

(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
2024年10月24日(24.10.2024)



(10) 国際公開番号

WO 2024/218924 A1

(51) 国際特許分類:
G16B 15/00 (2019.01) G16B 40/00 (2019.01)

(21) 国際出願番号: PCT/JP2023/015686

(22) 国際出願日: 2023年4月19日(19.04.2023)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人: 富士通株式会社(FUJITSU LIMITED)
[JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 Kanagawa (JP).

(72) 発明者: 片岡 正弘(KATAOKA, Masahiro);
〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP). 松村 量(MATSUMURA, Ryo); 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP). 藤井 泰(FUJII, Yasushi); 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP). 尾上 聡(ONOUE,

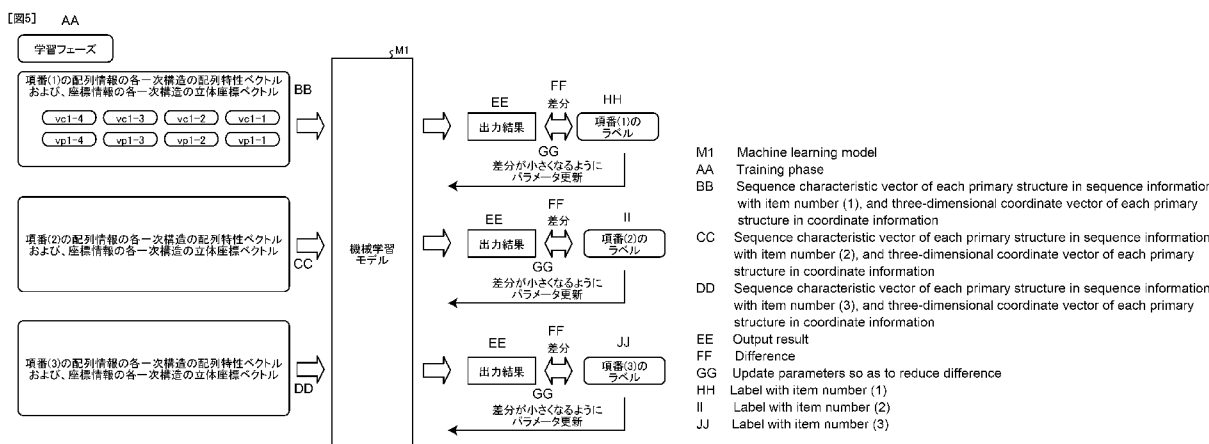
Satoshi); 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP). 城戸 まゆみ(KIDO, Mayumi); 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP). 古賀 裕美(KOGA, Hiromi); 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 弁理士法人酒井国際特許事務所(SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎ノ門ダイビルイースト 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, 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,

(54) Title: TRAINING PROGRAM, INFERENCE PROGRAM, TRAINING METHOD, INFERENCE METHOD, AND INFORMATION PROCESSING DEVICE

(54) 発明の名称: 学習プログラム、推論プログラム、学習方法、推論方法および情報処理装置



(57) Abstract: This information processing device acquires training data in which input data and a label are associated with each other, the input data containing a plurality of primary structures included in a higher-order structure of a ligand-coupled receptor and structure information relating to the plurality of primary structures, the label indicating whether or not the receptor and the ligand can be coupled to each other. The information processing device executes machine learning of a machine learning model on the basis of the training data.

(57) 要約: 情報処理装置は、リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、複数の一次構造の構造情報を有する入力データと、受容体とリガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを取得する。情報処理装置は、教師データを基にして、機械学習モデルの機械学習を実行する。

[続葉有]

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:

一 国際調査報告 (条約第21条(3))

明 細 書

発明の名称：

学習プログラム、推論プログラム、学習方法、推論方法および情報処理装置

技術分野

[0001] 本発明は、学習プログラム等に関する。

背景技術

[0002] 受容体は細胞に存在する制御タンパク質で、様々なシグナル分子を選択的に受け取る。主に細胞膜に埋まっているが、細胞質や核表面にも存在する。受容体と合体して生物学的反応を誘発するシグナル分子のことを「リガンド」という。

[0003] リガンドとなる物質は、ホルモン、一部のアミノ酸、神経伝達物質、毒素または医薬品等である。リガンドは受容体の特定の部位に対して選択的、または、特異的に高い親和性を示すことが知られている。なお、リガンド毎に異なる受容体が存在している場合が多く、細胞の種類によって、合体可能なリガンドと受容体との組み合わせは大きく異なる。

[0004] タンパク質の立体構造解析、および、化学特性解析の双方において、合体可能な受容体とリガンドとの組み合わせを推論する研究が進んでいる。

[0005] 受容体などのタンパク質の高次構造は、公開データであって、約20種類のアミノ酸の配列で構成される。図13Aは、アミノ酸の名称と略称、および記号との関係の一例を示す図である。たとえば、アミノ酸「アラニン」を示す略称と記号は「A | a」と「A」となる。他のアミノ酸の名称と略称、および記号との関係は、図13Aに示すとおりである。

