

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7596290号
(P7596290)

(45)発行日 令和6年12月9日(2024.12.9)

(24)登録日 令和6年11月29日(2024.11.29)

(51)国際特許分類

F I

G O 1 N 27/62 (2021.01)

G O 1 N 27/62 D

G O 1 N 30/72 (2006.01)

G O 1 N 30/72 A

G O 1 N 30/72 C

請求項の数 21 (全37頁)

(21)出願番号	特願2021-557093(P2021-557093)	(73)特許権者	521339843
(86)(22)出願日	令和2年3月27日(2020.3.27)		ヴェン バイオサイエンシーズ コーポレ ーション
(65)公表番号	特表2022-525427(P2022-525427 A)		アメリカ合衆国 9 4 0 8 0 カリフォル ニア州 サウスサンフランシスコ ツー・ タワー・プレイス 5階 フロア
(43)公表日	令和4年5月13日(2022.5.13)	(74)代理人	100107456
(86)国際出願番号	PCT/US2020/025502		弁理士 池田 成人
(87)国際公開番号	WO2020/205649	(74)代理人	100162352
(87)国際公開日	令和2年10月8日(2020.10.8)		弁理士 酒巻 順一郎
審査請求日	令和5年3月24日(2023.3.24)	(74)代理人	100123995
(31)優先権主張番号	62/826,228		弁理士 野田 雅一
(32)優先日	平成31年3月29日(2019.3.29)	(74)代理人	100123995
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	セリエ, ダニエル
			アメリカ合衆国 9 4 0 8 0 カリフォル ニア州 サウスサンフランシスコ ツー・ 最終頁に続く

(54)【発明の名称】 質量分析法データにおける境界の自動検出

(57)【特許請求の範囲】

【請求項1】

質量分析法のピークを識別する方法であって、
入力パラメータの関数として出力を生成する少なくとも1つの訓練されたニューラルネットワークモデルを提供することと、前記入力パラメータの値が、質量対電荷 (m/z) 比を有するグリコペプチドの保持時間値のシーケンスを含み、
入力パラメータの値として生体サンプルの質量分析法から前記保持時間値のシーケンスにわたる前記 m/z 比を含む質量分析法 (MS) データの値を提供することと、
前記少なくとも1つの訓練されたニューラルネットワークモデルから、前記生体サンプル中の1つ以上の分析物またはそのフラグメントの存在を示す1つ以上のピーク開始保持時間値およびピーク停止保持時間値の場所を含む、前記出力を受け取ることと、
ピーク開始およびピーク停止保持時間値の信頼スコアを識別することと、
前記それぞれの信頼スコアが閾値の値を下回ったことに応答して、ピーク開始およびピーク停止保持時間値の除外されたピーク開始およびピーク停止保持時間値に基づいて、ピーク開始およびピーク停止保持時間値を識別することを含む、
方法。

【請求項2】

前記生体サンプルから識別された前記ピーク開始保持時間値とピーク停止保持時間値との間で保持時間値に対する加算または積分によって1つ以上のグリコペプチドまたはペプチドの存在を数値化することをさらに含む、請求項1に記載の方法。

【請求項 3】

前記数値化することが、識別された前記ピーク開始保持時間値およびピーク停止保持時間値に対して重み付き平滑化関数を適用することを含む、請求項 2 に記載の方法。

【請求項 4】

前記除外されたピーク開始およびピーク停止保持時間の値は、人間による再評価の対象となる、請求項 1 ～ 3 のいずれかに記載の方法。

【請求項 5】

前記質量分析法（MS）データは、タンデムクロマトグラフィの質量分析法データである、請求項 1 ～ 4 のいずれかに記載の方法。

【請求項 6】

システムであって、
質量分析法（MS）データのデータリポジトリと、
前記データリポジトリに結合された 1 つ以上のプロセッサと、
前記 1 つ以上のプロセッサによって実行されるとき、請求項 1 ～ 5 のいずれかの方法を実行する、前記 1 つ以上のプロセッサがアクセスできる非一時的なメモリに常駐する機械実行可能コードと
を備える、システム。

【請求項 7】

方法であって、
未処理 MS データからデータをフィーチャ化することを含む方法であって、前記未処理 MS データは、各々が質量／電荷強度を表す複数のアナログサンプルを含み、前記フィーチャ化することが、前記アナログサンプルの各々に、
前記アナログサンプルを保持時間ウィンドウ内の中心に置くステップと、
前記アナログサンプルを、強度値を表す点のシーケンスに離散化するステップと、
前記強度値を標準化するステップと、
ラベル付きの点のシーケンスを生成するために、ピーク開始時間およびピーク停止時間に対応する前記点のシーケンスの中の点にラベルを割り当てるステップと
前記ラベル付きの点のシーケンスを用いて機械学習モデルを訓練して、生物学的サンプル中に存在する 1 つ以上のグリコペプチドまたはペプチド、またはそのフラグメントの存在を示す質量分析のピーク開始点とピーク停止点を識別するステップとを含む、
方法。

【請求項 8】

前記中心に置くステップが、前記保持時間ウィンドウの中心に曲線の頂点を配置することを含む、請求項 7 に記載の方法。

【請求項 9】

前記強度値を標準化するステップが、閾値量よりも大きい強度値を除外する、請求項 7 または請求項 8 に記載の方法。

【請求項 10】

前記閾値量が 5.00 である、請求項 9 に記載の方法。

【請求項 11】

前記未処理 MS データが、タンデムクロマトグラフィの質量分析法データである、請求項 7 ～ 10 のいずれかに記載の方法。

【請求項 12】

システムであって、
複数のアナログサンプルを含む未処理の質量分析法（MS）データのデータリポジトリと、
前記データリポジトリに結合された 1 つ以上のプロセッサと、
前記 1 つ以上のプロセッサによって実行されるとき、請求項 7 ～ 11 のいずれかの方法を実行する、前記 1 つ以上のプロセッサがアクセス可能な非一時的なメモリに常駐する機械実行可能コードと

10

20

30

40

50

を備える、システム。

【請求項 13】

前記機械学習モデルがリカレントニューラルネットワークモデルである、請求項 7 に記載の方法。

【請求項 14】

前記リカレントニューラルネットワークモデルが、アテンション層を有する双方向 LSTM リカレントニューラルネットワークモデルである、請求項 13 に記載の方法。

【請求項 15】

質量分析法のピーク開始点とピーク停止点が、タンデムクロマトグラフィの質量分析法データからのものである、請求項 7、13 および 14 のいずれかに記載の方法。

10

【請求項 16】

請求項 7、および 13～15 のいずれかに記載のプロセスによって訓練された機械学習モデルを使用して、グリコペプチドまたはペプチドについて存在量を計算することを含む方法。

【請求項 17】

システムであって、
ラベル付きシーケンシャルデータのデータリポジトリと、
前記データリポジトリに結合された 1 つ以上のプロセッサと、
前記 1 つ以上のプロセッサによって実行されるとき、請求項 13～16 のいずれかに記載の方法を実行する、前記 1 つ以上のプロセッサがアクセス可能な非一時的なメモリに常駐する機械実行可能コードと
を備える、システム。

20

【請求項 18】

前記少なくとも 1 つの訓練されたニューラルネットワークモデルが、シーケンシャルモデルであるモデル A、または基準ベースのモデルであるモデル B である、請求項 1 に記載の方法。

【請求項 19】

前記機械学習モデルが、シーケンシャルモデルであるモデル A、または基準ベースのモデルであるモデル B である、請求項 7 に記載の方法。

【請求項 20】

30

方法であって、
入力パラメータの関数として出力を生成する少なくとも 1 つの訓練されたニューラルネットワークモデルを提供することと、前記入力パラメータの値が、質量対電荷 (m/z) 比を有するグリコペプチドの保持時間値のシーケンスを含み、

入力パラメータの値として生体サンプルの質量分析法から前記保持時間値のシーケンスにわたる前記 m/z 比を含む質量分析法 (MS) データの値を提供することと、

前記少なくとも 1 つの訓練されたニューラルネットワークモデルから、前記生体サンプル中の 1 つ以上の分析物またはそのフラグメントの存在を示す 1 つ以上のピーク開始保持時間値およびピーク停止保持時間値の場所を含む、前記出力を受け取ることと、

ピーク開始およびピーク停止保持時間値の信頼スコアを識別することと、
前記それぞれの信頼スコアが閾値の値を下回ったことに応答して、ピーク開始およびピーク停止保持時間値の除外されたピーク開始およびピーク停止保持時間値に基づいて、ピーク開始およびピーク停止保持時間値を識別することと、

40

生物学的サンプルから識別されたピーク開始保持時間値とピーク停止保持時間値の間の保持時間値を加算または積分により、1 つ以上のグリコペプチドまたはペプチドの存在を数値化することを含む、
方法。

【請求項 21】

請求項 20 に記載の方法であって、
強度データのデータリポジトリと、

50

前記データリポジトリに結合された1つ以上のプロセッサと、

前記1つ以上のプロセッサによって実行されるとき、前記方法を実行する、前記1つ以上のプロセッサがアクセス可能な非一時的なメモリに常駐する機械実行可能コードとを備えるシステムによって実行される、方法。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願の相互参照

本願は、全体の内容が、すべての目的のためにその全体として参照により本明細書に組み込まれる、2019年3月29日に提出された米国仮特許出願第62/826,228号の優先権及び利益を主張する。

【0002】

本開示は、一般に、質量分析法を使用する物理的なものの検出に関し、特に、多重反応モニタリング質量分析法(MRM-MS)を使用する生物学的量の自動検出に関する。

【背景技術】

【0003】

多重反応モニタリング(MRM)は、他の分析物の中でも、タンパク質、脂質、及び翻訳後修飾(PTM)の高感度検出及び特異的検出を可能にするタンデム質量分析法(MS)で利用される技術である。この技術は、新しい臨床分析及び臨床診断における使用に大きな見込みがあるが、正確な分析ソフトウェアの可用性はボトルネックのままである。

【0004】

MRM実験の従来の分析は、手でクロマトグラフピークの開始及び停止を選び、データ依存取得技術または他の発見技術によって以前に特徴付けられた遷移から作業することを伴っていた。典型的な実施態様は、強度が、所与の前駆物質及び製品質量対電荷比に対する保持時間(RT)の事前に指定された範囲にわたって描かれるAgilentのMassHunterソフトウェアで確かめることができる。ピークの始まり及び終わりを選択することにより、観察RT、ピーク幅、及び統合存在量値が生じる。評価における個人差、人的エラー、及び大きな時間投資により、この方法は、研究以外での使用には不十分である。近年、多くのソフトウェアパッケージが、遷移のリストからピークを自動的に統合することによってこのギャップを埋めようと試してきた。

【0005】

既存のソフトウェアパッケージは、質量分析法データに高い相対的な存在量が存在するときに、例えばペプチドを数値化する上で非常に有用であるのが判明している。しかし、これらのソフトウェアパッケージは、糖化ペプチドなど、相対的に低い存在量で高度に不均質な種を統合する上では正確ではない。

【発明の概要】

【0006】

本開示は、時系列データの観点から表された、または時系列データから導出された数量の存在の自動検出のための方法及びシステムを説明する。また、本開示は、周波数依存性データの観点から表された、または周波数依存性データから導出された数量の存在の自動検出のための方法及びシステムも説明する。

【0007】

一実施態様では、自動検出は、強度対時間の時系列に適用される。強度は、数量の存在または不在を検出する検出器から生じる。例えば、自動検出は、強度が、例えば保持時間など、時間の関数として遷移値を検出するMS検出器から生じる質量分析法(MS)に適用され得る。したがって、本明細書に説明する自動検出は、時間の関数としてまたは周波数の関数として観察された強度値に基づいて多種多様のデータ入力に適用し得る。

【0008】

一実施態様では、自動検出は、質量分析法(MS)を伴うプロセスから生成された、時間に対比した質量対電荷比の強度の時系列に適用され得る。別の実施態様では、自動検出

10

20

30

40

50

は、液体クロマトグラフィまたはガスクロマトグラフィを伴うプロセスから生成された、時間に対比した強度の時系列に適用される。

【0009】

他の例によれば、自動検出は、時間データもしくは周波数データの観点から表された、または時間データもしくは周波数データから導出された以下の強度のいずれにも適用され得る。

- ・強度が、例えば 2θ など、走査角度が時間の関数として観察される x 線走査角度の関数として x 線の反射を検出する、例えば、電荷結合素子 (CCD) など、x 線検出器から生じる x 線回折。

- ・強度が、例えば cm^{-1} など、光波長の吸収または発射の関数として振動モードを検出する赤外線検出器から生じるラマン分光法。

- ・強度が、紫外光及び可視光の波長または周波数の関数として電子遷移の強度を検出する光検出器から生じる UV-VIS 分光法。発射された蛍光またはリン光の光強度が、入射、つまり吸収されたまたは励起の波長の関数として検出される蛍光定量分光法または燐光分析分光法。

- ・強度が、強力な磁場内での放射周波数の吸収または発射の関数として核遷移を検出する電波検出器から生じる核磁気共鳴 (NMR) 分光法。NMR では、強度対周波数の検出は、例えばフーリエ変換を使用して、同等な強度対時間に変換することができる。また、NMR に適用可能な同じ原理は、磁気共鳴映像法 (MRI) にも適用可能である。

【0010】

1つの特定の実施態様では、自動検出は、質量分析法を伴うプロセスから生成された、時間に対比した質量対電荷比の強度の時系列に適用される。この実施態様によれば、自動検出は、開始ピーク及び終了ピークが MS データで発生する場所を検出するために1つ以上の数学モデルを使用することを伴う。これらの境界点の正確な選択を提供することによって、存在量のより正確な評価が行われる。

【0011】

一態様によれば、質量分析法のピークを識別するように機械学習モデルを訓練する方法は、訓練セットとして、ピーク開始及びピーク停止がラベル付けされる質量／電荷比を表す点のシーケンスを使用する。訓練されたモデルは、生体サンプル中に存在する、1つ以上のグリコペプチドもしくはペプチド、またはそのフラグメントの存在を示す、質量分析法の開始ピーク及び停止ピークを識別するために適合される。

【0012】

別の態様によれば、質量分析法のピークを識別する方法は、入力パラメータの関数として出力を生成する、少なくとも1つの訓練されたニューラルネットワークモデルを使用し、入力パラメータの値は、 m/z (質量対電荷) 比を有するグリコペプチドの保持時間のシーケンスであってよい。

【0013】

別の態様によれば、未処理の MS データは、MS データ中のピーク開始時間及びピーク停止時間を識別するようにモデルを訓練する目的で、機械学習モデル用の訓練セットを生成するためにフィーチャ化される (featureized)。フィーチャ化 (featureization) は、MS データのアナログサンプルを点のシーケンスに離散化することであって、アナログサンプル中のピーク面積に対応する保持時間が点のシーケンスから除外される、離散化すること、及びピーク開始時間及びピーク停止時間に対応する点のシーケンスの中の点にラベルを割り当てて、ラベル付きの点のシーケンスを生成することを含み得る。ラベルは、表示装置及び入力装置を使用してピーク開始時間及びピーク停止時間を選択するように訓練された人間のアノデータを使用して見つけれられる場合がある。

【0014】

別の態様によれば、方法は、所与の m/z 比及び保持期間の MS データを表す点のシーケンスにわたって確率の分布を計算し、確率はありそうなピーク開始保持時間またはピーク停止保持時間を表す。これらの確率の分布から、存在量の統計値が計算される。統計値

10

20

30

40

50

は平均であってよい。

【0015】

別の態様によれば、病気を診断する方法は、質量分析法の遷移を数値化することであって、遷移が、アミノ酸配列を含む、グリコペプチドもしくはペプチド、またはそのフラグメントに特有な m/z 強度値ピークである、数値化すること、及び数値化に基づいてサンプルが病気分類内に入るまたは入らない確率を決定するためにモデルを使用して数値化を分析することを含む。いくつかの例では、アミノ酸配列を含むグリコペプチドもしくはペプチド、またはそのフラグメントは、国際特許出願公開第WO2019/046814号—Identification and use of glycopeptides as biomarkers for diagnosis and treatment monitoring、国際特許出願公開第WO201907639A1号—Identification and use of biological parameters for diagnosis and treatment monitoring、米国特許出願公開第US20190101544A1号—Identification and use of glycopeptides as biomarkers for diagnosis and treatment monitoring、2020年1月31日に出願された国際PCT特許出願第PCT/US2020/0162861号、及び2020年3月13日に出願された米国仮特許出願第62/989,510号に説明されるそれらの配列から選択される。これらの特許出願の各々の全体的な内容は、すべての目的のためにその全体として参照により本明細書に組み込まれる。

