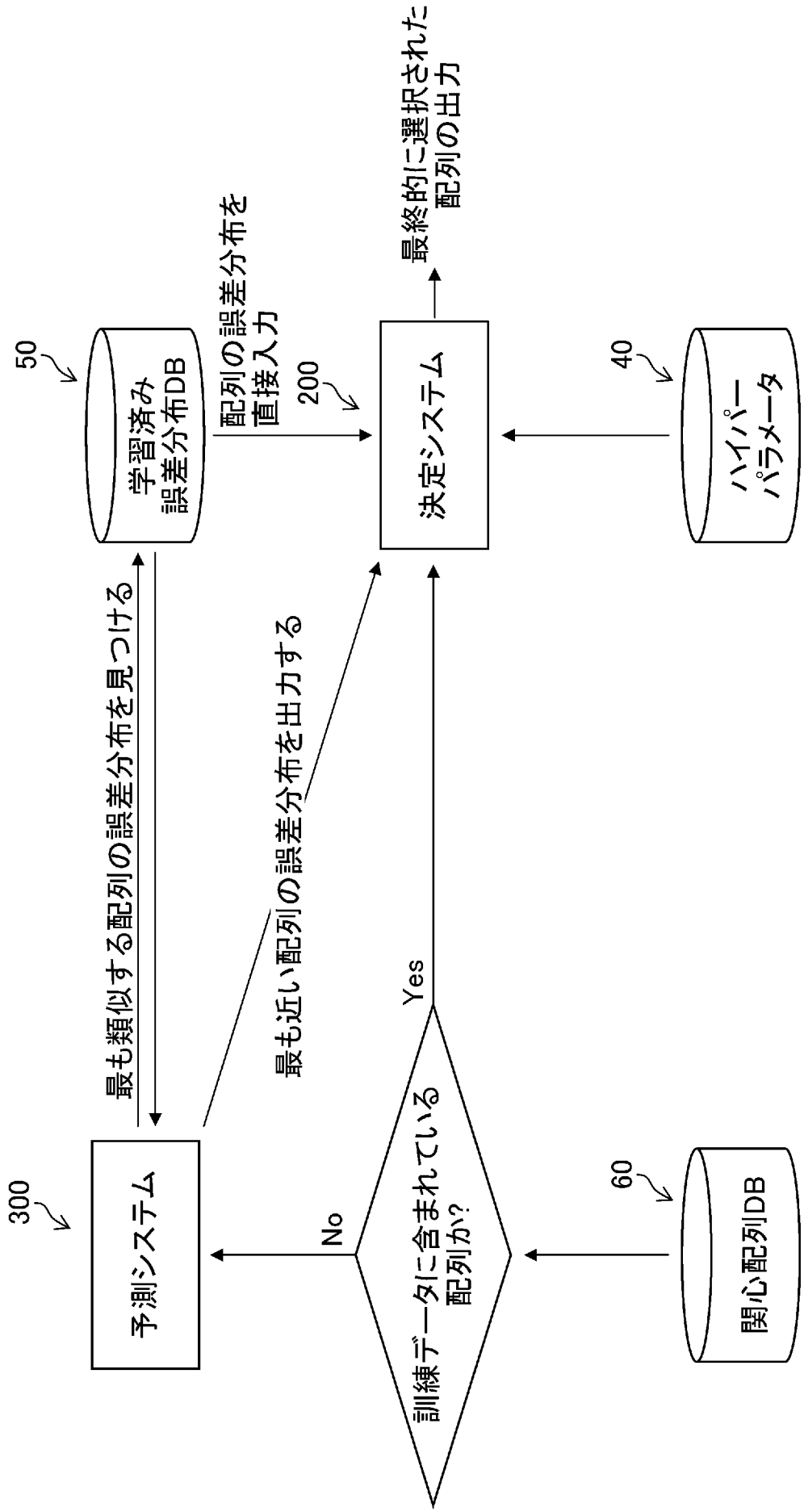
[図4]