[0006] アミノ酸とは、炭素（C）にアミノ基（ -NH_2 ）とカルボキシ基（ -COOH ）が結合したものであり、図13Bに、アミノ酸の一般構造式を示す。図13Bは、アミノ酸の化学構造式の一般構造式と側鎖との関係の一例を示す図である。また、中心の炭素（C）には「側鎖（R）」が結合しており

、この違いによりアミノ酸の種類が異なる。「アラニン」や「バリン」などの具体的な側鎖（R）とその化学構造式を示す。

[0007] 従来技術では、機械学習モデルを用いて、受容体とリガンドとの組み合わせが適切な組み合わせであるか否か（合体可能であるか否か）を判定している。

[0008] 図14は、従来技術を説明するための図（1）である。図14では、従来技術の学習フェーズの処理を説明する。説明の便宜上、従来技術を実行する装置を「従来装置」と表記する。従来装置は、入力データと、正解ラベルとの組を用いて、機械学習モデルM1の機械学習を実行する。

[0009] たとえば、入力データには、受容体の複数の化学構造式5とその各原子と、該受容体との合体対象となるリガンドの化学構造式6とその各原子とが含まれる。受容体の化学構造式5には、5-1, 5-2, 5-3が含まれる。正解ラベルには、入力データの受容体とリガンドとが合体可能であるか否かの情報が設定される。

[0010] 従来装置は、ベクトル辞書を用いて、受容体の複数の化学構造式5-1, 5-2, 5-3のベクトル v_{c5-1} , v_{c5-2} , v_{c5-3} をその各原子のベクトルをもとに算出する。従来装置は、ベクトル辞書を用いて、化学構造式6のベクトル v_{c6} をその各原子のベクトルをもとに算出する。従来装置は、ベクトル v_{c5-1} , v_{c5-2} , v_{c5-3} と、ベクトル v_{c6} とを積算したベクトル v_{c7} を算出する。

[0011] 従来装置は、ベクトル v_{c7} を機械学習モデルM1に入力することで、出力結果8を得る。従来装置は、従来装置は、出力結果8と、正解ラベルとの差分が小さくなるように、機械学習モデルのパラメータを更新する。

[0012] 従来装置は、他の入力データと、正解ラベルとの組についても、上記処理を繰り返し実行することで、機械学習モデルM1を訓練する。

[0013] 図15は、従来技術を説明するための図（2）である。図15では、従来技術の推論フェーズの処理を説明する。従来装置は、訓練済みの機械学習モデルM1を用いて、候補データの受容体とリガンドとが合体可能であるか否

かを推論する。

[0014] たとえば、候補データには、受容体の複数の化学構造式 10 とその各原子、該受容体との合体対象となるリガンドの化学構造式 11 とが含まれる。受容体の化学構造式 10 には、10-1, 10-2, 10-3 が含まれる。

[0015] 従来装置は、ベクトル辞書を用いて、化学構造式 10-1, 10-2, 10-3 のベクトル v_{c10-1} , v_{c10-2} , v_{c10-3} をその各原子のベクトルをもとに算出する。従来装置は、ベクトル辞書を用いて、化学構造式 11 のベクトル v_{c11} をその各原子のベクトルをもとに算出する。従来装置は、ベクトル v_{c10-1} , v_{c10-2} , v_{c10-3} と、ベクトル v_{c11} とを積算したベクトル v_{c12} を算出する。

[0016] 従来装置は、ベクトル v_{c12} を、訓練済みの機械学習モデル M1 に入力することで、出力結果 13 を得る。従来装置は、出力結果 13 が「OK (合体可能)」である場合には、候補データの受容体と、リガンドとの組み合わせが適切であると推定する。一方、従来装置は、出力結果 13 が「NG (合体可能ではない)」である場合には、候補データの受容体と、リガンドとの組み合わせが適切でないと推定する。

先行技術文献

特許文献

[0017] 特許文献1：特開 2019-028879 号公報

特許文献2：米国特許出願公開第 2022/0246233 号明細書

特許文献3：特表 2018-503171 号公報

特許文献4：米国特許出願公開第 2017/0323049 号明細書

発明の概要

発明が解決しようとする課題

[0018] たとえば、受容体とリガンドとが合体可能であるか否については、化学特性解析のためのタンパク質の配列情報だけではなく、立体構造解析のための受容体やリガンドを構成する原子の座標情報の影響を受ける。しかし、受容

体はタンパク質の高次構造（複数の一次構造）、リガンドはタンパク質の一次構造で構成されるにも係わらず、上述した従来技術では、アミノ酸の化学構造式の各原子に着目しており、推定を行うための粒度と情報量が最適ではなく、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定することができないという問題がある。

[0019] 1つの側面では、本発明は、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定することができる学習プログラム、推論プログラム、学習方法、推論方法および情報処理装置を提供することを目的とする。

課題を解決するための手段

[0020] 第1の案では、コンピュータに次の処理を実行させる。コンピュータは、リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、複数の一次構造の構造情報を有する入力データと、受容体とリガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを取得する。コンピュータは、教師データを基にして、機械学習モデルの機械学習を実行する。