【図面の簡単な説明】

【0016】

【図1】機械学習モデルを使用して質量分析法データから存在量を計算するためのプロセスを示す。

【図2】図1の機械学習モデルを訓練するためのプロセスを示す。テストセットA、境界予測での絶対誤差の棒グラフ及びテストセットB、境界予測誤差の累積分布でのモデル性能。19039（28.1%）のサンプルが規則ベースの予測の範囲外（>6秒）にあり、4824（7.1%）のサンプルがシーケンシャルNN内にあり、1470（2.2%）のサンプルが基準ベースのNN内にすることに留意されたい。C、テストセットでのピーク存在量予測／アノテーションの散布図。

【図3】図2の機械学習モデルに対する入力として使用されるMSサンプルから点のシーケンスを生成するためのプロセスを示す。

【図4】MSサンプルのピーク開始保持時間及びピーク停止保持時間の計算した確率に基づいて存在量を計算するためのプロセスである。

【図5A】A及びBは、長・短期記憶装置及びアテンション層を有する双方向リカレントニューラルネットワークである第1の機械学習モデル（モデルA）のモデルアーキテクチャを示す。

【図5B】A及びBは、長・短期記憶装置及びアテンション層を有する双方向リカレントニューラルネットワークである第1の機械学習モデル（モデルA）のモデルアーキテクチャを示す。

【図6A】長・短期記憶装置及びアテンション層を有する双方向リカレントニューラルネットワークである第2の機械学習モデル（モデルB）のためのモデルアーキテクチャを示す。

【図6B】長・短期記憶装置及びアテンション層を有する双方向リカレントニューラルネットワークである第2の機械学習モデル（モデルB）のモデルアーキテクチャを示す。

【図6C】モデルBを使用するサンプルのアテンションマップを示す。図6Cでは、サンプルピーク及び基準ベースモデルの予測は、すべての目的のためにその全体として参照により本明細書に組み込まれる、2020年1月31日に出願された国際特許PCT出願第PCT/US2020/0162861号のSEQ ID1番のMRM遷移、ならびに右下パネル、及び左下パネルのアテンションマップ（モデルから抽出される）に示すその基

10

20

30

40

50

準ピーク用である。

【図 7 A】モデル A 及びモデル B のサンプルラベル付き入力及び計算された確率（境界予測）を示す。星及び三角形のアノテーションは、モデルを訓練するための MS データのラベル付けを表す。緑の曲線はモデル B を使用する予測を表し、青の曲線はモデル A を使用する予測を表す。

【図 7 B】モデル A 及びモデル B のサンプルラベル付き入力及び計算された確率（境界予測）を示す。星及び三角形のアノテーションは、モデルを訓練するための MS データのラベル付けを表す。緑の曲線はモデル B を使用する予測を表し、青の曲線はモデル A を使用する予測を表す。

【図 8 A】同じデータ（規則ベースの方法－赤、1）に関して訓練された既知の機械学習アーキテクチャと対照したモデル A（青、2）及びモデル B（緑、3）の性能を示す。図 10 では、3 つの縦棒の各セットは、規則ベース、MAE 4.79 秒、シーケンシャル NN、MAE 2.33 秒、及び基準ベース NN、MAE 1.56 秒として左から右に提示される。

10

【図 8 B】同じデータ（規則ベースの方法－赤、1）に関して訓練された既知の機械学習アーキテクチャと対照したモデル A（青、2）及びモデル B（緑、3）の性能を示す。図 10 では、3 つの縦棒の各セットは、規則ベース、MAE 4.79 秒、シーケンシャル NN、MAE 2.33 秒、及び基準ベース NN、MAE 1.56 秒として左から右に提示される。

【図 8 C】同じデータ（規則ベースの方法－赤、1）に関して訓練された既知の機械学習アーキテクチャと対照したモデル A（青、2）及びモデル B（緑、3）の性能を示す。図 10 では、3 つの縦棒の各セットは、規則ベース、MAE 4.79 秒、シーケンシャル NN、MAE 2.33 秒、及び基準ベース NN、MAE 1.56 秒として左から右に提示される。

20

【図 9 A】モデル A 及びモデル B の追加のサンプルラベル付き入力及び計算された確率（境界予測）を示す。星及び三角形のアノテーションは、モデルを訓練するための MS データのラベル付けを表す。緑の曲線は、モデル B を使用する予測を表し、青の共線はモデル A を使用する予測を表す。グリコペプチド遷移の基準ピーク及びサンプルピーク。A：基準ピークでは、灰色の線は入力曲線を表し、赤の星及び赤の三角形は人間がアノテートしたピーク開始及び終了であり、青／緑の実線及び破線は、ピーク開始確率及びピーク終了確率のシーケンシャル／基準ベースのニューラルネットワークモデルの予測を示す。B：異なる入力曲線を有する同じ遷移のサンプルピーク。基準ベースのモデル（緑）は、はるかに多い一貫した予測を出力することに留意されたい。

30

【図 9 B】それぞれモデル A 及びモデル B の追加のサンプルラベル付き入力及び計算された確率（境界予測）を示す。星及び三角形のアノテーションは、モデルを訓練するための MS データのラベル付けを表す。緑の曲線は、モデル B を使用する予測を表し、青の共線はモデル A を使用する予測を表す。グリコペプチド遷移の基準ピーク及びサンプルピーク。A：基準ピークでは、灰色の線は入力曲線を表し、赤の星及び赤の三角形は人間がアノテートしたピーク開始及び終了であり、青／緑の実線及び破線は、ピーク開始確率及びピーク終了確率のシーケンシャル／基準ベースのニューラルネットワークモデルの予測を示す。B：異なる入力曲線を有する同じ遷移のサンプルピーク。基準ベースのモデル（緑）は、はるかに多い一貫した予測を出力することに留意されたい。

40

【図 10】人間のアノテータ（灰色、1）対モデル A（青、2）及びモデル B（緑、3）の境界点選択の相対的な標準偏差を示す。縦棒プロットは、13 人の人間のアノテータの相対的な標準偏差（RSD）、及びシーケンシャル／基準ベースのニューラルネットワーク予測の相対誤差（RE）のものである。図 10 では、3 つの縦棒の各セットは、シーケンシャル NN、平均 RE 0.034、基準ベースの NN、平均 RE 0.019、及び人間のアノテータ、平均 = 0.025 として左から右に提示される。

【図 11 A】モデル A とモデル B の間の存在量計算を比較する散布図である。シーケンシャル（A）及び基準ベース（B）のニューラルネットワークのピーク存在量予測／アノテ

50

ーションの散布図。シーケンシャルモデルの予測では、重みを付けた存在量（紫の点）は、グラウンドトゥールズ及び最大点集合存在量（青の点）よりも少ない外れ値を有するより高い一貫性を示し、ピアソンの r は 0.9969 でより高い。重みを付けた存在量が、低存在量範囲で上方に偏向されることに留意すること。基準ベースの予測の間、結果は非常に類似する。

【図 1 1 B】それぞれ、モデル A とモデル B の間の存在量計算を比較する散布図である。シーケンシャル（A）及び基準ベース（B）のニューラルネットワークのピーク存在量予測／アノテーションの散布図。シーケンシャルモデルの予測では、重みを付けた存在量（紫の点）は、グラウンドトゥールズ及び最大点集合存在量（青の点）よりも少ない外れ値を有するより高い一貫性を示し、ピアソンの r は 0.9969 でより高い。重みを付けた存在量が、低存在量範囲で上方に偏向されることに留意すること。基準ベースの予測の間、結果は非常に類似する。

10

【図 1 2】モデル B の信頼スコア、及び信頼スコアを計算するために使用される方程式を示す。信頼スコアの図：青の破線（*）－モデルが予測したピーク終了確率、えんじ色の線（#）－最大点集合予測、明るい赤の線（!）－他の極大値。信頼スコア（0.634）は、黒いボックス内の領域で除算された陰影をつけた領域の面積である。

【図 1 3 A】モデル A について予測信頼性のヒストグラム、境界予測の平均絶対誤差、及び異なる信頼スコアのピンの散布図を示す。A：シーケンシャルニューラルネットワークの予測信頼性のヒストグラム。4つのピンのサンプルの存在量予測は、以下のパネルに示す。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。C、D、E、F：異なる信頼スコアのピン内でのピーク存在量の予測／アノテーションの散布図。確実性が高い方のサンプルが、より優れた相関スコアを達成した。報告された対数－存在量ピアソンの r の値が、異なる遷移からのサンプルの組み合わせに対して計算されることに留意されたい。

20

【図 1 3 B】モデル A について予測信頼性のヒストグラム、境界予測の平均絶対誤差、及び異なる信頼スコアのピンの散布図を示す。A：シーケンシャルニューラルネットワークの予測信頼性のヒストグラム。4つのピンのサンプルの存在量予測は、以下のパネルに示す。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。C、D、E、F：異なる信頼スコアのピン内でのピーク存在量の予測／アノテーションの散布図。確実性が高い方のサンプルが、より優れた相関スコアを達成した。報告された対数－存在量ピアソンの r の値が、異なる遷移からのサンプルの組み合わせに対して計算されることに留意されたい。

30

【図 1 3 C】モデル A について予測信頼性のヒストグラム、境界予測の平均絶対誤差、及び異なる信頼スコアのピンの散布図を示す。A：シーケンシャルニューラルネットワークの予測信頼性のヒストグラム。4つのピンのサンプルの存在量予測は、以下のパネルに示す。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。C、D、E、F：異なる信頼スコアのピン内でのピーク存在量の予測／アノテーションの散布図。確実性が高い方のサンプルが、より優れた相関スコアを達成した。報告された対数－存在量ピアソンの r の値が、異なる遷移からのサンプルの組み合わせに対して計算されることに留意されたい。

40

【図 1 3 D】モデル A について予測信頼性のヒストグラム、境界予測の平均絶対誤差、及び異なる信頼スコアのピンの散布図を示す。A：シーケンシャルニューラルネットワークの予測信頼性のヒストグラム。4つのピンのサンプルの存在量予測は、以下のパネルに示す。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。C、D、E、F：異なる信頼スコアのピン内でのピーク存在量の予測／アノテーションの散布図。確実性が高い方のサンプルが、より優れた相関スコアを達成した。報告された対数－存在量ピアソンの r の値が、異なる遷移からのサンプルの組み合わせに対して計算されることに留意されたい。

【図 1 3 E】モデル A について予測信頼性のヒストグラム、境界予測の平均絶対誤差、及び異なる信頼スコアのピンの散布図を示す。A：シーケンシャルニューラルネットワーク

50

の予測信頼性のヒストグラム。4つのビンのサンプルの存在量予測は、以下のパネルに示す。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。C、D、E、F：異なる信頼スコアのビン内でのピーク存在量の予測／アノテーションの散布図。確実性が高い方のサンプルが、より優れた相関スコアを達成した。報告された対数－存在量ピアソンの r の値が、異なる遷移からのサンプルの組み合わせに対して計算されることに留意されたい。

【図13F】モデルAについて予測信頼性のヒストグラム、境界予測の平均絶対誤差、及び異なる信頼スコアのビンの散布図を示す。A：シーケンシャルニューラルネットワークの予測信頼性のヒストグラム。4つのビンのサンプルの存在量予測は、以下のパネルに示す。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。C、D、E、F：異なる信頼スコアのビン内でのピーク存在量の予測／アノテーションの散布図。確実性が高い方のサンプルが、より優れた相関スコアを達成した。報告された対数－存在量ピアソンの r の値が、異なる遷移からのサンプルの組み合わせに対して計算されることに留意されたい。

10

【図14A】モデルBについて、予測信頼性のヒストグラム、及び境界予測の平均絶対誤差を示す。A：基準ベースのニューラルネットワークの予測信頼性のヒストグラム。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。

【図14B】モデルBについて、予測信頼性のヒストグラム、及び境界予測の平均絶対誤差を示す。A：基準ベースのニューラルネットワークの予測信頼性のヒストグラム。B：境界予測の平均絶対誤差（MAE）は、確実性が高まるにつれ、サンプル上で着実に減少する。

20

【発明を実施するための形態】

【0017】

明細書では、図面及び説明に表示される参照番号は、異なる図の間で対応する要素または類似する要素を指定する。

【0018】

定義

本明細書で使用する場合、以下の単語及び語句は、それらが使用される文脈が明確に別段に示す場合を除き、以下に述べる意味を有すると意図される。

30

【0019】

「遷移」は、質量分析法の質量対電荷比の強度の時系列を指す。本明細書で使用する場合、語句「遷移」または「多重反応モニタリング（MRM）遷移」は、グリコペプチドまたはそのフラグメントがMRM-MSによって検出されるときに観察される質量対電荷（ m/z ）ピークまたは信号を指す。MRM遷移は、前駆物質及びプロダクトイオンの遷移として検出される。遷移検出の時間間隔は、タンデム液体クロマトグラフィ計器の送り速度、及びMRM分析計内の流量に依存する。

【0020】

「生物学的量」は、生体サンプル中の分析物の量またはタンデムクロマトグラフィ計器内の溶質の量を意味するとして理解される。分析物及び溶質は、多糖、ペプチド、グリコペプチド、糖化ペプチド、タンパク質、糖タンパク質、及びそれらのフラグメントを含むが、これに限定されるものではない。

40

【0021】

本明細書で使用する場合、語句「生体サンプル」は、有機体からまたは有機体の流体もしくは組織から導出された、有機体からまたは有機体の流体もしくは組織によって取得された、有機体からまたは有機体の流体もしくは組織から生成された、有機体からまたは有機体の流体もしくは組織から提供された、有機体からまたは有機体の流体もしくは組織から採取された、または有機体からまたは有機体の流体もしくは組織から除去されたサンプルを指す。生体サンプルは、生検によって取得された滑液、全血、血清、血漿、尿、痰、組織、唾液、涙、脊髄液、組織切片（複数可）、組織培養に配置または適合された細胞（

50

複数可)、汗、粘膜、糞便物質、胃液、腹水、羊水、囊胞液、腹膜水、腓液、乳汁、肺洗浄液、骨髓、胃酸、胆汁、精液、膿、水性体液、漏出液など、及び上述の派生物、部分、及び組み合わせを含むが、これらに限定されるものではない。いくつかの例では、生体サンプルは、血液及び／または血漿を含むが、これらに限定されるものではない。いくつかの例では、生体サンプルは、尿または糞便を含むが、これらに限定されるものではない。生体サンプルは、唾液を含むが、これに限定されるものではない。生体サンプルは、組織解剖及び組織生検を含むが、これらに限定されるものではない。生体サンプルは、上記の生体サンプルの任意の派生物または一部を含むが、これに限定されるものではない。

【0022】

本明細書で使用する場合、用語「多糖」は、グリコペプチド、糖タンパク質、糖脂質、またはプロテオグリカンの炭水化物部分など、複合糖質の炭水化物残基を指す。

10

【0023】

本明細書で使用する場合、用語「グリコペプチド」は、それに結合された少なくとも1つの多糖残基を有するペプチドを指す。

【0024】

本明細書で使用する場合、語句「糖化ペプチド」は、多糖残基に結合されたペプチドを指す。

【0025】

本明細書で使用する場合、語句「グリコペプチドフラグメント」または「糖化ペプチドフラグメント」は、糖化タンパク質のアミノ酸配列の部分（であるが、すべてではない）と同じであるアミノ酸配列を有する糖化ペプチド（またはグリコペプチド）を指す。糖化ペプチドは、糖化タンパク質から、例えば1つ以上のプロテアーゼ（複数可）による消化によって、または例えばMRM-MS計器内でのイオンフラグメンテーションなどのフラグメンテーションによって得られる。MRMは、多重反応モニタリングを指す。

20

【0026】

本明細書で使用する場合、語句「グリコペプチドをフラグメント化する」は、MRM-MS計器内で発生するイオンフラグメンテーションのプロセスを指す。フラグメンテーションは、同じ質量を有するが、その電荷に関して変化する様々なフラグメントを生成し得る。