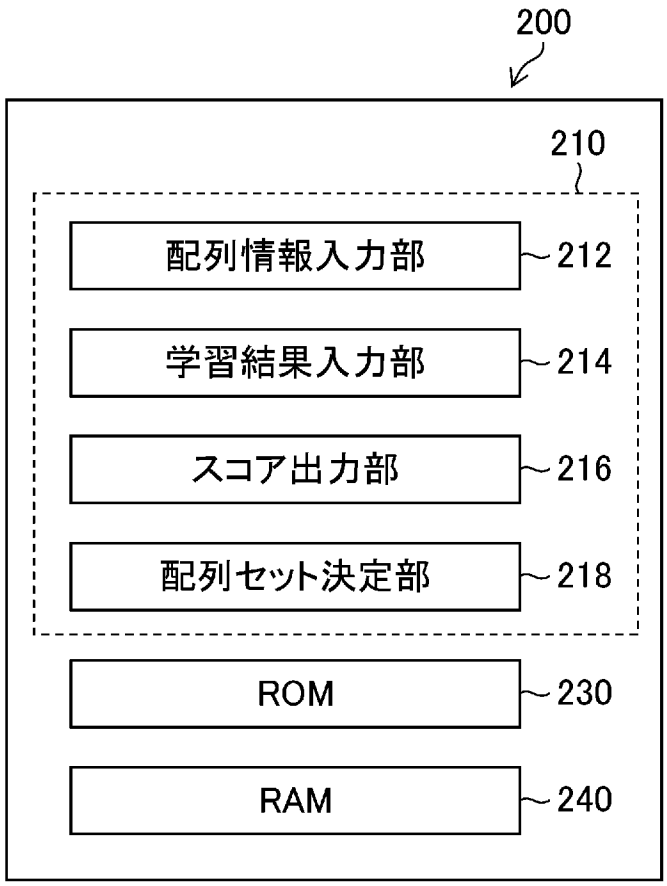
[図5]


120
バイサルファイト
プロトコル
PCRプロトコル
多重化パネル
バイサルファイト
共変量
バイサルファイト
誤差分布事前変数
PCR誤差分布
事前変数
PCR共変量
全体分布モデル
対数変換比
配列相互依存性
事前変数
観測された
テンプレート対
プロダクト比

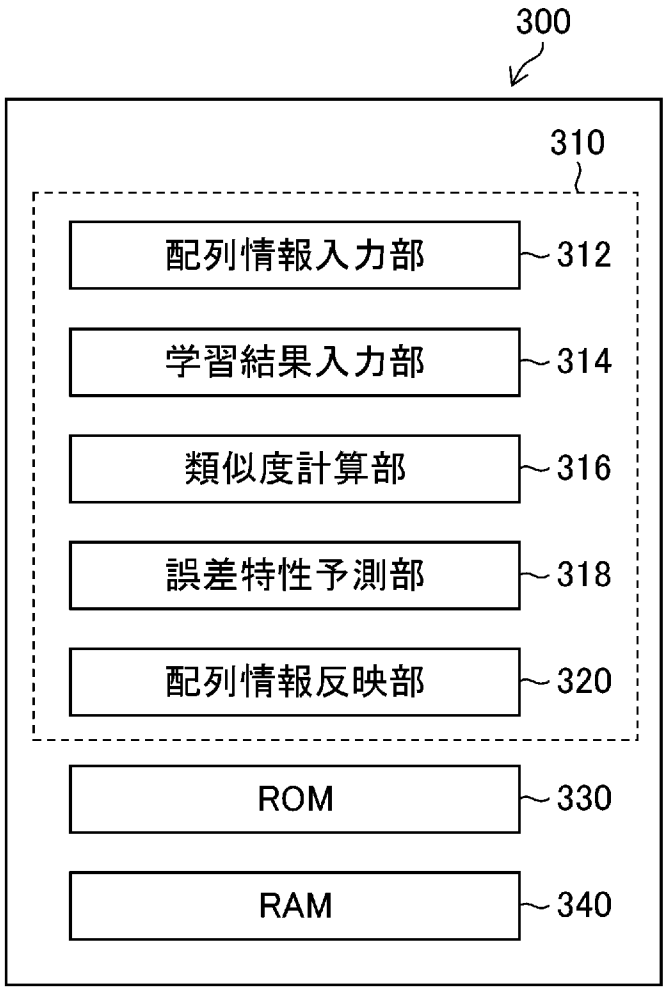
[図6]