発明の効果

[0021] 対象となる受容体とリガンドとが合体可能であるか否かを適切に推定することができる。

図面の簡単な説明

[0022] [図1]図1は、タンパク質構造データの一例を示す図である。

[図2]図2は、前処理フェーズの処理を説明するための図である。

[図3]図3は、Postscriptプログラムの一例を示す図である。

[図4]図4は、学習フェーズを説明するための図（1）である。

[図5]図5は、学習フェーズを説明するための図（2）である。

[図6]図6は、推論フェーズの処理を説明するための図（1）である。

[図7]図7は、推論フェーズの処理を説明するための図（2）である。

[図8]図8は、本実施例に係る情報処理装置の構成を示す機能ブロック図である。

[図9]図 9 は、前処理フェーズの処理手順を示すフローチャートである。

[図10]図 10 は、学習フェーズの処理手順を示すフローチャートである。

[図11]図 11 は、推論フェーズの処理手順を示すフローチャートである。

[図12]図 12 は、実施例の情報処理装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。

[図13A]図 13 A は、アミノ酸の名称と記号との関係の一例を示す図である。

[図13B]図 13 B は、アミノ酸の化学構造式の一般構造式と側鎖との関係の一例を示す図である。

[図14]図 14 は、従来技術を説明するための図（1）である。

[図15]図 15 は、従来技術を説明するための図（2）である。

発明を実施するための形態

[0023] 以下に、本願の開示する学習プログラム、推論プログラム、学習方法、推論方法および情報処理装置の実施例を図面に基づいて詳細に説明する。なお、この実施例によりこの発明が限定されるものではない。

実施例

[0024] 本実施例に係る情報処理装置の処理の説明を行う前に、情報処理装置が取り扱う「タンパク質構造データ」の一例について説明する。タンパク質構造データは、PDB (Protein Data Bank)から入手することができる。

[0025] 図 1 は、タンパク質構造データの一例を示す図である。たとえば、図 1 に示すタンパク質構造データ 30 は、タンパク質の構造情報として、ヘッダ領域 30 a と、配列情報領域 30 b と、座標情報領域 30 c とが含まれる。ヘッダ領域 30 a には、タンパク質に対応する分子名等が設定される。

[0026] 配列情報領域 30 b には、タンパク質に含まれるアミノ酸の配列情報が設定される。アミノ酸の配列情報は、図 13 A で説明したように、タンパク質を構成するアミノ酸の略称（3 文字）が配列された情報である。

[0027] タンパク質構造データの一連の複数のアミノ酸の配列が、タンパク質の一次構造に対応する。一次構造に含まれるアミノ酸の配列は様々なパターンが存在するが、本実施例では、各一次構造に含まれるアミノ酸の配列は、事前

に定義されているものとする。連続する複数の一次構造の並びが、タンパク質の高次構造に対応する。また、PDBには、配列情報と座標情報に、受容体とリガンドが合体した状態でデータが格納されている場合がある。たとえば、タンパク質構造データ30には、受容体の高次構造を構成する複数のタンパク質の一次構造と、リガンドを構成するタンパク質一次構造とが含まれる。

[0028] 座標情報領域30cは、タンパク質に含まれるアミノ酸を構成する複数の原子の位置（3次元座標）が設定される。本実施例では、配列情報領域30bに含まれるアミノ酸を構成する各原子の位置が、座標情報領域30cに設定されているものとする。本実施例では、複数の原子のうち、所定の原子に着目する。たとえば、所定の原子を、（1）アミノ基「N」、（2）中心の炭素（C）の側鎖の先端に位置する原子（たとえば、アミノ酸のバリン「Val」は原子「C」）、（3）カルボキシル基「O」とする。以下の説明では、着目する所定の原子を、第1原子、第2原子、第3原子と表記する。一つの一次構造に複数の第1原子が存在し、各第1原子の位置が異なる場合もある。第2原子、第3原子についても同様である。

[0029] 上記のように、リガンドと合体した受容体のタンパク質構造データを利用することで、受容体の高次構造（複数の一次構造）、リガンドの一次元構造、および、受容体とリガンドを構成するアミノ酸の所定の原子の座標情報を特定することができる。

[0030] 続いて、本実施例に係る情報処理装置の処理について説明する。本実施例に係る情報処理装置は、前処理フェーズの処理、学習フェーズの処理、推論フェーズの処理を順に実行する。以下の説明では、本実施例に係る情報処理装置を「情報処理装置100」と表記する。

[0031] 情報処理装置100が実行する前処理フェーズの処理について説明する。図2は、前処理フェーズの処理を説明するための図である。情報処理装置100は、前処理フェーズの処理を実行することで、第1ベクトル辞書142aと、第2ベクトル辞書142bとを生成する。

[0032] 情報処理装置100は、タンパク質構造データベースPDB141を有する。タンパク質構造データベースPDB141は、複数のタンパク質（受容体、あるいはリガンドと合体した受容体）に対応するタンパク質構造データを格納する。タンパク質構造データについては、図1で説明した。

[0033] 情報処理装置100が、第1ベクトル辞書142aを生成する処理について説明する。情報処理装置100は、タンパク質構造データベースPDB141の各タンパク質構造データから、複数の一次構造を抽出する。図1で説明したように、一次構造の情報は、配列情報領域30bに格納されている。

[0034] 図2の説明では、複数の一次構造（配列情報）をまとめて、「一次構造41」と表記する。図1で説明したように、一次構造は、アミノ酸配列の文字列の情報である。情報処理装置100は、アミノ酸配列の文字列に分解する。