【0027】

本明細書で使用する場合、語句「多重反応モニタリング質量分析法(MRM-MS)」は、生体サンプル中の多糖及びペプチドのターゲット数値化のための高感度かつ選択的な方法を指す。従来の質量分析法と異なり、MRM-MSはきわめて選択的であり（ターゲットを定めた）、研究者が計器を微調整して、関心のある特定のペプチドフラグメントを具体的に探すことを可能にする。MRMは、潜在的なバイオマーカなど、関心のあるペプチドフラグメントのより大きい感度、特異性、速度、及び計量を可能にする。MRM-MSは、三連四重極(QQQ)質量分析計及び四重極飛行時間(qTOF)質量分析計の1つ以上を使用することを伴う。

30

【0028】

本明細書で使用する場合、「ペプチド」は、特に明記しない限りグリコペプチドを含むことを意味する。

40

【0029】

「保持時間(RT)」は、溶質がクロマトグラフィカラムを通過するのに要する時間を意味すると理解される。RTは、注入から検出までの時間として計算される。RTは、ガス流量、炉とカラム内との温度差、カラム劣化、カラム長さ、カラムタイプ、キャリアガスを含むが、これらに限定されないいくつかの要因に依存する。

【0030】

質量分析法(MS)データまたはデータセットは、遷移数の関数として遷移の強度を描画するスペクトルの集合体を意味すると理解される。所与の遷移値のMSデータは、保持時間(RT)である期間(t)にわたる電荷(z)強度レベルに比較した質量(m)

50

(m/z)である。

【0031】

「存在量」は、質量分析法データの中で表されるなんらかの生物学的量の計算された量を意味すると理解される。存在量は、開始ピーク境界と停止ピーク境界との間の離散時点での値を合計することによって計算し得る。また、存在量は、開始ピーク境界と停止ピーク境界との間の離散時間ステップにわたる積分によっても計算し得る。積分は、任意の標準的な方法、例えば、 $\sum (t_{i+1} - t_i) * (1/2) * [I(t_i) + I(t_{i+1})]$ 、 $i = 1 \dots n-1$ (n は、境界間の離散点の数であり、 $I(t_i)$ は、保持時間 t_i での質量対電荷強度である)によって実行され得る。特別の定めのない限り、存在量は、別のMRMピークに対する1つの統合されたMRMピークの相対存在量を指す。

10

【0032】

非線形性により、最小値を見つけることに基づいて最適化された関数／モデルパラメータのセットは存在せず、むしろ最適化された関数／モデルパラメータは極小値にのみ常駐する(極大値は存在しない)という意味で、本発明の態様を実践するために開発され得る数学モデルは非線形数学モデルに制限される。企図されたモデルアーキテクチャは、人工ニューラルネットワーク、グラフモデル、及びサポートベクトルマシンを含むが、これらに限定されない。

【0033】

「機械学習モデル」または「学習モデル」は、(1つ以上のアフィン変換に加えて)少なくとも1つの非線形活性化層を有する任意の種々の数学モデルを意味すると理解される。機械学習モデルは、さらに、モデル自体とは別個である1つ以上の損失関数の最小化(例えば、交差エントロピー損失または負の対数尤度の最小化)によって訓練または最適化される。訓練または最適化のプロセスは、既知の結果(低バイアス)を再現するためにモデルを最適化し、モデルが目に見えない経験(低分散)から正確な予測を行うことを可能にしようとする。モデルの出力は、例えば、確率または統計値、分類、シーケンスなど、タスクに関連する任意の種々のものであってよい。1つの特定の例では、モデルを訓練及び最適化するために使用するデータは、ターゲットの結果が既知である時系列データなどのシーケンシャルデータであり、既知のターゲット出力を有するそのシーケンシャルデータは、ラベル付きのシーケンシャルデータまたはアノテーション付きのシーケンシャルデータを意味すると解釈される。

20

30

【0034】

ラベル付きのシーケンシャルデータに基づいて予測を行うために適した周知のモデルの一例は、長・短期記憶(LSTM)装置またはゲートリカレントユニット(GRU)を有し、任意選択でアテンションメカニズムを有する、一方向、または双方向のリカレントニューラルネットワーク(RNN)である。アテンションメカニズム(またはアテンション層)は、将来の状態予測に影響を与えるために、過去の状態から蓄積した知識を表す双方向RNNとともに、シーケンスの最終的なLSTM装置もしくはGRU、またはLSTM装置もしくはGRUの組み合わせで考えられる得るものよりもより重要な情報を過去の状態変数から保持し得る。アテンションメカニズムは、学習した重みを入力状態の各々に適用して、出力予測を生成するときにシーケンスに表示される入力されたMSデータ点のより重要ではないまたはより重要な特徴を強調する。これらの学習された重みは、所望の出力に応じて、対応する出力状態、または最終的な出力状態の各々に多かれ少なかれ直接的に結合される。

40

【0035】

ラベル付きデータだけを使用して訓練及び最適化されたモデルは、教師ありモデルとして知られる。また、学習モデルは、教師なしモデルまたは半教師ありモデルであってもよい。これらのモデルは、ラベルなしデータまたはラベル付きデータとラベルなしデータとの組み合わせを使用し、訓練及び最適化される。ラベル付きデータとラベルなしデータの組み合わせまたはおもにラベルなしデータを使用して訓練及び最適化されたモデルは、それぞれ半教師ありモデル及び教師なしモデルとして知られる。学習モデルは、確率的勾配

50

降下法（SGD）の周知の手法によって効率的に訓練及び最適化され得る。学習モデルのハイパーパラメータは、訓練及び／またはバリデーションの間に手動の調整によって最適化される場合もあれば、学習モデルのグローバル損失関数の最小化と並行して、ハイパーパラメータの1つ以上に関して代理損失関数を最小限に抑える方法によって学習される場合もある。

【0036】

「訓練セット」は、既知のターゲット出力を再現するために学習モデルでパラメータを訓練または調整するために使用されるデータセットである。

【0037】

「バリデーションセット」は、ラベル付きデータ（低バイアス）に一致することと、目に見えない遷移（低分散）の開始／停止場所を予測することとの間の最適なバランスに到達するために、新しい経験と対照してモデルの予測機能を調整するために使用されるデータセットである。バリデーションセットは、モデルが訓練セットに基づいて調整された後に適用される。

【0038】

「テストセット」（または除外バリデーションセット）は、最適化されたモデルの予測能力を評価するために使用されるデータセットである。テストセットと対照したモデルの性能は、新しい目に見えない遷移の正確な予測を行うモデルの能力を評価するために使用されるメトリクスである。

【0039】

「フィーチャ化」は、入力として使用するための分布から、サンプルを表す未処理データを機械学習モデルに変換することと関連するプロセスを意味すると理解される。フィーチャ化の例は、アナログ信号のデジタル信号への離散化、データの部分のフィルタリング、またはノイズ除去、またはクリッピング（例えば、信号対雑音比の低減）、信号の正規化及び信号の標準化（例えば、ウィンドウの中心にターゲットピークを置くこと）を含む。また、フィーチャ化は、一般化のためにサンプルのモデルまたは関連する増強を確認する、及び／または新しい（または目に見えない）経験から予測を行うモデルの能力を高めるためにデータセットに雑音を導入することを含んでもよい。

【0040】

本開示によれば、訓練データセット、検証データセット、及び試験データセットがその中からサンプリングされたシーケンシャルデータのデータセットは、遷移のために少なくともラベル付きの開始場所及び停止場所を含み得、その場所の各々は、境界点の間で積分により存在量を計算するための境界点を示す。好ましい実施形態では、ターゲットとされる出力は、遷移のシーケンシャルデータの中の各点の開始場所及び停止場所の確率である。

【0041】

説明

上述のように、MRM実験の従来の分析は、手でクロマトグラフのピークの開始及び停止を選び、データ依存取得技術または他の発見技術によって以前に特徴付けられた遷移から作業することを伴う。AgilentのMassHunterソフトウェアなどのソフトウェアスイートは、所与の前駆物質及び製品質量対電荷比について保持時間（RT）の事前に指定された範囲にわたって信号強度を描画する。ピークの始まり及び終わりを手動で選択することによって、観察RT、ピーク幅、及び統合された存在量値が生じる。しかしながら、評価における個人差、人的エラー、及び大きな時間投資により、この方法は、研究以外での使用には不十分である

【0042】

明確に定義された単一のピーク及び高い信号対雑音比によって特徴付けられた遷移の場合、存在量を計算するために境界点を選ぶ上記の手動のプロセスは効果的である場合があるが、この方法は、低い信号対雑音比によって特徴付けられた生物学的量の存在量を検出するとき、正確な結果を提供しない、または一貫性のある結果を生じさせない場合がある。

【0043】

図1を参照すると、本開示によれば、指定された保持時間ウィンドウ及び質量／電荷強度についてマススペクトルで境界点を自動的に選択するために、訓練された機械学習モデルが代わりに使用される。機械学習モデルに対する入力、点のシーケンス（MSデータのサンプル）である。各点は、対応する保持時間の質量／電荷強度を表す。

【0044】

図2を参照すると、一実施態様によれば、機械学習モデルは、入力として提供された対応する点の各々のピーク開始時間及びピーク停止時間の確率を生成するニューラルネットワークである。ニューラルネットワークは、ラベル付きのターゲットピーク開始保持時間及びピーク停止保持時間を有するマススペクトルを使用して訓練される。訓練及びバリデーションのプロセスを通じて、最適なモデルパラメータが見つけられる。ラベル付きデータは、グラフィカルユーザーインターフェース及び入力装置を使用してピーク開始時間／停止時間を選択する人間のデータから取得されてもよい。

【0045】

図3を参照すると、図1の機械学習モデルへの入力のための未処理のマススペクトルデータをフィーチャ化するためのプロセスの流れがある。フィーチャ化ステップは、（1）特定の閾値量を超える質量／電荷の強度を削除すること、または代わりにほとんど雑音であり、MSデータの信号対雑音遷移領域に近いそれらの点だけを選ぶこと、及び（2）シーケンシャルデータを生成するためにアナログ信号を点のセットに離散化することを含み得る。

【0046】

図4を参照すると、1つの特定の実施態様では、モデルは、積分の加重和に基づいて存在度の量を計算する。最も直接的な方法の1つは、機械学習モデルの予測結果を確率分布として解釈することである。言い換えると、モデルに、ピーク開始／終了マーカとしての機能を果たすであろうシーケンス中の点のすべての考えられる対の確率を評価させることである。N個の点のシーケンスの場合、これは、順序の制約を課した後に、合計で $N \times (N - 1) / 2$ 対を含む。各対の確率は、2つの点でピーク開始／終了タスクの予測値を乗算することによって容易に計算できる。確率は、次に、対応する対の間で積分に対して加重和を計算するために使用され、真の境界の最高の確率を有する2つの端点から計算される単一の存在度の値の代わりに、平均推定値を提供する。存在量の平均は、よりロバストな存在量推定値を提供し得る。

【0047】

実施形態

いくつかの例では、複数のラベル付きシーケンシャルデータを使用することであって、複数のラベル付きシーケンシャルデータの各々が、グリコペプチドもしくはペプチド、またはそのフラグメントに特有な保持時間値を表すピーク開始ラベル及びピーク停止ラベルを含む、使用すること、及びラベル付きシーケンシャルデータを使用して少なくとも1つの機械学習モデルを訓練することであって、訓練されたモデルが、生体サンプルに存在する、1つ以上のグリコペプチドもしくはペプチド、またはそのフラグメントの存在を示す質量分析法の開始ピーク及び停止ピークを識別するために適合された、訓練することを含む、質量分析法のピークを識別するようにモデルを訓練する方法が、本明細書に説明されている。

【0048】

上記のいずれかを含むいくつかの例では、方法は、生体サンプルから質量分析法の開始ピークと停止ピークとの間で経時的に積分によって1つ以上のグリコペプチドまたはペプチドの存在を数値化することをさらに含む。

【0049】

上記のいずれかを含むいくつかの例では、機械学習モデルはリカレントニューラルネットワークモデルである。

【0050】

上記のいずれかを含むいくつかの例では、リカレントニューラルネットワークは、アテ

10

20

30

40

50

ンション層を有する双方向 L S T M リカレントニューラルネットワークモデルである。

【0051】

いくつかの例では、本明細書に説明するプロセスを実行するように訓練された機械学習モデルを使用して、グリコペプチドまたはペプチドの存在量を計算することを含む方法が、本明細書に説明されている。

【0052】

いくつかの例では、ラベル付きシーケンシャルデータのデータリポジトリ、リポジトリに結合された1つ以上のプロセッサ、1つ以上のプロセッサによって実行されるとき、本明細書に説明する方法を実行する、1つ以上のプロセッサがアクセス可能な非一時的なメモリに常駐する機械実行可能コードを含むシステムが、本明細書に説明されている。

10

【0053】

いくつかの例では、入力パラメータの関数としての出力を生成する少なくとも1つの訓練されたニューラルネットワークモデルを提供することであって、入力パラメータの値が m/z (質量対電荷) 比を有するグリコペプチドの保持時間値のシーケンスを含む、提供すること、及び入力パラメータの値として、生体サンプルの質量分析法の保持時間値のシーケンスに対する質量対電荷 (m/z) 比を含む質量分析法 (MS) データを提供することを含み、出力が、生体サンプル中の1つ以上の分析物またはそのフラグメントの存在を示す1つ以上のピーク開始保持時間値及びピーク停止保持時間値の場所を含む、質量分析法のピークを識別する方法が、本明細書に説明されている。

【0054】

上記のいずれかを含むいくつかの例では、方法は、生体サンプルから質量分析法の開始ピークと停止ピークとの間の保持時間値に対する加算または積分によって1つ以上のグリコペプチドまたはペプチドの存在を数値化することをさらに含む。

20

【0055】

上記のいずれかを含むいくつかの例では、方法は、生体サンプルから質量分析法の開始ピーク及び停止ピークの間の保持時間値に対する加算によって1つ以上のグリコペプチドまたはペプチドの存在を数値化することをさらに含む。上記のいずれかを含むいくつかの例では、方法は、生体サンプルから質量分析法の開始ピーク及び停止ピークの間の保持時間値に対する積分によって1つ以上のグリコペプチドまたはペプチドの存在を数値化することをさらに含む。

30

【0056】

上記のいずれかを含むいくつかの例では、数値化するステップは、予測された開始ピーク及び停止ピークに対して重み付き平滑化関数を適用することを含む。

【0057】

いくつかの例では、質量分析法 (MS) データのデータリポジトリ、リポジトリに結合された1つ以上のプロセッサ、1つ以上のプロセッサによって実行されるとき、本明細書に説明する方法を実行する、1つ以上のプロセッサがアクセス可能な非一時的なメモリに常駐する機械実行可能コードを含むシステムが、本明細書に説明されている。

【0058】

いくつかの例では、未処理の MS データからデータをフィーチャ化することであって、未処理の MS データが、各々が質量／電荷強度を表す複数のアナログサンプルを含む、フィーチャ化することを含み、フィーチャ化することが、アナログサンプルの各々について、保持時間ウィンドウ内の中心にサンプルを置くステップと、サンプルを、強度値を表す点のシーケンスに離散化するステップと、強度値を標準化するステップと、ラベル付きの点のシーケンスを生成するために、ピーク開始時間及びピーク停止時間に対応する点のシーケンスの中で点にラベルを割り当てるステップと、を含み、ラベル付きの点のシーケンスが、目に見えない MS データの存在量を予測するように機械学習モデルを訓練するために構成される、方法が本明細書に説明されている。

40

【0059】

上記のいずれかを含むいくつかの例では、中心に置くことは、保持ウィンドウの中心に

50

曲線の頂点を配置することを含む。

【0060】

上記のいずれかを含むいくつかの例では、強度値を標準化することは、閾値量よりも大きい強度値を除外する。上記のいずれかを含むいくつかの例では、閾値量は約500である。