50
学習済み
誤差分布DB
配列の誤差分布を
直接入力
200
決定システム
最終的に選択された
配列の出力
40
ハイパー
パラメータ
最も類似する配列の誤差分布を見つける
最も近い配列の誤差分布を出力する
300
予測システム
No
訓練データに含まれている
配列か?
Yes
60
関心配列DB

[図7]


200
210
配列情報入力部
212
学習結果入力部
214
スコア出力部
216
配列セット決定部
218
ROM
230
RAM
240

[図8]


300
310
配列情報入力部
312
学習結果入力部
314
類似度計算部
316
誤差特性予測部
318
配列情報反映部
320
ROM
330
RAM
340

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2023/011772

A. CLASSIFICATION OF SUBJECT MATTER

G16B 40/00(2019.01)i; ***C12M 1/00***(2006.01)i; ***C12M 1/34***(2006.01)i; ***C12N 15/11***(2006.01)i; ***C12Q 1/6844***(2018.01)i; ***C12Q 1/6869***(2018.01)i; ***G01N 33/50***(2006.01)i; ***G06N 20/00***(2019.01)i; ***G16B 30/00***(2019.01)i
FI: G16B40/00; C12M1/00 A; C12M1/34 Z; C12N15/11 Z; C12Q1/6844 Z; C12Q1/6869 Z; G01N33/50 P; G06N20/00; G16B30/00

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)

G16B40/00; C12M1/00; C12M1/34; C12N15/11; C12Q1/6844; C12Q1/6869; G01N33/50; G06N20/00; G16B30/00

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

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2023
Registered utility model specifications of Japan 1996-2023
Published registered utility model applications of Japan 1994-2023

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2021-521536 A (FREENOME HOLDINGS, INCORPORATED) 26 August 2021 (2021-08-26) entire text, all drawings	1-14
A	WO 2020/008192 A2 (CHRONOMICS LIMITED) 09 January 2020 (2020-01-09) entire text, all drawings	1-14

☐ Further documents are listed in the continuation of Box C.
☑ 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
19 May 2023	**30 May 2023**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2023/011772

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP 2021-521536 A	26 August 2021	US 2021/0174958 A1 entire text, all drawings US 2021/0210205 A1 WO 2019/200410 A1 EP 3776381 A1 KR 10-2020-0143462 A CN 112292697 A AU 2019253118 A1 CA 3095056 A SG 11202009696W A	
WO 2020/008192 A2	09 January 2020	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2023/011772

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G16B 40/00(2019.01)i; C12M 1/00(2006.01)i; C12M 1/34(2006.01)i; C12N 15/11(2006.01)i; C12Q 1/6844(2018.01)i; C12Q 1/6869(2018.01)i; G01N 33/50(2006.01)i; G06N 20/00(2019.01)i; G16B 30/00(2019.01)i
FI: G16B40/00; C12M1/00 A; C12M1/34 Z; C12N15/11 Z; C12Q1/6844 Z; C12Q1/6869 Z; G01N33/50 P; G06N20/00; G16B30/00

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

G16B40/00; C12M1/00; C12M1/34; C12N15/11; C12Q1/6844; C12Q1/6869; G01N33/50; G06N20/00; G16B30/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０２３年
日本国実用新案登録公報	１９９６－２０２３年
日本国登録実用新案公報	１９９４－２０２３年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2021-521536 A (フリーノーム・ホールディングス・インコーポレイテッド) 26.08.2021 (2021-08-26) 全文、全図	1-14
A	WO 2020/008192 A2 (CHRONOMICS LIMITED) 09.01.2020 (2020-01-09) 全文、全図	1-14

☐ Ｃ欄の続きにも文献が列挙されている。

☑ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
“O” 口頭による開示、使用、展示等に言及する文献
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの
“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
“Y” 特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
“&” 同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
19.05.2023	30.05.2023

名称及びあて先
日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）
塩田　徳彦 5R 4533
電話番号 03-3581-1101 内線 3562

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2023/011772

引用文献	公表日	パテントファミリー文献	公表日
JP 2021-521536 A	26.08.2021	US 2021/0174958 A1 全文、全図 US 2021/0210205 A1 WO 2019/200410 A1 EP 3776381 A1 KR 10-2020-0143462 A CN 112292697 A AU 2019253118 A1 CA 3095056 A SG 11202009696W A	
WO 2020/008192 A2	09.01.2020	（ファミリーなし）	