[0035] 情報処理装置100は、一次構造41を複数のアミノ酸配列（あるいは、有機化合物の官能基配列）の文字列に分解した後、各一次構造の文字列を順番に並べる。情報処理装置100は、順番に並べた各文字列に対して、CBoWやskip-gram (Word2vec) のアルゴリズムを適用し、各アミノ酸（あるいは、官能基）をそれぞれ単語に見立て、文に相当する一次構造の各文字列のベクトルを計算する。情報処理装置100は、一次構造41の基準単位の文字列と、ベクトルとの関係を、第1ベクトル辞書142aに登録する。なお、情報処理装置100は、一次構造41を事前に定義した基準単位で分割してもよい。

[0036] 情報処理装置100は、他の一次構造に対しても、上記処理を繰り返し実行することで、他の一次構造に含まれるアミノ酸配列の文字列と、ベクトルとの関係を、第1ベクトル辞書142aに登録する。なお、情報処理装置は、アミノ酸の単位を基準単位として、ベクトルを割り当ててもよい。

[0037] 続いて、情報処理装置100が、第2ベクトル辞書142bを生成する処理について説明する。情報処理装置100は、タンパク質構造データベースPDB141の各タンパク質立体構造データから、複数の一次構造の座標情

報を抽出する。図1で説明したように、一次構造の座標情報は、座標情報領域30cに格納された情報であって、各アミノ酸の第1原子、第2原子、第3原子の位置の情報を含む。座標情報は、「構造情報」の一例である。

[0038] 情報処理装置100は、座標情報に含まれる第1原子、第2原子、第3原子の位置を結んだ3次元の線の形状を描画するPostscriptプログラムの文字列を生成する。図3は、Postscriptプログラムの一例を示す図である。たとえば、図3に示す例では、アミノ酸のバリン「Val」の第1原子50a1と、第2原子50a2と、第3原子50a3とがそれぞれ接続された3次元の線50を示す。図3に示すように、同一の原子が複数の位置に存在する場合もあり得る。たとえば、情報処理装置100は、複数の原子のうち、最も近い原子同士を接続することを繰り返すことで、3次元の線50を生成してもよい。

[0039] 情報処理装置100は、3次元の線50を描画するPostscriptプログラム51を生成する。Postscriptプログラム51には、線50を描画するための命令文（文字列）が含まれる。なお、情報処理装置100は、3次元の線を、所定の方向から、2次元平面に投影し、2次元平面に投影された線を描画するPostscriptプログラムを生成してもよい。

[0040] 図2の説明に戻る。情報処理装置100は、Postscriptプログラムに対して形態素解析を実行することで、複数の形態素（トークン）に分解する。情報処理装置100は、Postscriptプログラムを複数のトークンに分解した後に、各トークンを順番に並べる。情報処理装置100は、順番に並べた各トークンに対して、CBoWやskip-gramのアルゴリズムを適用し、各トークンをそれぞれ単語に見立て、各トークンのベクトルを計算する。情報処理装置100は、Postscriptプログラムのトークンと、ベクトルとの関係を、第2ベクトル辞書142bに登録する。

[0041] 情報処理装置100は、他の一次構造の座標情報に対しても、上記処理を繰り返し実行することで、他の一次構造の座標情報から得られるPostscriptプログラムのトークンと、ベクトルとの関係を、第2ベクトル辞書142b

に登録する。

[0042] 上記のように、情報処理装置100は、前処理フェーズの処理を実行することで、第1ベクトル辞書142a、第2ベクトル辞書142bを生成する。なお、情報処理装置100は、アミノ酸配列の文字列とベクトルとの関係を定義した第1ベクトル辞書142a、Postscriptプログラムのトークンとベクトルとの関係を定義した第2ベクトル辞書142bを外部装置等から事前に取得してもよい。情報処理装置100は、第1ベクトル辞書142a、第2ベクトル辞書142bを事前に取得する場合には、前処理フェーズの処理をスキップしてもよい。

[0043] ここではタンパク質の一次構造のアミノ酸配列をテキストの文に見立て、各アミノ酸の記号をテキストの単語に見立て、ベクトル辞書の作成方法を示した。日本語の「て」「に」「を」「は」などの平仮名は意味を持つ単語である。約20種類のアミノ酸も酸性、塩基性、中性・親水性、中性・疎水性などの化学特性を持っている。また、「あい（愛）」「あい（藍）」などの複数の平仮名で構成される単語も固有な意味を持つ。従って、タンパク質の α ヘリックスや β シートなどの規則的な立体構造を構成するモチーフと呼ばれるアミノ酸配列をベクトル辞書に追加することで、ベクトル辞書の高精度化を図ることができる。

[0044] これまで、アミノ酸配列で構成されるバイオ医薬品のリガンドについて、ベクトル辞書の作成方法を示した。一方、従来の官能基配列で構成される有機化合物の医薬品に対し、数十種類の官能基をもとに、アミノ酸配列と同様な方法で算出されたベクトル辞書を作成することで、化学特性分析や立体構造分析が可能となる。なお、約20種のアミノ酸にA～Zの文字の割り当てと同様に、官能基に対し、a～zの文字、および「！」などの記号を割り当てる方法を適用しても良い。