【0061】

いくつかの例では、複数のアナログサンプルを含む未処理の質量分析法(MS)データのデータリポジトリ、リポジトリに結合された1つ以上のプロセッサ、1つ以上のプロセッサによって実行されるとき、いずれかの方法を実行する、1つ以上のプロセッサがアクセス可能な非一時的なメモリに常駐する機械実行可能コードを含むシステムが、本明細書に説明されている。いくつかの例では、所与の m/z 比及び保持時間期間の質量分析法(MS)データを表す点のシーケンスにわたって確率の分布を計算することであって、確率がありそうなピーク開始保持時間またはピーク停止保持時間を表す、計算すること、及び確率の分布を使用して存在量の統計値を計算することであって、統計値がMSデータの中の生体サンプルの存在を示す、計算することを含む方法が、本明細書に説明されている。

【0062】

上記のいずれかを含むいくつかの例では、統計値は、積分を評価するための開始点及び終了点のすべての考えられる対を使用して、積分の加重和として計算された平均存在量である。

【0063】

上記のいずれかを含むいくつかの例では、機械学習モデルは、シーケンシャルモデル(モデルA)または基準ベースモデル(モデルB)である。

【0064】

いくつかの例では、複数のラベル付きシーケンシャルデータを使用すること、及びラベル付きシーケンシャルデータを使用して少なくとも1つの機械学習モデルを訓練することであって、訓練されたモデルが、強度の開始ピーク及び停止ピークが目に見えないサンプルで発生する期間または周波数範囲を識別するために適合された、訓練することを含む方法が、本明細書に説明されている。

【0065】

上記のいずれかを含むいくつかの例では、開始ピーク及び停止ピークは、雑音から信号への遷移期間に含まれる。

【0066】

上記のいずれかを含むいくつかの例では、機械学習モデルは、アテンション層を有するリカレントニューラルネットワークである。

【0067】

上記のいずれかを含むいくつかの例では、強度の開始ピークから停止ピークまでの時間または周波数に対する加法または積分により、物理的なものが数値化されるように、サンプルは物理的なものを表す。

【0068】

上記のいずれかを含むいくつかの例では、サンプルは、指定された期間にわたって質量対電荷強度の連続するシリーズであり、加法または積分は物理的なものの存在量を表す。

【0069】

いくつかの例では、入力パラメータの関数として出力を生成する少なくとも1つの訓練されたニューラルネットワークモデルを提供することであって、入力パラメータの値が、数量を表す強度値の時系列またはスペクトル系列である、提供すること、及び入力パラメータの各々に値を提供することであって、値は数量の目に見えない強度のシーケンスを表す、提供することを含み、出力は、数量を数値化するために使用される1つ以上のピーク開始及びピーク停止の時間または周波数を含む方法が、本明細書に説明されている。

【0070】

上記のいずれかを含むいくつかの例では、数量は、それぞれピーク開始の時間または周

10

20

30

40

50

波数とピーク停止の時間または周波数との間の時間または周波数に対する加算または積分によって数値化される。

【0071】

上記のいずれかを含むいくつかの例では、強度は、質量分析法、MRM、液体クロマトグラフィもしくはガスクロマトグラフィ、x線回折、ラマン分光法、UV-VIS分光法、蛍光定量分光法もしくは蛍光分析分光法、核磁気共鳴（NMR）分光法、または磁気共鳴映像法（MRI）のセットから選択されたプロセスから取得される。

【0072】

実施例

以下の実施例は、本開示のいくつかの態様に従って実施された研究を説明する。

10

【0073】

この研究では、2つの学習モデルが、訓練され、検証され、次に、ピークの場所を予測する際の精度（またはより正確には、関連付けられた遷移の存在量をより正確に計算するためにピークの開始／停止境界）を評価するために試験された。学習モデルのために選ばれた特定のアーキテクチャは、長・短期記憶装置及びアテンション層を有する双方向リカレントニューラルネットワーク（RNN-LSTM-AL）であった。モデルアーキテクチャ及び訓練方法は、図5A、図5B、図6A、図6B、及び図6Cに示されている。モデルの入力パラメータまたは入力状態変数は、質量対電荷比（ m/z ）強度を表す時系列に対応する。

【0074】

20

研究は、ペプチドと無傷のグリコペプチドの両方を含む広いダイナミックレンジに及ぶ質量分析法（MS）遷移からの手動でラベル付けしたピークの大きいデータセットにアクセスした。データセットは、各遷移の保持時間（RT）内の強度の時系列を含んでいた。これらの遷移の各遷移は、RNN-LSTM-ALに入力されたシーケンシャルデータであり、訓練遷移、バリデーション遷移、及び試験遷移の各遷移とともに含まれる、関連する人間によりアノテーションを付けられた（またはラベルを付けられた）（存在量を導出するために積分のための境界点を示す）開始RT及び停止RTは、それぞれ教師あり学習、バリデーション、及び試験のために使用されるラベルである。

【0075】

2.1 データセット

30

2つの質量分析法データセットが研究のために使用された。第1の実験用の未処理のAgilent QQQ MS出力は、210のヒト血清サンプル内で評価された716の遷移から成り立っていた。これにより、正の存在量及びRTシフトについてフィルタリング後、106355の有効なサンプルピークが生じた。このセットは、572の遷移を有する訓練セット、及び144のピークを有するバリデーションセットに分割された。

【0076】

第2の実験は、商用のバイオリポジトリから購入した135のヒト血清サンプルから生成されたMS出力を使用した。503の遷移が評価され、試験専用で使用された67672の高品質のピークが生じた。313のテストセット遷移は、訓練セット／バリデーションセットに存在していなかった。313のテストセット遷移は、モデル出力予測を比較するために使用された。

40

【0077】

2.2 ベースライン方法

訓練され、確認されたRNN-LSTM-ALモデルの性能は、2つのベースライン予測因子に比較された。これらのベースライン予測因子の第1の予測因子は、多重反応モニタリングを含む質量分析法アプリケーションのデータ分析に使用される周知のオープンソースのソフトウェアツールであるSkylineである。Skylineソフトウェアは、テストサンプルのピーク面積（存在量）を計算する。RNNモデルとの比較用の第2のベースライン予測因子は予測の決定論的方法であり、以後「規則ベースの」方法と呼ぶ。規則ベースの方法のアーキテクチャは、Ken Aoshima、Kentaro Tak

50

ahashi, Masayuki Ikawa, Takayuki Kimura, Mitsuru Fukuda, Satoshi Tanaka, Howell E Parry, Yuichiro Fujita, Akiyasu C Yoshizawa, Shin-ichi UtsunomiyaらのA simple peak detection and label free quantitation algorithm for chromatography-mass spectrometry, BMC bioinformatics, 15 (1) : 376, 2014に説明されている。規則ベースの方法モデルは、そのモデルパラメータを調整して、訓練セット及びバリデーションセットの性能を最大限にすることによってこの研究のために開発された。そしてこれらの調整されたパラメータは、次にテストセットの評価中に適用された。

【0078】

2.3 ピーク積分のためのシーケンシャルモデル

RNN-LSTM-ALアーキテクチャは、研究においてMSデータオンシーケンスから予測するために選択された。

【0079】

研究で使用された2つのRNN-LSTM-ALモデルのうちの第1のモデル（以後、「モデルA」または「シーケンシャルモデル」）の訓練／予測プロセス及び流れの視覚的なブロック説明は、図5A及び図5Bに示されている。入力（時間の関数としての質量／電荷比の強度）は、別々の位置を関連付けるためにbi-LSTM層及びアテンションメカニズムを通して符号化される。予測は点単位で行われ、あらゆる入力について、モデルが出力を予測することを意味する。モデルは、 $N \times 2$ の次元性で出力のシーケンスを生成するように訓練された。ここで、「N」は、モデルAの入力パラメータの数（サンプル中の曲線の性質に応じて50～300の範囲の、強度値がモデルに入力された保持時間に対応する点、以下5.3を参照）、シーケンス中のデータポイントの数、及び各入力値に対応する出力値である。N個の点の各々について出力された二つ（2）の値は、対応する点がピーク開始またはピーク停止を表す確率（0から1の範囲）である。所与の遷移について値が1（高確率）に最も近いモデルから出力された1つの点または点の近傍は、次に存在量積分（継続的に積分された）が計算される間隔の境界を定めるために使用される。

【0080】

モデルからの出力確率は、このようにして境界の分布を形成する。モデルは、予測された境界とグラウンドトゥールズとの間の交差エントロピー損失に対する勾配降下法、または手作業でラベル付けしたピーク開始及びピーク停止によってエンドツーエンドで訓練される。例えば、モデルのハイパーパラメータに関して最小限に抑えられた目的関数に適用されたベジアン最適化など、最小ハイパーパラメータ検索は、優れた一般化を達成し、訓練セットに対する過剰適合を回避する目的で、除外バリデーションセットまたはテストセットで性能を最適化するために適用される。フィーチャ化及びハイパーパラメータは、以下により詳細に説明する。

【0081】

2.4 基準ベースのシーケンシャルモデル

いくつかの用途では、遷移ピークの高度に可変的な形状及び幅のため、モデルAなどの統一した予測モデルは、目に見えない遷移に対して十分に一般化しない場合がある。同じ遷移での異なる患者のサンプルの形状が一貫している旨の観察に基づいて、第2のRNN-LSTM-ALモデル（以後「モデルB」または「基準ベースのモデル」）も開発された。

【0082】

モデルBの訓練／予測プロセス及び流れの視覚的なブロック説明は、図6A、図6B、及び図6Cに示されている。モデルBは、入力としてそれぞれクエリサンプル及び基準を採取する。両方の符号化された特徴は比較され、サンプル-基準アテンション層にマージされ、予測を生成する。クエリピーク特徴、基準ピーク特徴、及び基準ラベルを提供され、アテンション層は、基準の中のラベル付きの開始／停止点に非常に類似しているクエリ

の中の点がより高い予測確率を取得するように、クエリ及び基準を逐一比較する。2つの一貫したサンプルを所与として、この操作の理想的なアテンションマップは、同じ周辺を有する点がきわめて類似していると思われる（図6Cを参照）「単位行列」となる。アテンションマップと保持時間マッピング行列との間の発散項は、優れた正則化を達成するために最適化プロセスに含まれた。

【0083】

（クエリベースの予測と対照的に）基準ベースの予測の状況では、タスクは、雑音－信号遷移点を、モデルAと同様に識別しようとしなくなるが、類似性のマッピングを最適化するために、むしろシーケンスの中の点及びそのコンテキストの優れた符号化機構を識別しようとすることに留意されたい。アプリケーションでは、基準ベースのモデルは、フューショットラーニングモデルと同様に作動する。基準ベースのモデルは、クエリ／基準の対で訓練され、クエリサンプルでの予測に基準ラベルを正確にマッピングすることを目指す。これにより、当然、単一の（またはいくつかの）正しくラベル付けした基準が提供される限り、モデルは目に見えない遷移ピークに対してよりうまく機能できるようになる。

【0084】

モデルBの採用された構造は、いくつかの高品質の基準ピークの手動アノテーションが妥当な前提条件である、遷移の同じセットでの高スループット実験に対して最も適用可能かつ費用効率が高い場合がある。大部分の発見または小規模の実験では、スタンドアロン基準なし予測因子が好ましい場合がある。

【0085】

図7A及び図7Bを参照すると、各図は、2つのプロットを有している。第1のプロットは、それぞれ星及び三角形で示されたラベル付きの開始時間及び停止時間を有する遷移のプロットである。下部のプロットは、訓練され、検証されたモデルによって予測された開始時間／停止時間の確率のプロットである。

【0086】

3.1 境界及び存在量の予測性能

除外サンプルまたはテストセットサンプルが、モデルを評価するために使用された。解決されている問題の包括的な図を提示するためにいくつかのメトリクスが使用された。予測は、2つの基準、つまり（1）境界（ピーク開始／停止）を予測する上での精度、及び（2）存在量（ピーク面積）を予測する上での精度について評価される。前者は問題の設定に直接的に関連付けられるが、後者は、分析のために分析物の存在量を利用する下流のアプリケーションにとってよりはるかに重要になる。

【0087】

予測では、2つのタスク、つまりピーク開始時間及びピーク終了時間を見つけるタスクの最大予測確率を有する曲線中の点は、以下の文中では「最大点集合境界」と呼ぶ境界のマーカとして使用される。我々は、その境界予測がグラウンドトゥルースアノテーション（人間がラベルを付けた点）の周辺の1.2秒（1シーケンス内で2つの点）の誤差閾値の範囲内であるサンプルの比率として定義された精度スコアとともに、最大点集合境界と人間のアノテーションまたはラベルとの間の平均絶対誤差（MAE）を計算した。ピーク持続時間がテストセットで20+/-8秒（シーケンス中の33+/-14点）であることを所与とすると、誤差閾値は、ウィンドウ内の任意の境界予測が存在量を大幅に変更しないほど十分に小さい。

【0088】

存在量の評価の場合、最初に計算されるのは、それぞれグラウンドトゥルース（人間がラベル付けした）からの境界と、モデル予測との間のベースライン調整済みの積分であり、次に各個別遷移での2つの間のスピアマン順位相関係数（スピアマンの r ）及びピアソン相関係数（ピアソンの r ）が評価される。広範囲の存在量（ $1 \sim 10^{15}$ ）のため、ピアソンの r は、存在量が多いサンプルによって支配され、ここでは、バイアスを回避するために、代わりに対数存在量に対して計算された相関が報告される。全遷移にわたる平均スコアが報告される。さらに、実用的な使用法の表示のために、その予測存在量がアノ

テートされた存在量の $\pm 5\%$ の範囲内にあるサンプルの比率として定義された精度スコアも報告される。この精度メトリクスに1つの注意事項がある。つまり、開始／停止境界予測誤差は、互いを相殺する場合があるため、アノテートされた存在量を逸脱する境界予測によって定義された間隔にわたって計算された積分は、アノテートされた存在量と非常に似た存在量を有するよう見える場合があることに留意されたい。すなわち、強度値が積分で合計されるとき、予測されたピーク開始時間の1%の誤差は、予測されたピーク停止時間の対応する%の誤差を相殺する場合がある。

【0089】

【表1】

表1: 除外テストセットでの予測性能 (67672のピーク)					
	モデル境界予測		存在量予測		
	MAE (秒)	精度*	対数－存在量 ピアソンの r	スピアマンの r	精度**
Skyline	該当なし	該当なし	0.638	0.7372	0.436
規則ベース	4.79	0.285	0.987	0.9708	0.584
モデルA	2.33	0.432	0.989	0.9835	0.688
モデルB	1.56	0.584	0.997	0.9917	0.799

【0090】

【表2】

表2: 訓練セットでの予測性能 (86311のサンプル)					
モデル	境界予測		存在量予測		
	MAE (秒)	精度*	対数－存在量ピア ソンの r	スピアマン の R h o	精度**
規則ベース	3.97	0.31	0.9097	0.992	0.645
モデルA	1.79	0.60 1	0.9477	0.996 5	0.743
モデルB	1.54	0.66	0.9592	0.996	0.779

*境界予測が1.2秒の誤差の範囲内であるサンプルの比率

**存在量予測が5%の誤差の範囲内であるサンプルの比率

【0091】

【表 3】

表 3:バリデーションセットでの予測性能（20044のサンプル）					
モデル	境界予測		存在量予測		
	MAE（秒）	精度	対数－存在量	スピアマン	精度
			ピアソンの r	R h o	
規則ベース	4.08	0.32	0.9141	0.902	0.6593
モデル A	2.57	0.44	0.9332	0.927	0.6884
モデル B	2.80	0.37	0.9541	0.939	0.6563

【0092】

表 1 は、3つの試験した方法及び S k y l i n e 予測の性能スコアを提示する。S k l i n e が、存在量の予測のみを生成し、境界予測タスクに対して評価及び比較されなかったことに留意されたい。モデル A とモデル B の両方とも、規則ベースの方法及び S k y l i n e 予測と比べてすべてのタスクで著しい性能の向上を明示した。モデル B は、追加の基準データを用いて最高のスコアを達成した。境界予測では、規則ベースの方法は訓練／バリデーションセットの性能を最適化するために調整されているので（表 3 を参照）、強力な過剰適合が観察される。比較すると、より優れたサンプル品質のおかげで、モデル A 及びモデル B は、（基準あり及び基準なしで）それぞれテストセットで、規則ベースの方法よりも 2 倍小さい 2.33 及び 1.56 秒の MAE を達成した。同様に、精度スコア性能では、モデル B は、規則ベースの方法に比べて 30% 有利になった。図 8 A 及び図 8 B は、それぞれ、境界予測誤差の棒グラフ及び正規化されていない累積分布を示し、モデル A 及びモデル B は、誤差がより低く、外れ値がより少ない分布を提示した。