[0045] 次に、情報処理装置100が実行する学習フェーズの処理について説明する。図4および図5は、学習フェーズを説明するための図である。情報処理装置100は、事前に準備された教師データテーブル143を用いて、学習

フェーズの処理を実行する。

- [0046] 図4について説明する。たとえば、教師データテーブル143は、項番と、配列情報と、座標情報と、ラベルとを対応付ける。項番は、教師データテーブル143のレコード（教師データ）を識別する番号である。配列情報は、リガンドと合体した受容体の高次構造であって、かかる高次構造には、一連の複数の一次構造が含まれる。また、項番により、合体した状態で、リガンドの一次構造の前後に、受容体の高次構造を構成する、どの一次構造が隣接するか、識別することができる。座標情報は、タンパク質の高次構造に含まれる複数の一次構造の各アミノ酸の第1原子、第2原子、第3原子の位置を示す情報である。なお、一つの一次構造に対して、一つの座標情報が設定される。ラベルは、リガンドと合体した受容体が適切であるか否かを示す。たとえば、リガンドと合体した受容体が適切である場合、ラベルには「OK<たとえば、1>」が設定される。一方、リガンドと合体した受容体が適切でない場合、ラベルには「NG<たとえば、0>」が設定される。
- [0047] たとえば、項番（1）の配列情報には、先頭から順に、一次構造c1-1, c1-2, c1-3, c1-4が含まれる。たとえば、一次構造c1-1, c1-2, c1-3, c1-4のうち、一次構造c1-3を、リガンドの一次構造とする。
- [0048] 項番（1）の座標情報には、先頭から順に、座標情報e1-1, e1-2, e1-3, e1-4が含まれる。たとえば、座標情報e1-1は、一次構造c1-1に含まれる第1原子、第2原子、第3原子の位置の情報である。座標情報e1-2は、一次構造c1-2に含まれる第1原子、第2原子、第3原子の位置の情報である。座標情報e1-3は、一次構造c1-3に含まれる第1原子、第2原子、第3原子の位置の情報である。座標情報e1-4は、一次構造c1-4に含まれる第1原子、第2原子、第3原子の位置の情報である。
- [0049] ここで、情報処理装置100は、準備フェーズで作成した第1ベクトル辞書142aを用いて、化学特性解析のための各一次構造のベクトルを算出す

る。以下の説明では、一次構造のベクトルを「配列特性ベクトル」と表記する。

[0050] 項番(1)の配列情報に含まれる一次構造 $c1-1$ を用いて説明する。情報処理装置100は、一次構造 $c1-1$ を、基準単位(たとえば、予め設定された原子の単位、あるいはあらかじめ設定された原子の単位)の文字列に分解する。情報処理装置100は、一次構造 $c1-1$ の基準単位の各文字列と、第1ベクトル辞書142aとを比較することで、基準単位の各文字列のベクトルを特定する。情報処理装置100は、基準単位の各文字列のベクトルを積算することで、一次構造 $c1-1$ の配列特性ベクトル v_{c1-1} を算出する。

[0051] 情報処理装置100は、一次構造 $c1-1$ と同様にして、一次構造 $c1-2$ の配列特性ベクトル v_{c1-2} を算出する。情報処理装置100は、一次構造 $c1-3$ の配列特性ベクトル v_{c1-3} を算出する。情報処理装置100は、一次構造 $c1-4$ の配列特性ベクトル v_{c1-4} を算出する。

[0052] 続いて、情報処理装置100は、準備フェーズで作成した第2ベクトル辞書142bを用いて、各座標情報のベクトルを算出する。以下の説明では、座標情報のベクトルを「立体座標ベクトル」と表記する。

[0053] 項番(1)の座標情報に含まれる座標情報 $e1-1$ を用いて説明する。情報処理装置100は、座標情報 $e1-1$ に含まれる第1原子、第2原子、第3原子の位置を結んだ3次元の線の形状を描画するPostscriptプログラムの文字列 $p1-1$ を生成する。情報処理装置100は、Postscriptプログラム $p1-1$ に対して形態素解析を実行することで、複数の形態素(トークン)に分解する。

[0054] 情報処理装置100は、Postscriptプログラム $p1-1$ の各トークンと、第2ベクトル辞書142bとを比較することで、各トークンのベクトルを特定する。情報処理装置100は、Postscriptプログラム $p1-1$ の各トークンのベクトルを積算することで、立体座標ベクトル v_{p1-1} を算出する。

[0055] 情報処理装置100は、座標情報 $e1-1$ と同様にして、座標情報 $e1-$

2のPostscriptプログラム p_{1-2} を生成し、Postscriptプログラム p_{1-2} と第2ベクトル辞書142bとを基にして、立体座標ベクトル $v_{p_{1-2}}$ を算出する。情報処理装置100は、座標情報 e_{1-3} のPostscriptプログラム p_{1-3} を生成し、Postscriptプログラム p_{1-3} と第2ベクトル辞書142bとを基にして、立体座標ベクトル $v_{p_{1-3}}$ を算出する。情報処理装置100は、座標情報 e_{1-4} のPostscriptプログラム p_{1-4} を生成し、Postscriptプログラム p_{1-4} と第2ベクトル辞書142bとを基にして、立体座標ベクトル $v_{p_{1-4}}$ を算出する。