【0093】

存在量の予測は同様の結果を提示した。つまりモデル A 及びモデル B は、ピーク存在量の評価時により優れ、よりロバストな性能を達成し（図 8 C）、精度では 20% を超えて上昇した。この問題では、境界点は通常全体的な積分にほとんど寄与しないため、境界予測の誤差の影響が大きく軽減されることに留意されたい。相関スコアは、モデル A 及びモデル B が、現在の既製の質量分析法データ分析ソフトウェアよりも著しくよく機能し、スピアマンの相関係数は 0.9975 及び 0.9990 でほぼ完璧であることを示した。興味深いことに、規則ベースの方法も S k y l i n e の予測よりもはるかに高い相関スコアを達成した。いかなる理論にも拘束されることを望むものではないが、この結果は、テストサンプルの複雑な組成に起因していた可能性があり、これにより保存時間が近い共溶出ピークまたは化合物が生成された。R T ウィンドウが所望の遷移の周りに密接に設定されたとき（通常、人間のアノデータによって使用される設定でもある）、S k y l i n e は誤って異なる範囲からピークを選び、結果的に性能が著しく悪くなった可能性がある。

【0094】

モデル A とモデル B を比較すると、追加の「基準データ」は、モデル B の訓練段階及びバリデーション段階の間に含まれたため、このモデルはモデル A よりもよく機能し、境界予測及び存在量予測の精度で 15% 及び 10% の向上を提供した。モデル B は、存在量相関プロットでの観察と一致して、外れ値がより少ない（2.2% 対 7.1%、図 8 B を参照）より優れた境界誤差分布を提示した。緑の点は中心線の周りにより集中し、一方、青の点はいくつかの外れた予測を含んでいる（図 8 C）。

【0095】

しかしながら、必要とされる訓練セットがより少ないので、用途によってはモデル A が好ましい場合もある。結果が示すように、モデル A は優れた精度／予測存在量を提供し得

るので、モデルBとともに使用される追加の基準データは、必要とされない場合もあれば、小規模の用途または実験に実用的である場合もある。モデルBは、より大規模な実験に、またはより精度が必要とされるときに好ましい場合がある。

【0096】

これらの実験では、モデルBは、プロットに表示されていない別の優位点、つまり複数のサンプルにわたる同じ遷移に対する予測のより高い一貫性を提供した。特に、境界が肩の点及び谷の点のために変わる場合がある遷移の場合、モデルBは、そのモデル構造設計の主要な目的である、その基準の同じ予測形式に従う傾向がある（図9A）。これは、存在量が、まさに同じ標準に基づいて生成され、したがってより正確な比較及び数値化を可能にするという点で、高スループット実験で非常に望ましい。しかし、保持時間のシフトまたは強力なバッチ効果の場合、モデルでの強力な以前の仮定がその性能を損なわせる場合があることにも留意されたい。

【0097】

3.2 人間のアノテータとの比較

ニューラルネットワークベースのモデルの実用的な使用法をさらに確証するために、我々は、モデル予測を、ピーク存在量を計算する人間のアノテータのグループと比較した。

【0098】

テストセットの中では、135すべての血清サンプルに対して12の遷移から成り立つサブセットが分析され、12人のアノテータが独立してピーク開始／終了をマーキングした。元のテストセットラベルと結合され、人間のアノテータの13のセットがこのサブセットのために調製された。次に、我々は、異なるアノテータ間の変動のための代理として、13のアノテーション全体でサンプルピークごとに相対的な標準偏差を計算し、モデルAの及びモデルBの予測相対誤差と比較した。

【0099】

結果は図10に要約されている。図10の灰色の縦線で明示されるように、大部分のサンプルでアノテータに優る高い一貫性が観察される。人間のアノテータの場合のすべてのサンプルピークでの平均RSDは、2.5%である。同時に、モデルAとモデルBの両方ともサブセットに対して良好に機能し、人間のアノテータの平均と比較して、それぞれ3.5%と2.1%の平均相対誤差を達成した。誤差の分布は、青の縦棒及び緑の縦棒に示すように、類似しているがアノテータ全体でのRSDよりもわずかに悪い。ウィルコクソンの符号付検定は、両方のモデルに対して有意性（ $p = 0.11$ 、 0.50 ）を示さず、差が有意ではないことを示している。

【0100】

個々の各アノテータの予測をグループとして見て、我々は次に、アノテータ間の対ごとの差、つまり各サンプルピークでの相対的な存在量の差の平均を計算した。結果は、 $0.025 + / - 0.009$ でアノテータ間の不一致を示し、RSD値と一致している。同時に、モデルBの予測と13人のアノテータとの間の差は、 $0.026 + / - 0.007$ であり、アノテータ間の差と重複している。モデルAの対応する値は $0.039 + / - 0.005$ である。モデル対人間の差が著しく大きくない、または基準ベースの場合にはさらに近いことを所与とすると、ニューラルネットワークベースのモデルが、このタスクでの手動のアノテーションの優れた代用品の機能を果たすことができることが明らかである。また、人間のアノテータが、開始及び停止が読み取りの基準としての機能を果たすために選択されたピークを有し、そのタスクをモデルBに直接的に類似させていたことも留意する。

【0101】

3.3 存在量計算のための加重ピーク開始／終了

肩及び谷の点のあるピークでは、境界の決定は、はるかに高い可変性に苦しむ。人間のアノテーションでさえ、異なるアノテータの中で変動が存在するため、モデルA及びモデルBからの出力は、訓練サンプル全体での平均化のために、多峰型の形状を示すことが予想される（図9B、図7B）。強力な基準の事前確率が存在しないので、これは、基準が

ない場合（モデル A）でもやや深刻であり、最大点集合境界を適用するとき結果は不定になる。これらの場合、境界の単一点の推定値の代わりに完全な予測によりうまく対応できる方法が検討された。

【0102】

最も直接的な方法の 1 つは、予測結果を確率分布として解釈することである。言い換えると、モデルに、ピーク開始／終了マーカとしての機能を果たすであろうシーケンス中の点のすべての考えられる対の確率を評価させることである。N 個の点のシーケンスの場合、これは、順序の制約を課した後に、合計で $N \times (N - 1) / 2$ 対を含む。各対の確率は、2 つの点でピーク開始／終了タスクの予測値を乗算することによって容易に計算できる。確率は、次に、対応する対の間で積分に対して加重和を計算するために使用され、MAP 推定値の代わりに、平均推定値を提供する（存在量を計算するためのこの代替方法は、以下「加重和法」と呼ばれる）。MAP 推定値は、最大点集合値をエンドポイントとして使用して 1 つの存在量値を計算する。

【0103】

存在量の平均は、特にモデル B よりもモデル A により類似したモデルアーキテクチャが使用されるとき、最大点集合境界を使用するよりもよりロバストな存在量推定値を提供し得る。明示的な境界は使用されないの、方法は、ピーク開始予測の前に（または非常に近くに）現れるピーク終了予測の問題を回避し得る。モデル A 及びモデル B の存在量相関の比較は、それぞれ図 11 A 及び図 11 B に提示される。

【0104】

特にモデル A の場合、加重和法は、より正確な存在量を生成し、存在量相関を 0.9897 から 0.9969 に上昇させるのに役立った。モデル B での増加は、相対的に小さい（0.9974 から 0.9977）。後処理ステップは、その最大点集合境界が人間のアノテーションよりも著しく低い存在量を生じさせた外れ値を回避する上でおもに役に立った。これらのサンプルは、通常、次に最大点集合操作によって選ばれる（谷または肩の点での）ピークの範囲内の異常に高い境界確率に苦しむ。同時に、加重和法により、全体的な予測は、低存在量ピークの場合、上方に偏向した。これはおもに雑音の多い信号及び予測に起因する。信号対雑音比が低い場合、予測境界確率は、ピークの周辺でのより長い保持時間範囲で変動し（図 7 A 及び図 7 B、ならびに図 12）より高い加重和につながるであろう。雑音の多いピークは、下流のアプリケーションにとってはあまり価値がないので、我々は、この影響を許容できると見なした。

【0105】

これらの規則から、信頼値は以下の通りに計算された。N 個の点のあるサンプル曲線での予測の場合、範囲内のすべての極大値を計算する。最大値は、以下としてラベル付けされ、

【0106】

【数 1】

$$(\hat{x}^*, \hat{y}^*)$$

、ここでは

【0107】

【数 2】

$$\hat{x}^* \in \{1, 2, \dots, N\}, \hat{y}^* \in [0, 1]$$

である。

【0108】

（極大値が存在する場合）その左側にある極大値は、

【0109】

【数 3】

$$(\hat{x}_i, \hat{y}_i), i = 1, 2, \dots, p$$

としてラベル付けされ、ここでは

【 0 1 1 0 】

【数 4】

$$\hat{x}_1 < \hat{x}_2 < \dots < \hat{x}_p$$

10

であり、（最大値が存在する場合）右側にある最大値は、

【 0 1 1 1 】

【数 5】

$$(\tilde{x}_j, \tilde{y}_j), j = 1, 2, \dots, q$$

としてラベル付けされ、ここでは

【 0 1 1 2 】

【数 6】

$$\tilde{x}_1 > \tilde{x}_2 > \dots > \tilde{x}_q$$

20

である。さらに、

【 0 1 1 3 】

【数 7】

$$\hat{x}_0 = \hat{x}^*$$

及び

【 0 1 1 4 】

30

【数 8】

$$\tilde{x}_0 = \tilde{x}^*$$

を定義することによって、確実性は以下の通り計算される。

【 0 1 1 5 】

【数 9】

$$\begin{aligned} \text{信頼値} = \hat{y}^* - \frac{1}{N-1} & \left(\sum_{i=1}^p (\hat{x}_i - \hat{x}_{i-1}) \times \max(\hat{y}_i, \hat{y}_{i+1}, \dots, \hat{y}_p) + \right. \\ & \left. \sum_{j=1}^q (\tilde{x}_{j-1} - \tilde{x}_j) \times \max(\tilde{y}_j, \tilde{y}_{j+1}, \dots, \tilde{y}_q) \right) \end{aligned}$$

40

図解は図 1 2 に示される。

【 0 1 1 6 】

3. 4 予測信頼性

点ごとの予測構造のもう 1 つの中心的な優位点は、予測確実性を推定する際の容易さである。ニューラルネットワーク予測の不確実性を数値化する内部の方法は依然として複雑

50

であり、制限を伴うが、このタスクでは、境界についてモデル予測を検査することによって確実／不確実を推論することはかなり簡単である。大部分の予測された確率はガウス混合の形をとり、その数量、及び高さはモデルの確実性を表す。研究では、以下の指針に基づいて予測信頼性を数値化するために発見的方法が採用された。

項目より低い予測値→より低い確実性

項目複数の極大値→より低い確実性

項目より分散した極大値→より低い確実性

【0117】

これらの規則から、信頼値が計算された。図12は、図解とともに信頼水準を計算するために使用された計算の説明を提供する。信頼スコアは、陰影領域に比例する。予測値を高く（1に近く）、単峰型にするか、複数の極大値が表示される場合は閉じることを必要とする。

【0118】

信頼スコアは、モデルA及びモデルBのすべての予測について計算された（図13A及び図14A）。異なる信頼カットオフ全体で、性能の変化の傾向を見ることができる。図13B及び図14Bに示すように、境界での予測誤差は、確実性が高まるにつれて着実に減少し、最も自信があるサンプルでは1秒近くまで低下する。

【0119】

異なる信頼区間での存在量の相関は、図13A～図13Fに示されている。高存在量サンプルでの予測は、通常、より高い確実性を有し、人間のアノテーションとより良く一致する。上位2つのビンの予測は、0.999を超えるピアソンのrを達成した。対照的に、より低い予測信頼性のサンプルはより悪い性能を生じさせ、境界の平均絶対誤差（MAE）はより高く、存在量相関はより低い。これは、部分的には、低い存在量によって示される入力サンプルのより悪い信号対雑音比に起因する。全体的には、確実性の測定はノイズスケール及びモデル性能の優れた指標としての機能を果たした。

【0120】

この後処理ステップは、ワークフロー品質管理を実施する上で多用途性を提供し得る。現実世界の用途では、直接的な使用法は、値を下回る予測が除外される、または再評価のために人間の専門家に提示されるように閾値信頼スコアを設定することであろう。また、これは実験データの異常を検出するためにも役立つであろう。

【0121】

5.0 実験

実験パラメータ

化学薬品及び試薬：ヒト血清、ジチオスレイトール（DTT）、及びヨードアセトアミド（IAA）は、Sigma-Aldrich（St. Louis, MO）から購入した。シーケンシンググレードのトリプシンは、Promega（Madison, WI）から購入した。アセトニトリル（LC-MSグレード）はHoneywell（Muskegon, MI）から購入した。

サンプル調製：血清サンプルは、DTTで還元し、IAAでアルキル化した後、水槽内でトリプシンを用いて37℃で18時間温浸した。温浸を冷ますために、インキュベーション後、各サンプルにギ酸を添加し、最終濃度を1%（v/v）にした。

【0122】

5.1 LC-MS分析

温浸した血清サンプルは、Agilent 1290無限UHPLCシステム及びAgilent ZORBAX Eclipse Plus C18カラム（2.1mm×150mm内径1.8µm粒子径）を備えたAgilent 6490三連四重極質量分析計に注入した。ペプチド及びグリコペプチドの分離は、70分のバイナリグラジエントを使用し実行された。水性移動相Aは3%アセトニトリル、0.1%水中のギ酸（v/v）、及び有機移動相Bは90%アセトニトリル、0.1%水中のギ酸（v/v）であった。流量は0.5mL/分に設定された。イオン化源としてエレクトロスプレーイオン化（ESI

10

20

30

40

50

）を使用し、陽イオンモードで操作した。三連四重極は、動的多重反応モニタリング（dMRM）モードで操作した。訓練及び試験で用いたペプチド及びグリコペプチドの遷移は、その内容全体が、すべての目的のためにその全体として参照により本明細書に組み込まれる、Li et al., Site-Specific Glycosylation Quantification of 50 serum Glycoproteins Enhanced by Predictive Glycopeptidomics for Improved Disease Biomarker Discovery, Anal. Chem. 2019, 91, 5433–5445; DOI: 10.1021/acs.analchem.9b00776に公開されたものの選択であった。

【0123】

5.2 データセットの準備

訓練セット及びバリデーションセットのピーク、ならびにテストセットのピークは、2セットのマススペクトル実験から収集された。セット1は、2017年後半にUC DavisのCarlito Lebrillaの実験室で測定されたヒト血清サンプル及び遷移から成り立っていた。一方、セット2は、2018年半ばにIndivumed GmbHから購入した卵巣がん及び良性腫瘍患者からの市販の血清であった。サンプルの両方のセットは、上記に概略した実験プロトコルを受けた。ポストランAgilent.Dフォーマットファイルが、バイオインフォマティクス処理用のProteowizard 3.0ソフトウェアスイートの中のmsConvertを介してmzMLに変換された。

【0124】

人間のアノテーションを取得するために、すべての遷移は、1つの基準血清サンプル（プールされたシグマ血清）のためにマススペクトル専門家によってラベルを付けられ、これらのラベルが「基準」ピークを構成している。残りのサンプルについて、アノテータは、基準ピークに基づいて、及び保持時間の任意のシフトを考慮して、AgilentのMassHunter Quantitative Analysis B5.0でピーク開始及びピーク終了の保持時間マークをラベル付けした。

【0125】

2つの実験から、正の存在量でラベル付けされたピークだけが選択された。また、保持時間の大きいシフト（> 0.2分）のあるサンプルも除外された。合計106355（訓練／バリデーション）及び67672（試験）ピークがサニティテストに合格し、この研究で使用された。遷移ごとに、我々は、基準保持時間開始−0.2分から基準時間停止+0.2分までの保持時間ウィンドウの中のすべての信号を収集し、所望の前駆物質及び製品m/zの周囲で±0.1の質量−電荷比ウィンドウを可能にした。信号は、質量−電荷比ウィンドウに沿って合計され、抽出イオンクロマトグラム（XIC）を生成した。シーケンシャルモデルと人間のアノテータの両方とも、ターゲット遷移の周囲でトリミングされたXIC曲線を提示された（図7A、図7B、及び図9Bは、いくつかの例示的なXIC入力を示している）。

【0126】

「人間のアノテータとの比較」と題する上記の項で使用したテストセットは、1619のピークを含んでいた。すべてのピークは、整合性のために12人のアノテータによって独立してさらにアノテートされたが、正の存在量のアノテーションだけが、遷移ごとの相対的な標準偏差及びアノテータ間の対ごとの不一致を計算するために使用された。

【0127】

5.3 フィーチャ化

入力は、一定の開始／停止または長さがなく、強度−保持時間平面上の点のシーケンスとして策定された。一様な表現を生成するために、フィーチャ化が各サンプル曲線で適用された。単一のサンプルの中のすべての点の保持時間が最初にサンプル曲線の頂点位置で中心に置かれ、次に−1分から1分に及ぶ、128の等しく離間されたガウスビンによって拡大され、時点を表す256の実数のベクトルFRTを生成した。

【0128】

質量対電荷の強度値は広範囲を有し、強度値は、256のガウス関数を0から500までの範囲に適用することによって同様に離散化され、アナログ値を、強度を表す256の実数のベクトル I_{RT} に変換した。上位閾値よりも大きい強度を有するあらゆる点は、ピークの一部と見なさなければならない旨の仮定を採用して、雑音段階または雑音－信号遷移段階に位置するそれらの点だけがフィーチャ化された。 F_{RT} 及び I_{RT} を形成した後の各サンプル曲線の完全な入力 X （つまり、 RNN モデルに対する入力パラメータまたは状態）は、サイズ $N \times 384$ の行列に形成された。変数 N は50～300に及び、曲線の長さを表していた。

【0129】

モデルBの場合、各サンプルは基準と対にされた。入力の対は、上記と同じフィーチャ化プロセスを有し、寸法 $N \times 384$ 及び $N' \times 384$ の2つの行列 X 及び X_{ref} として形成された。 N 及び N' は近くなければならないが、それらは必ずしも互いに等しくはない。

【0130】

サンプルのラベル付けは、ピーク開始／終了の人間のアノテーションが対応するタスクで1の値を有する同じ連続構造として策定された。安定性及びロバスト性を強化するために、さらなるラベル平滑化が適用された。つまり、人間によってアノテートされた

【0131】

【数10】

$$\hat{y} = \exp\left(-\frac{d_{RT}^2}{\sigma_{RT}} - \frac{d_{ABD}^2}{\sigma_{ABD}}\right)$$

ラベル／グラウンドトゥールズ開始／終了アノテーションに近いサンプル点も以下として計算された正のラベル値を有する。

d_{RT} 及び d_{ABD} は、保持時間（ RT ）の差及び存在量（ ABD ）の差であり、 σ_{RT} 及び σ_{ABD} は、訓練中のそれぞれの減衰率を制御するハイパーパラメータである。平滑化の目的は、雑音の多いラベルに対するモデル適応性を可能にする、及び／または単一の点を指定しない人間のラベル付け／アノテーションを説明することである。しかしながら、ラベルに近い点の近傍は、雑音から信号への遷移も示す妥当な点を表す可能性がある。

【0132】

5.4 訓練設定

再び図5Aと図5B（モデルA）及び図6Aと図6B（モデルB）を参照すると、モデルA及びモデルBはPytorchに実装されていた。

【0133】

モデルAは、隠れ状態のための128の特徴を有する2つの $bi-LSTM$ 層を含み、 $LSTM$ 出力は、次に8つのヘッド及び各ヘッドに32の特徴を有するマルチヘッドアテンション層に適用される。ヘッドからの出力は連結され、32のユニットを有する完全に接続された隠れ層を通過し、次に各々がアフィン層を通して2つのクラスを有する2つのタスクにマッピングする。ソフトマックスは、ピーク開始／終了の確率を生成するために各タスク及びシーケンス中の各点に適用される。モデルは、平滑化したラベルと、バッチサイズが8、及び学習率が0.001である20のエポックに対して86311のサンプルを含む訓練サブセットに対するソフトマックス予測との間の交差エントロピー損失によって訓練される。早期停止は、バリデーションセットの性能に基づいて適用される。

【0134】

モデルBは、それぞれサンプル及び基準を符号化するために使用される、完全に接続された層を除き、上記と同一の2セットのシーケンシャルネットワーク構造を含んでいた。重みは2つの構造の間で共有されない。サンプル－基準アテンション層は、（クエリとして）サンプル符号化、（キーとして）基準符号化、及び（値として）基準ラベルで適用され、その出力はサンプル符号化と同じ長さであり、タスク寸法で2である。これは、ピーク開始／終了確率のためのモデル予測として直接的に使用される。別の損失項は、サンプ

ルー基準アテンションマップと、（単位行列と同様に）点をサンプル及び基準の同じ保持時間と接続する標準マッピングとの間のKL発散として計算される。モデルは、同じ設定で同じ訓練セットの追加の発散項を加えた交差エントロピー損失によって訓練される。

【0135】

追加の定義

本明細書で使用する場合、「プロセッサ」は、SASプログラムコードを実行すること、GUIのためにデータを処理すること、またはメモリから読み書きすることなどのタスクを実行するために使用される1つまたは複数のプロセッサを含むと解釈されるべきであり、汎用コンピュータを含む。

【0136】

本明細書で使用する場合、「ユーザー」は、アプリケーションをローカルで使用する、または遠隔に位置するコンピュータを使用するウェブブラウザを遠隔で使用してアプリケーションにアクセスする実際の人物を含むと解釈されるべきである。

【0137】

本明細書で使用する場合、「コンピュータ」は、ハードディスクから読み取り、ハードディスクに書き込むためのハードディスクドライブ、取り外し可能な磁気ディスクから読み取るまたは取り外し可能な磁気ディスクに書き込むための磁気ディスクドライブ、及びCDROM、DVD、または他の光学媒体などの取り外し可能な光ディスクから読み取るまたは取り外し可能な光ディスクに書き込むための光ディスクドライブを有するコンピュータを含むと解釈されるべきである。ハードディスクドライブ、磁気ディスクドライブ、及び光ディスクドライブは、それぞれハードディスクドライブインタフェース、磁気ディスクドライブインタフェース、及び光ドライブインタフェースによってシステムバスに接続される。ドライブ及びそれらの関連するコンピュータ可読媒体は、コンピュータ可読命令、データ構造、プログラム、及びコンピュータシステムのための他のデータの非揮発性（非一時的な）ストレージを提供する。

【0138】

本明細書で使用する場合、「データベース」は、集中データベースまたは分散データベースを含むと解釈されるべきであり、SQLデータベースを含む。

【0139】

本明細書で使用する場合、用語「データ」は、ファイルシステムの機械可読メモリに含まれる、もしくはデータベースで使用される任意の情報、または他の構造化データ、または一時的なメモリもしくは非一時的なメモリ内の非構造化データ表現を含むと解釈されるべきである。

【0140】

本明細書で使用する場合、「入力装置」は、キーボード、マウス、ポインタ、タッチパッド、音声認識装置、またはメモリに格納され得る電子コンテンツに修正を加える、探索する、アクセスする、または生成するための（ディスプレイ付きのまたはディスプレイなしの）コンピュータに接続された他の形の入力装置を含むと解釈されるべきである。

【0141】

本明細書で使用する場合、「ディスプレイ」は、コンピュータに接続され、GUIを使用して入力をレンダリング及び受信することが可能である、または機械実行可能コードを実行するコンピュータで実行中のアプリケーションによって生成される他の情報を表示するLCD/LEDタイプのディスプレイまたは他のタイプの表示装置である。

【0142】

本明細書で使用する場合、「モジュール」または「コンポーネント」は、ソフトウェアもしくはソースコードファイル、またはソースコードファイルに含まれているもしくは含まれておらず、非一時的メモリ媒体もしくは一時的メモリ媒体に存在する機械実行可能命令のセットを含むと解釈されるべきである。モジュールは、スタンドアロンプログラムとして存在する、または別のプログラムによって呼び出されるプログラムを表す場合がある。「コンポーネント」は、モジュールが、モジュールの1つのタスクまたは複数のタスク

10

20

30

40

50

の部分を実行するためにコンポーネントを呼び出すという意味では「モジュール」の一部分である。

【0143】

使用され、解釈される「データリポジトリ」または「リポジトリ」は、データベース、データストア、データウェアハウス、データレイク、またはこれらのものの組み合わせを含む場合がある。

【0144】

本明細書で使用する場合、「メモリ」は、機械可読媒体、一時的なコンピュータ可読記憶媒体または非一時的なコンピュータ可読記憶媒体のどちらかを含むと解釈されるべきである。様々な実施形態では、メモリの揮発性（一時的な）部分は、スタティックランダムアクセスメモリ（SRAM）、同期ダイナミックRAM、または任意の他のタイプのメモリなど、任意の適切なメモリ技術を使用して実装され得る。図示され、説明されるメモリの例は単一の機械可読媒体を指す場合があるが、メモリは、機械実行可能コードまたはデータの1つ以上のセットを格納する、単一の媒体または複数の媒体（例えば、集中データベースもしくは分散データベース、及び／または関連するキャッシュとサーバ、ならびにプロセッサレジスタ、メモリ、及び記憶装置などの種々の記憶媒体）を含むと解釈されるべきである。また、用語「機械可読媒体」は、機械による実行のために命令のセットを格納する、符号化する、または運ぶことができ、機械に本明細書の方法の任意の1つ以上を実行させる、またはそのような命令のセットによって利用されるもしくはそのような命令のセットと関連付けられるデータ構造を格納する、符号化する、または運ぶことができる任意の媒体を含むと解釈されるものとする。用語メモリは、相応して、ソリッドステートメモリならびに光媒体及び磁気媒体など、有形媒体を含むと解釈されるものとする。

【0145】

本明細書で使用する場合、「モジュール」は、他のモジュールに情報を提供する、及び他のモジュールから情報を受け取ることができる異なる機能のユニットを含むと解釈されるべきである。したがって、説明するモジュールは、通信で結合されていると見なし得る。また、モジュールは、入力装置または出力装置との通信を開始し得、リソース（例えば、データベースなど、情報の集合体）に作用することができる。モジュールは、様々な実施形態の特定の実装のために、必要に応じて、ハードウェア回路、光学部品、シングルプロセッサ回路もしくはマルチプロセッサ回路、メモリ回路、ソフトウェアプログラムモジュール及びオブジェクト、ならびにそれらの組み合わせを含み得る。用語「モジュール」は、特定の機能、操作、処理、または手順を達成するために、コード、データ、または計算オブジェクトの識別可能な部分を含む場合がある。

【0146】

いくつかの例では、質量分析法の遷移を数値化することであって、遷移が、アミノ酸配列を含むグリコペプチドまたはペプチドに特有な m/z 強度値のピークである、数値化すること、及び数値化に基づいて、サンプルが病気分類の範囲内にある、または範囲内にならない確率を決定するために、モデルを使用して数値化を分析することを含む、病気を診断するための方法が、本明細書に説明される。これらの例のいくつかでは、遷移、グリコペプチド、ペプチド、アミノ酸配列、またはそれらの任意の組み合わせは、2020年1月31日に出願された国際PCT特許出願第PCT/US2020/0162861号に説明されるものであってよい。

【0147】

要約書に説明されることを含む本発明の示されている実施形態の上記説明は、網羅的になる、または本発明を開示される正確な形に限定することを意図していない。本発明の具体的な実施形態及び実施例は、例示の目的で本明細書に説明されているが、関連技術の当業者が認識するように、様々な修正が本発明の範囲内で考えられる。