[0056] 図5の説明に移行する。情報処理装置100は、各項番の各一次構造の配列情報の各配列特性ベクトルおよび座標情報の各立体座標ベクトルの組を順番に機械学習モデルM1に入力し、機械学習モデルM1から出力される値が、該当のラベルに近づくように、機械学習モデルM1を訓練（パラメータを更新）する。

[0057] 機械学習モデルM1は、BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding)、Next Sentence Prediction、Transformers等のNN (Neural Network) である。

[0058] 情報処理装置100が、教師データテーブル143の項番(1)に対応する配列情報、座標情報、ラベルを用いてパラメータを更新する場合について説明する。

[0059] 情報処理装置100は、図4で説明した処理を実行し、項番(1)に対応する配列情報の各一次構造の配列特性ベクトル $v_{c_{1-1}} \sim v_{c_{1-4}}$ を算出する。また、情報処理装置100は、項番(1)に対応する座標情報の各立体座標ベクトル $v_{p_{1-1}} \sim v_{p_{1-4}}$ を算出する。

[0060] 情報処理装置100は、配列情報の配列特性ベクトルと、座標情報の立体座標ベクトルとの組を、機械学習モデルM1に順番に入力する。たとえば、情報処理装置100は、1番目に、配列特性ベクトル $v_{c_{1-1}}$ 及び立体座標ベクトル $v_{p_{1-1}}$ を、機械学習モデルM1に入力する。情報処理装置100は、2番目に、配列特性ベクトル $v_{c_{1-2}}$ 及び立体座標ベクトル $v_{p_{1-2}}$ を、機械学習モデルM1に入力する。

1-2を、機械学習モデルM1に入力する。情報処理装置100は、3番目に、配列特性ベクトル $v_c 1-3$ 及び立体座標ベクトル $v_p 1-3$ を、機械学習モデルM1に入力する。情報処理装置100は、4番目に、配列特性ベクトル $v_c 1-4$ 及び立体座標ベクトル $v_p 1-4$ を、機械学習モデルM1に入力する。

[0061] 情報処理装置100は、最後の一次構造の配列特性ベクトルと、座標情報の立体座標ベクトルとの組を機械学習モデルM1に入力した時点で、機械学習モデルM1から出力される出力結果と、項番(1)のラベルとの差分が小さくなるように、機械学習モデルM1のパラメータを更新する。

[0062] 情報処理装置100は、教師データテーブル143の項番(2)以降の配列情報、座標情報、ラベルについても、上記と同様の処理を繰り返し実行することで、機械学習モデルM1のパラメータを更新する。

[0063] 次に、情報処理装置100が実行する推論フェーズの処理について説明する。図6および図7は、推論フェーズの処理を説明するための図である。まず、図6について説明する。情報処理装置100は、ユーザから、推論対象のある受容体であって、あるリガンドと合体したある受容体の配列情報と、座標情報とを受け付ける。

[0064] たとえば、配列情報には、先頭から一次構造 $c 10-1$ 、 $c 10-2$ 、 $c 10-3$ 、 $c 10-4$ が含まれる。たとえば、一次構造 $c 10-1$ 、 $c 10-2$ 、 $c 10-3$ 、 $c 10-4$ のうち、一次構造 $c 10-3$ を、リガンドの一次構造とする。

[0065] 座標情報には、先頭から順に、座標情報 $e 10-1$ 、 $e 10-2$ 、 $e 10-3$ 、 $e 10-4$ が含まれる。たとえば、座標情報 $e 10-1$ は、一次構造 $c 10-1$ に含まれる第1原子、第2原子、第3原子の位置の情報である。座標情報 $e 10-2$ は、一次構造 $c 10-2$ に含まれる第1原子、第2原子、第3原子の位置の情報である。座標情報 $e 10-3$ は、一次構造 $c 10-3$ に含まれる第1原子、第2原子、第3原子の位置の情報である。座標情報 $e 10-4$ は、一次構造 $c 10-4$ に含まれる第1原子、第2原子、第3原

子の位置の情報である。

[0066] ここで、情報処理装置100は、第1ベクトル辞書142aを用いて、一次構造 $c10-1$ 、 $c10-2$ 、 $c10-3$ 、 $c10-4$ の配列特性ベクトル $vc10-1$ 、 $vc10-2$ 、 $vc10-3$ 、 $vc10-4$ を算出する。情報処理装置100が、第1ベクトル辞書142aを用いて、一次構造の配列特性ベクトルを算出する処理は、学習フェーズで説明した処理と同様である。

[0067] 続いて、情報処理装置100は、座標情報 $e10-1$ 、 $e10-2$ 、 $e10-3$ 、 $e10-4$ を基にして、Postscriptプログラム $p10-1$ 、 $p10-2$ 、 $p10-3$ 、 $p10-4$ を生成する。情報処理装置100が、座標情報を基にして、Postscriptプログラムを生成する処理は、学習フェーズで説明した処理と同様である。

[0068] 情報処理装置100は、第2ベクトル辞書142bを用いて、Postscriptプログラム $p10-1$ 、 $p10-2$ 、 $p10-3$ 、 $p10-4$ の立体座標ベクトル $vp10-1$ 、 $vp10-2$ 、 $vp10-3$ 、 $vp10-4$ を算出する。情報処理装置100が、第2ベクトル辞書142bを用いて、立体座標ベクトルを算出する処理は、学習フェーズで説明した処理と同様である。

[0069] 図7の説明に移行する。情報処理装置100は、図6で説明した、配列情報の配列特性ベクトル $vc10-1 \sim vc10-4$ と、座標情報の立体座標ベクトル $vp10-1 \sim vp10-4$ との組を、機械学習モデルM1に順番に入力する。

[0070] たとえば、情報処理装置100は、1番目に、配列特性ベクトル $vc10-1$ 及び立体座標ベクトル $vp10-1$ を、機械学習モデルM1に入力する。情報処理装置100は、2番目に、配列特性ベクトル $vc10-2$ 及び立体座標ベクトル $vp10-2$ を、機械学習モデルM1に入力する。情報処理装置100は、3番目に、配列特性ベクトル $vc10-3$ 及び立体座標ベクトル $vp10-3$ を、機械学習モデルM1に入力する。情報処理装置100は、4番目に、配列特性ベクトル $vc10-4$ 及び立体座標ベクトル $vp1$

0-4を、機械学習モデルM1に入力する。

[0071] 情報処理装置100は、最後の一次構造の配列特性ベクトルと、座標情報の立体座標ベクトルとの組を機械学習モデルM1に入力した時点で、機械学習モデルM1から出力される出力結果を取得する。たとえば、本実施例の機械学習モデルM1は、合体OKの確からしさを示すスコアを出力してもよい。

[0072] 情報処理装置100は、出力結果のスコアが閾値以上（合体OK）である場合には、図6の配列情報に示される受容体であって、リガンドと合体した受容体が適切である（受容体の一連の一次構造の並びも適切である）と推定する。一方、情報処理装置100は、出力結果のスコアが閾値未満（合体NG）である場合には、図6の配列情報に示される受容体であって、リガンドと合体した受容体が適切でないと推定する。

[0073] 上記のように、情報処理装置100は、学習フェーズにおいて、リガンドと合体した受容体の高次構造に含まれる複数の一次構造の配列情報、および、複数の一次構造の座標情報を有する入力データと、受容体とリガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを基にして、前記機械学習モデルの機械学習を実行する。これによって、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定するための機械学習モデルM1を生成することができる。

[0074] 情報処理装置100は、訓練済みの機械学習モデルM1に対し、対象となるリガンドと合体した受容体の高次構造に含まれる複数の一次構造の配列情報、および、複数の一次構造の座標情報を有するデータを入力する。情報処理装置100は、出力結果を利用することで、リガンドと合体した受容体が適切であるか否かを推定することができる。また、本実施例では、更に、受容体の複数の一次構造（リガンドの一次構造を含む）の並びも適切であるか否かを判定することも可能である。

[0075] 続いて、上記の推論フェーズの処理、学習フェーズの処理、推論フェーズの処理を実行する情報処理装置100の構成例について説明する。図8は、

本実施例に係る情報処理装置の構成を示す機能ブロック図である。図8に示すように、情報処理装置100は、通信部110、入力部120、表示部130、記憶部140、制御部150を有する。

[0076] 通信部110は、有線又は無線で外部装置等と接続され、外部装置等との間で情報の送受信を行う。通信部110は、NIC (Network Interface Card) 等によって実現される。通信部110は、図示しないネットワークに接続されていてもよい。通信部110は、外部装置から、タンパク質構造データベースPDB141、第1ベクトル辞書142a、第2ベクトル辞書142bの情報を受信して、記憶部140に登録してもよい。

[0077] 入力部120は、各種の情報を、情報処理装置100に入力する入力装置である。入力部120は、キーボードやマウス、タッチパネル等に対応する。たとえば、推論フェーズにおいて、ユーザは、入力部120を操作して、推論対象となる配列情報、座標情報を入力してもよい。

[0078] 表示部130は、制御部150から出力される情報を表示する表示装置である。表示部130は、液晶ディスプレイ、有機EL (Electro Luminescence) ディスプレイ、タッチパネル等に対応する。たとえば、制御部150による推定フェーズの推定結果を表示する。

[0079] 記憶部140は、タンパク質構造データベースPDB141、第1ベクトル辞書142a、第2ベクトル辞書142b、教師データテーブル143、機械学習モデルM1を有する。記憶部140は、たとえば、RAM (Random Access Memory)、フラッシュメモリ (Flash Memory) 等の半導体メモリ素子、または、ハードディスク、光ディスク等の記憶装置によって実現される。

[0080] タンパク質構造データベースPDB141は、複数のタンパク質 (受容体、あるいはリガンドと合体した受容体) に対応するタンパク質構造データを格納する。タンパク質構造データベースPDB141に関する説明は、図2で行った説明と同様である。