【0148】

これらの修正は、上記の詳細な説明を鑑みて本発明に加えることができる。特許請求の範囲で使用する用語は、本発明を、明細書に開示する特定の実施形態に限定すると解釈さ

10

20

30

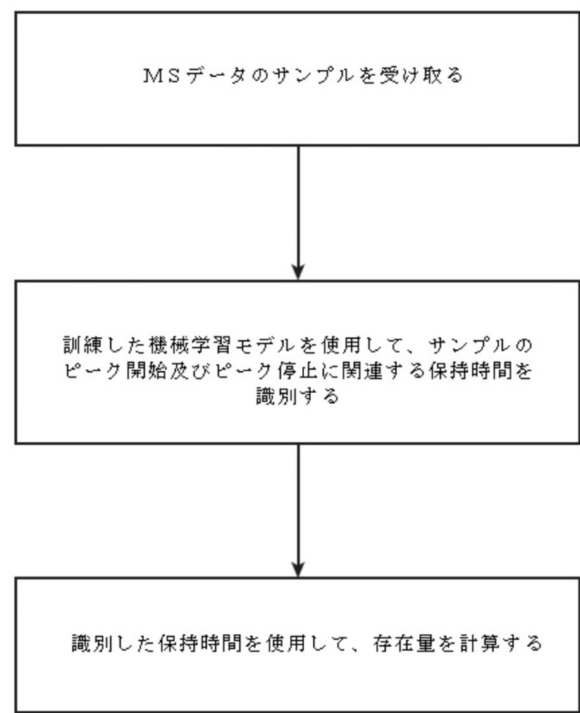
40

50

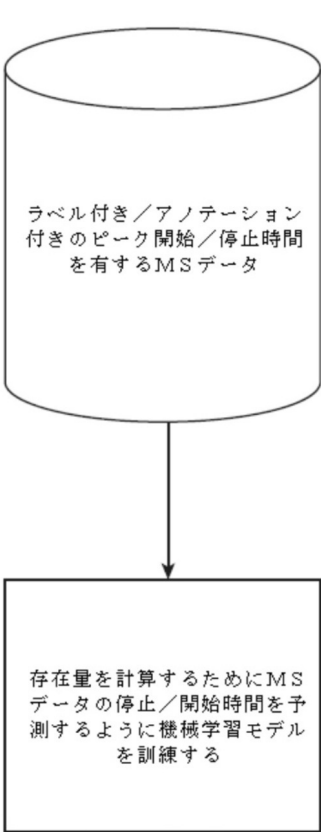
れるべきではない。

【図面】

【図 1】



【図 2】



10

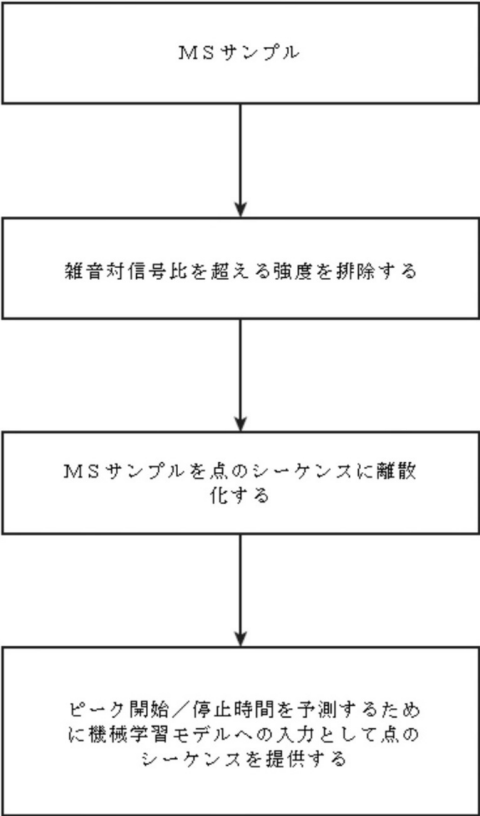
20

30

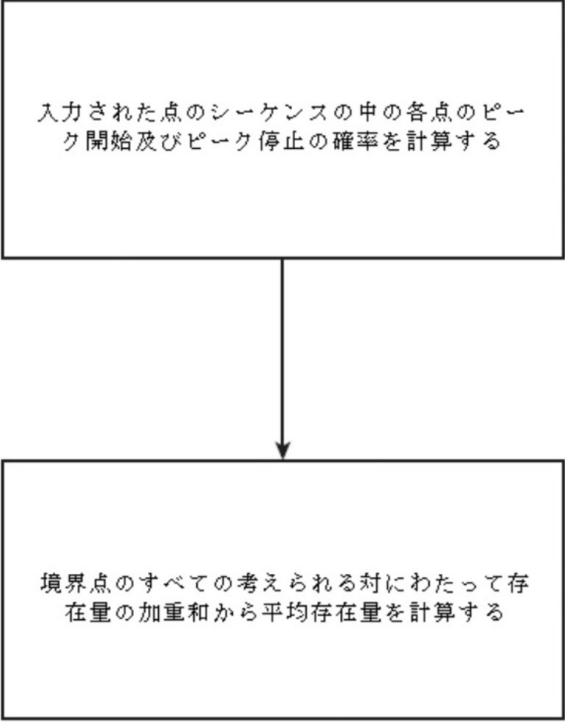
40

50

【図 3】



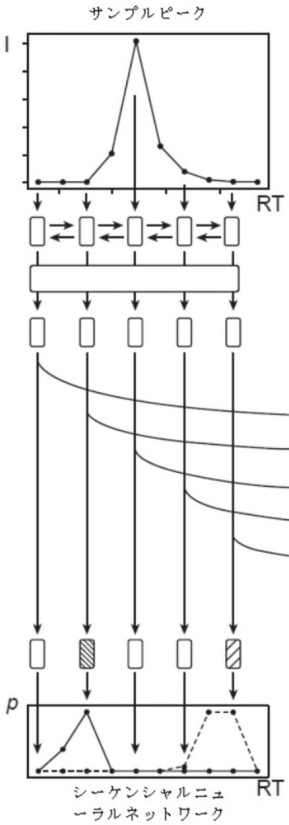
【図 4】



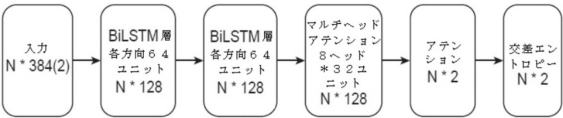
10

20

【図 5 A】



【図 5 B】

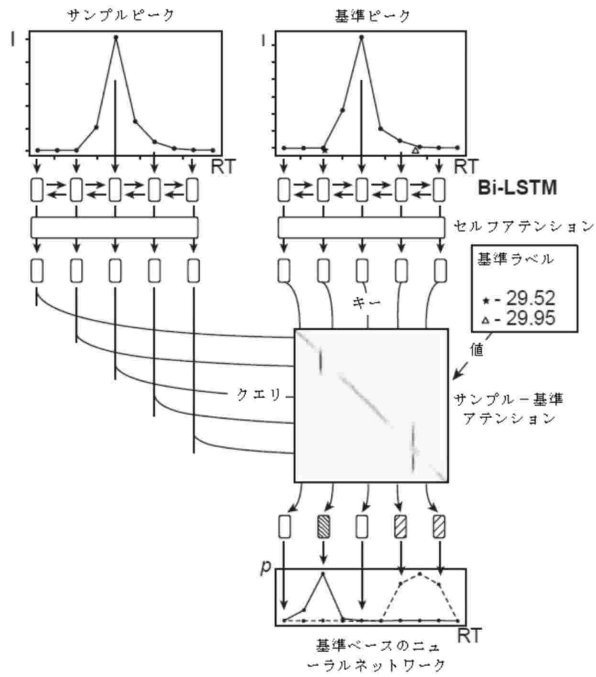


30

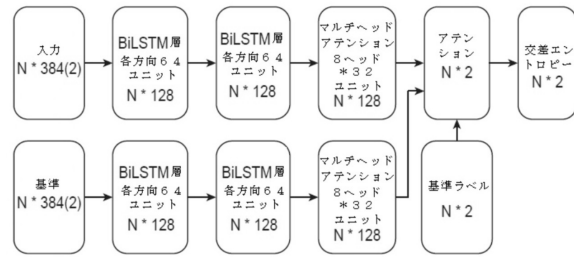
40

50

【図 6 A】

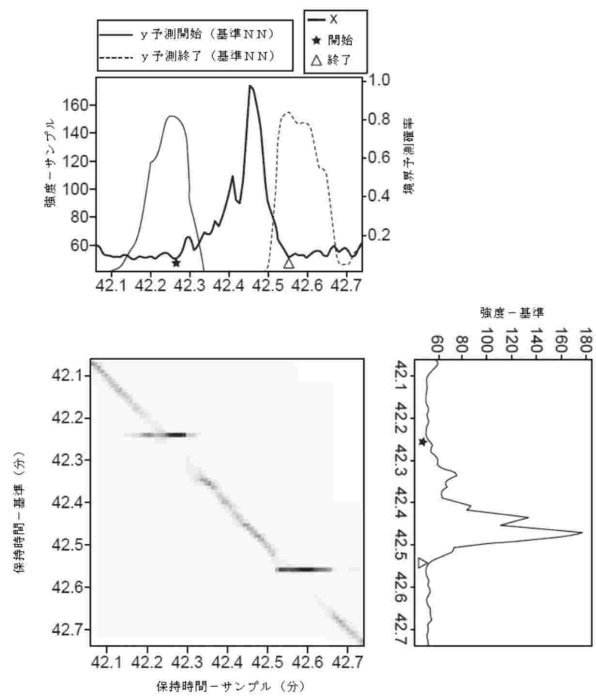


【図 6 B】

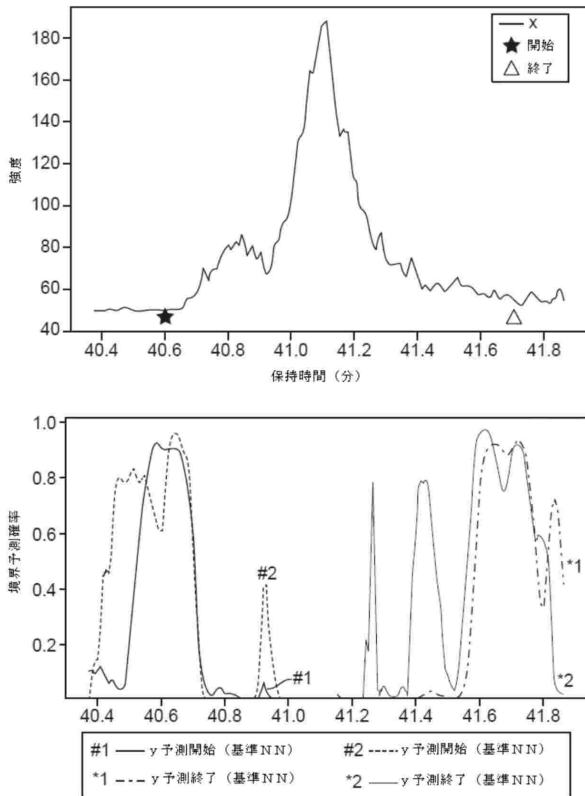


10

【図 6 C】



【図 7 A】

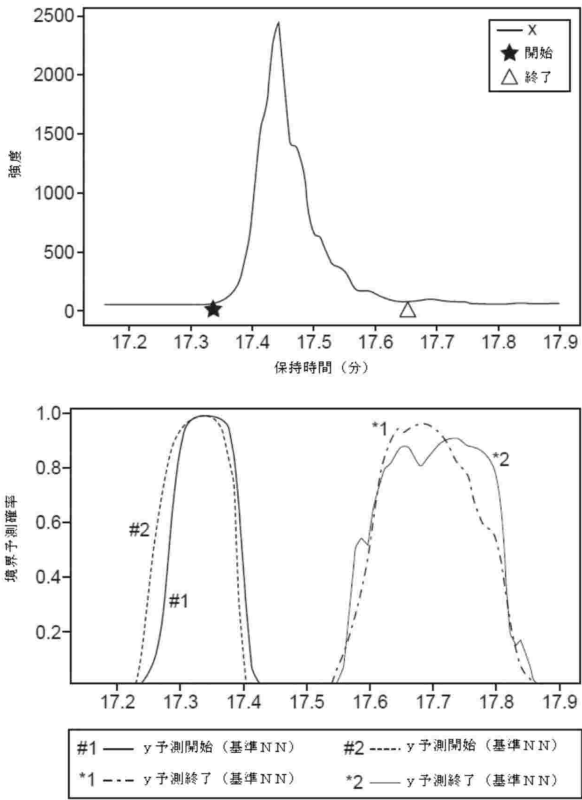


30

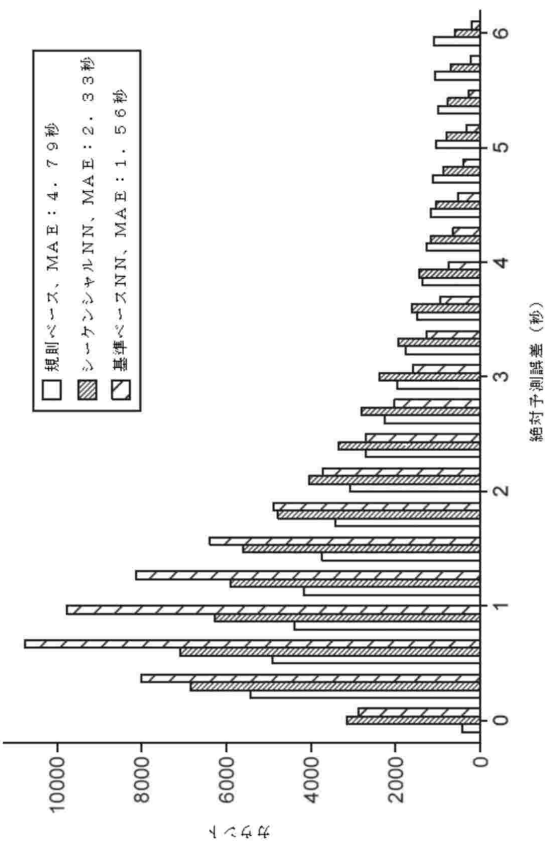
40

50

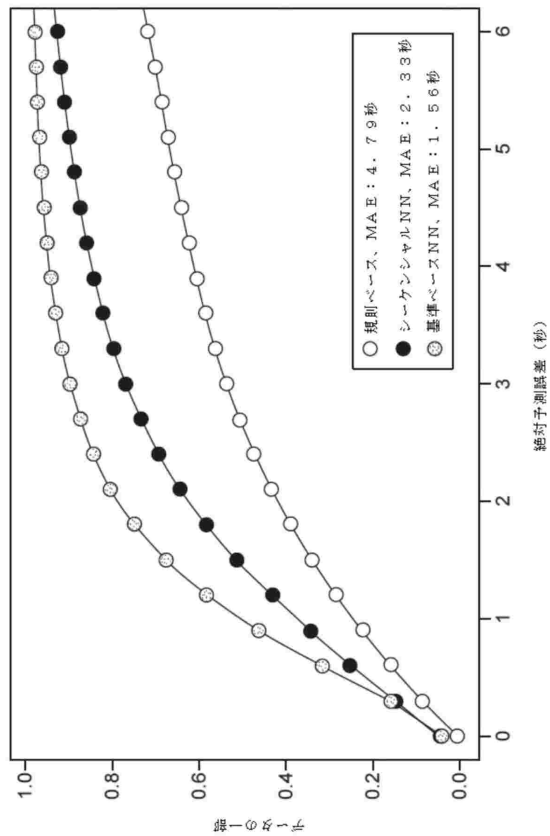
【図 7 B】



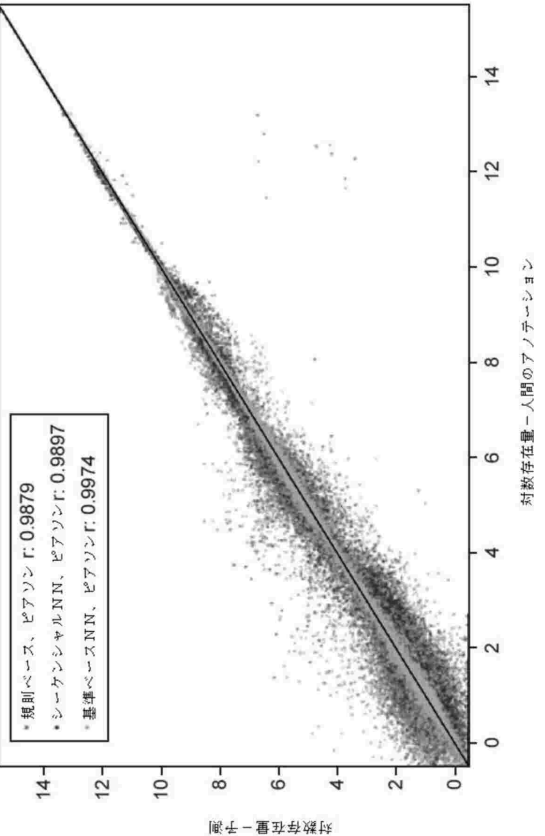
【図 8 A】



【図 8 B】



【図 8 C】



10

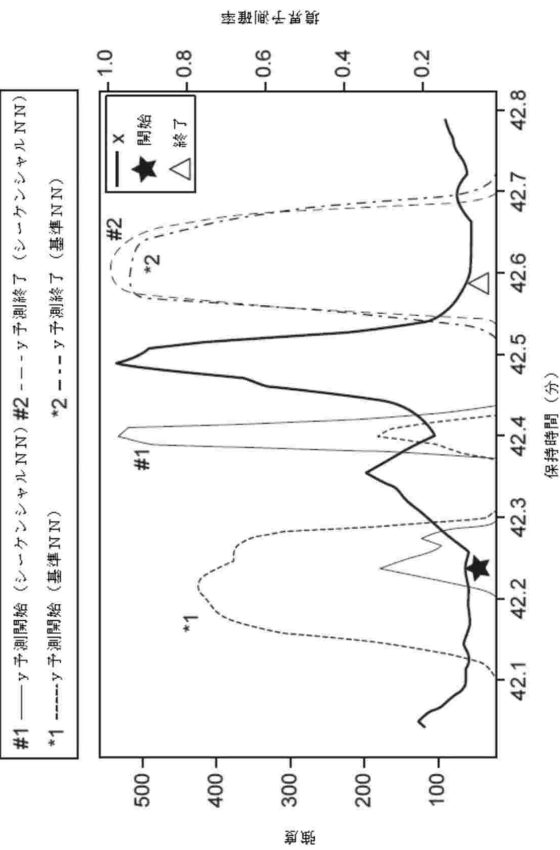
20

30

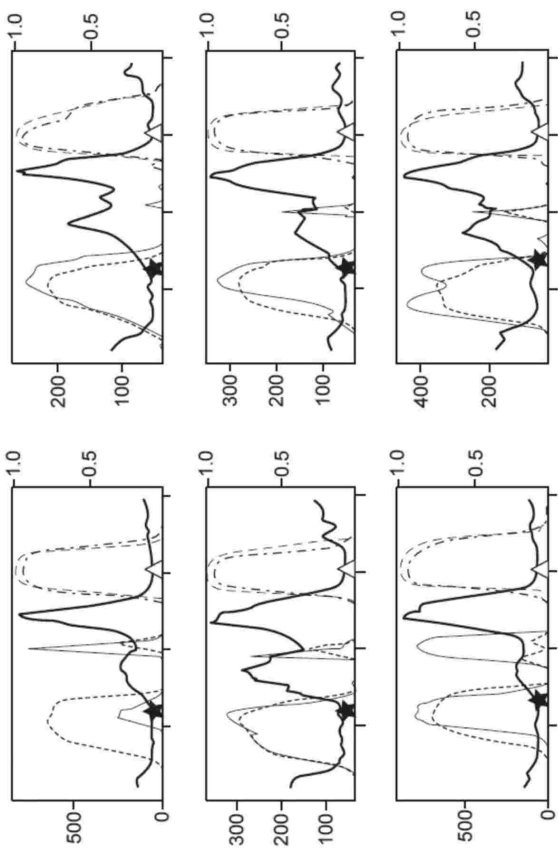
40

50

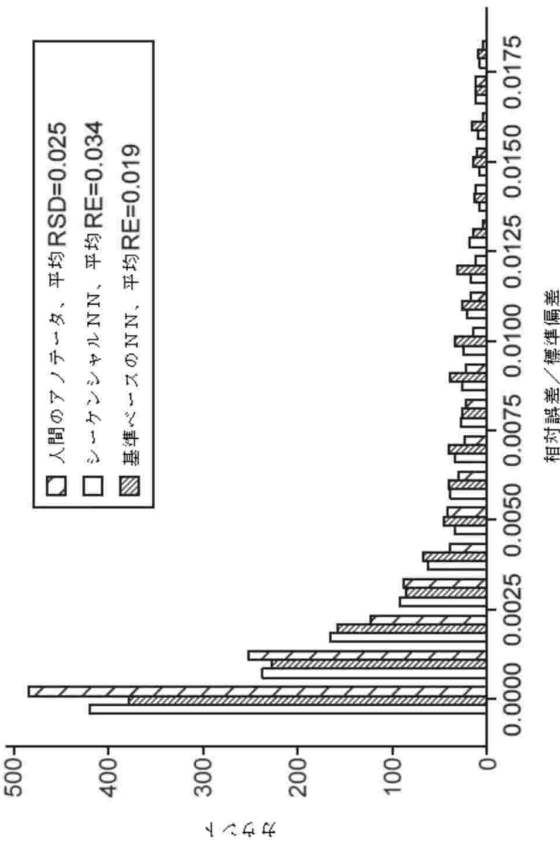
【図 9 A】



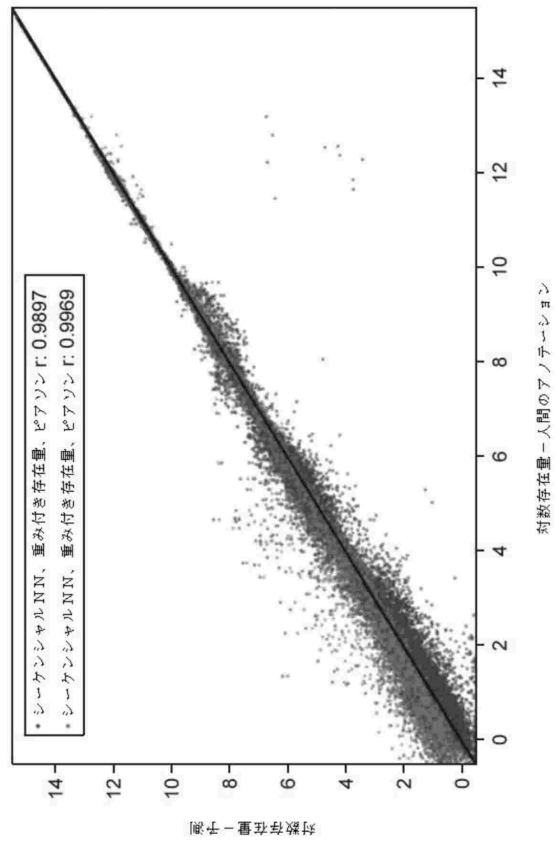
【図 9 B】



【図 10】



【図 11 A】



10

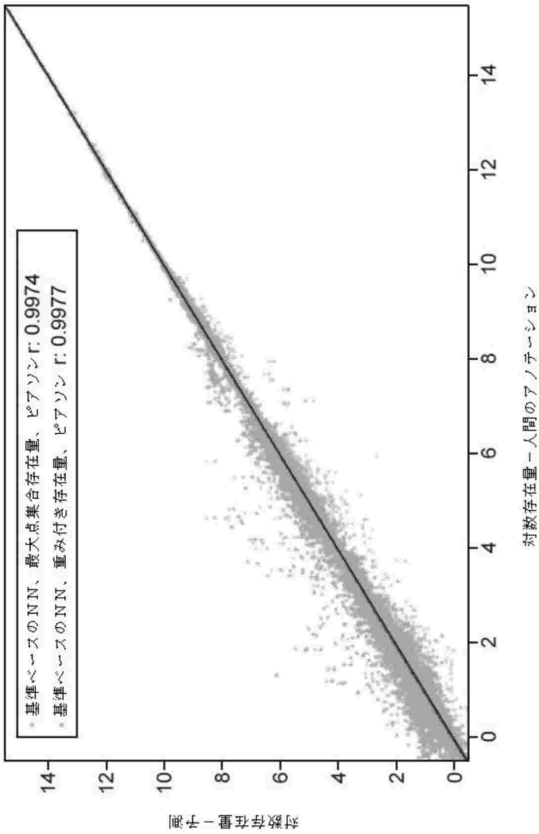
20

30

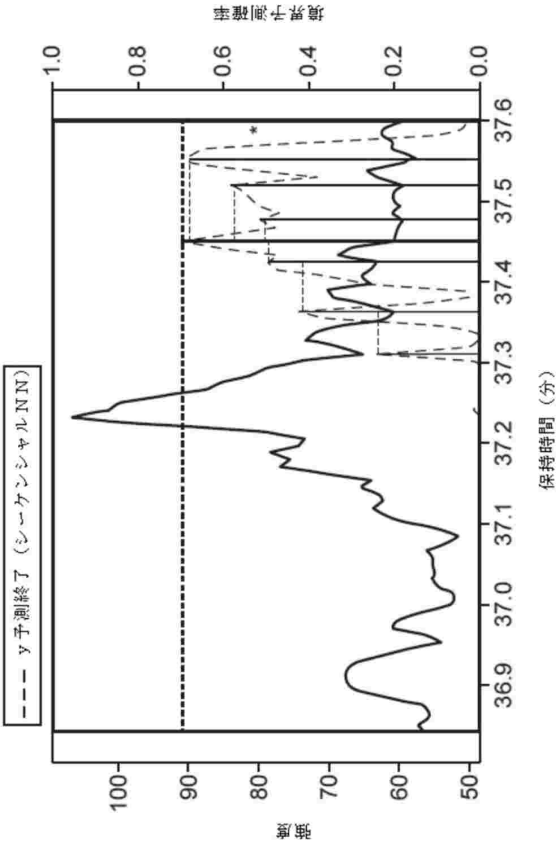
40

50

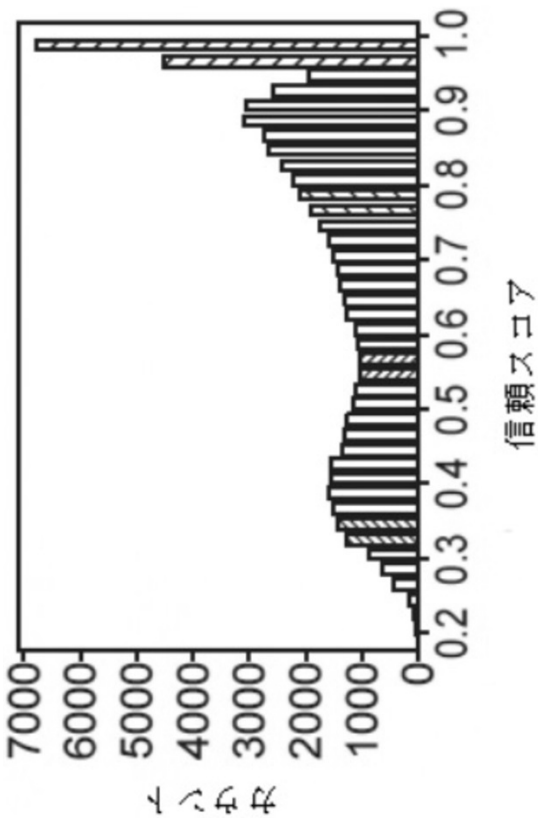
【図 1 1 B】



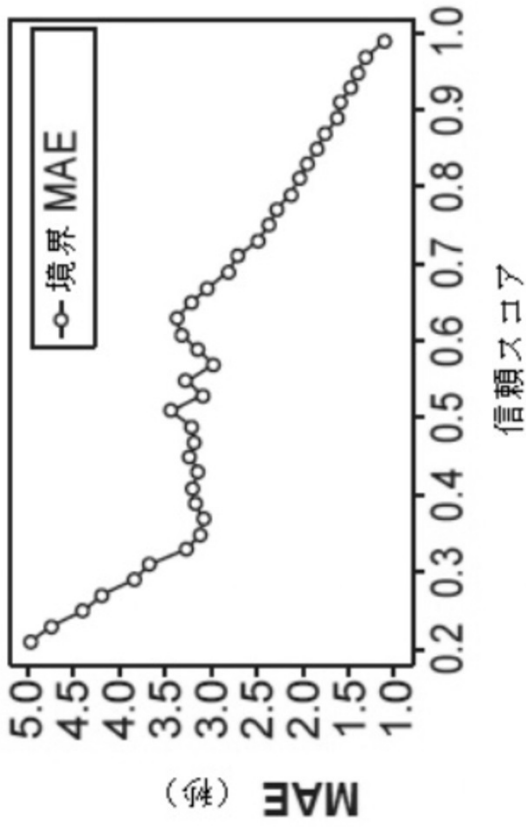
【図 1 2】



【図 1 3 A】



【図 1 3 B】



10

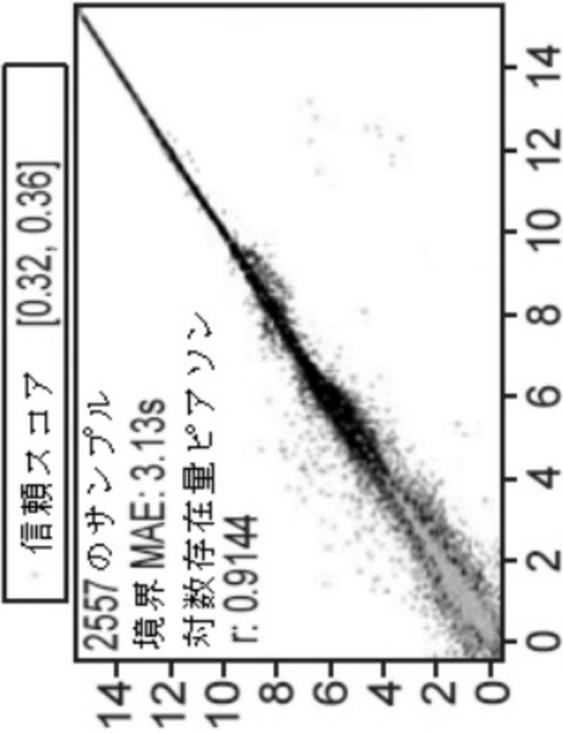
20

30

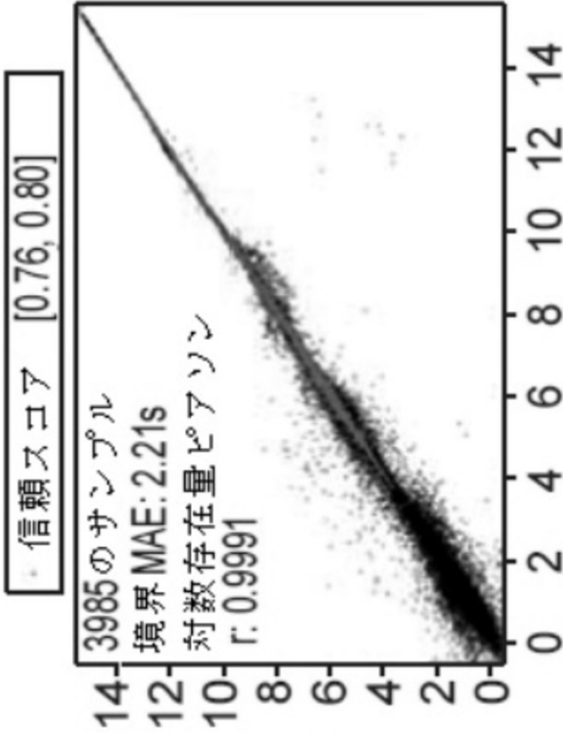
40

50

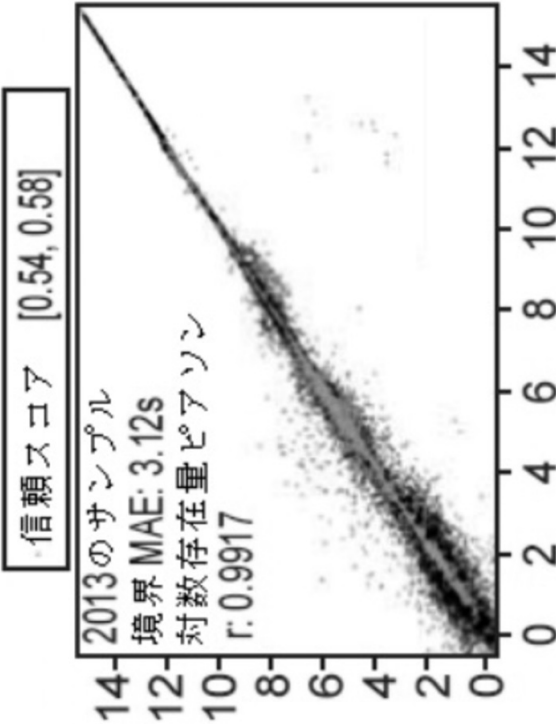
【図 1 3 C】



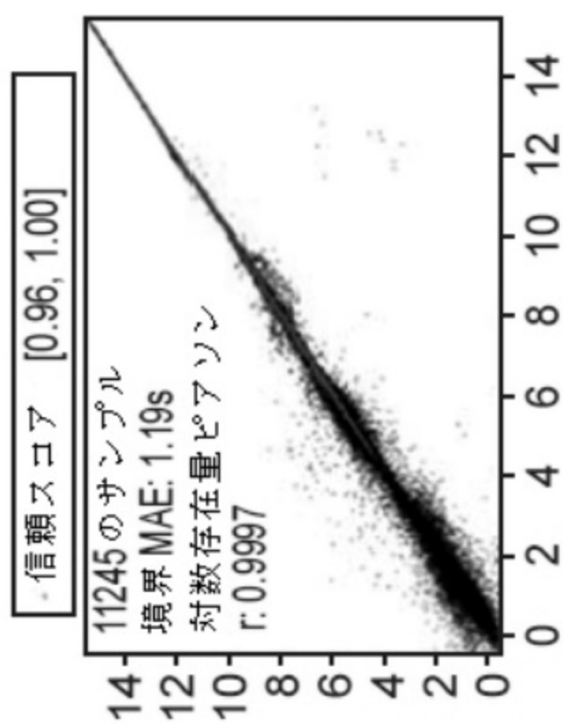
【図 1 3 D】



【図 1 3 E】



【図 1 3 F】



10

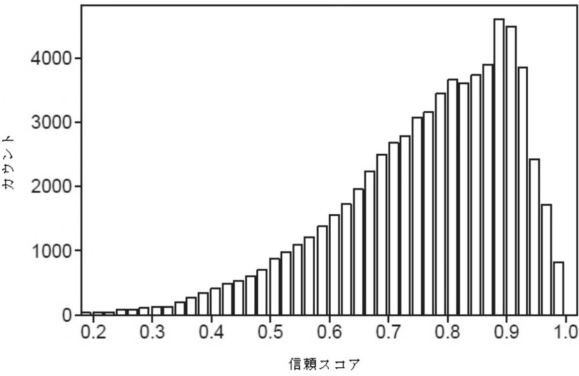
20

30

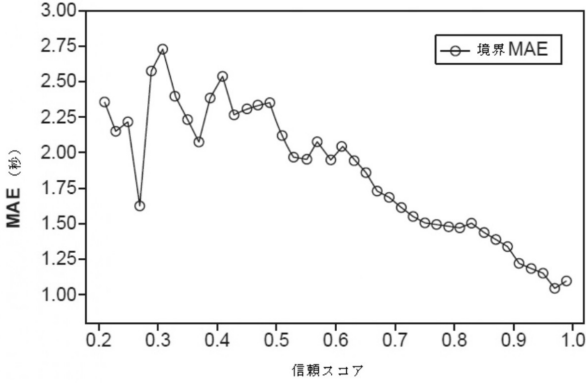
40

50

【図 1 4 A】



【図 1 4 B】



10

20

30

40

50

フロントページの続き

 タワー・プレイス 5階 シー／オー ヴェン バイオサイエンシーズ コーポレーション
(72)発明者 ウー, ジェンチン
 アメリカ合衆国 94080 カリフォルニア州 サウスサンフランシスコ ツー・タワー・プレイス
 5階 シー／オー ヴェン バイオサイエンシーズ コーポレーション
 審査官 伊藤 裕美
(56)参考文献 特表2008-536147 (JP, A)
 特開平06-094696 (JP, A)
 特開2018-189607 (JP, A)
 国際公開第2019/092836 (WO, A1)
 特表2019-502212 (JP, A)
 米国特許出願公開第2010/0299078 (US, A1)
(58)調査した分野 (Int.Cl., DB名)
 G01N 27/60-27/70
 H01J 49/00-49/48
 G01N 33/48-33/98
 G01N 30/00-30/96