[0081] 第1ベクトル辞書142aは、一次構造の基本単位の文字列と、ベクトル

とを対応付けて保持する辞書である。第1ベクトル辞書142aに関するその他の説明は、図2等で示した第1ベクトル辞書142aに関する説明と同様である。

[0082] 第2ベクトル辞書142bは、座標情報から生成されるPostscriptプログラムのトークンと、ベクトルとを対応付けて保持する辞書である。第2ベクトル辞書142bに関するその他の説明は、図2等で示した第2ベクトル辞書142bに関する説明と同様である。

[0083] 教師データテーブル143は、複数の教師データを保持する。教師データは、配列情報、座標情報、ラベルを対応付ける。各教師データは、機械学習モデルM1に対する機械学習を実行する場合に利用される。教師データテーブル143のデータ構造に関する説明は、図4で示した教師データテーブル143のデータ構造に関する説明と同様である。

[0084] 機械学習モデルM1は、図5で説明したBERT、Next Sentence Prediction、Transformers等のNN等である。

[0085] 制御部150は、前処理部151と、学習処理部152と、推論処理部153とを有する。制御部150は、たとえば、CPU (Central Processing Unit) やMPU (Micro Processing Unit) により実現される。また、制御部150は、例えばASIC (Application Specific Integrated Circuit) やFPGA (Field Programmable Gate Array) 等の集積回路により実行されてもよい。

[0086] 前処理部151は、図2等で説明した前処理フェーズの処理を実行する。前処理部151は、タンパク質構造データベースPDB141から、タンパク質構造データを取得し、タンパク質構造データから、複数の一次構造と、各一次構造に対応する座標情報を取得する。

[0087] 前処理部151は、一次構造を、複数の基準単位（たとえば、アミノ酸配列）の文字列に分解し、CBoWやskip-gram (Word2vec) のアルゴリズムを適用し、各文字列にベクトルを割り当てる。前処理部151は、文字列とベクトルとの関係を、第1ベクトル辞書142aに設定する。

- [0088] 前処理部 151 は、座標情報に含まれる第 1 原子、第 2 原子、第 3 原子の位置を結んだ 3 次元の線の形状を描画する Postscript プログラムを生成する。前処理部 151 は、Postscript プログラムに対して形態素解析を実行することで、複数の形態素（トークン）に分解する。前処理部 151 は、CBoW や skip-gram (Word2vec) のアルゴリズムを適用し、各トークンにベクトルを割り当て、Postscript プログラムのトークンとベクトルとの関係を、第 2 ベクトル辞書 142b に設定する。
- [0089] 前処理部 151 のその他の処理に関する説明は、図 2 および図 3 で説明した前処理フェーズの処理の説明と同様である。なお、前処理部 151 は、第 1 ベクトル辞書 142a、第 2 ベクトル辞書 142b が事前取得されている場合には、前処理フェーズの処理をスキップしてもよい。
- [0090] 学習処理部 152 は、図 4 および図 5 で説明した学習フェーズの処理を実行する。学習処理部 152 は、教師データテーブル 143 から教師データを取得する。学習処理部 152 は、第 1 ベクトル辞書 142a を基にして、配列情報の各一次構造の配列特性ベクトルを算出する。学習処理部 152 は、座標情報に含まれる複数の座標情報から、Postscript プログラムの文字列を生成し、第 2 ベクトル辞書 142b を基にして、各立体座標ベクトルを算出する。
- [0091] 学習処理部 152 は、配列特性ベクトルと立体座標ベクトルとの組を、機械学習モデル M1 に入力し、機械学習モデル M1 の出力結果とラベルとの差分が小さくなるように、誤差逆伝播法等に基づいて、機械学習モデル M1 のパラメータを更新する。学習処理部 152 は、教師データテーブル 143 に格納された各教師データを用いて、上記処理を繰り返し実行する。
- [0092] 学習処理部 152 が実行するその他の処理に関する説明は、図 4 および図 5 で説明した学習フェーズの処理の説明と同様である。
- [0093] 推論処理部 153 は、図 6 および図 7 で説明した推論フェーズの処理を実行する。推論処理部 153 は、外部装置または入力部 120 から、推論対象となる配列情報および座標情報を取得する。

- [0094] 推論処理部 153 は、第 1 ベクトル辞書 142 a を用いて、配列情報の各一次構造の配列特性ベクトルを算出する。学習処理部 152 は、座標情報に含まれる複数の座標情報から、Postscript プログラムを生成し、第 2 ベクトル辞書 142 b を基にして、各立体座標ベクトルを算出する。
- [0095] 推論処理部 153 は、配列特性ベクトルと、立体座標ベクトルとの組を先頭から順番に、機械学習モデル M1 に入力する。推論処理部 153 は、最後の一次構造の配列特性ベクトルと、座標情報の立体座標ベクトルとの組を機械学習モデル M1 に入力した時点で、機械学習モデル M1 から出力される出力結果を取得する。
- [0096] 推論処理部 153 は、出力結果のスコアが閾値以上（合体 OK）である場合には、推定対象の配列情報に示される受容体であって、リガンドと合体した受容体が適切である（受容体の一連の一次構造の並びも適切である）と推定する。一方、推論処理部 153 は、出力結果のスコアが閾値未満（合体 NG）である場合には、推定対象の配列情報に示される受容体であって、リガンドと合体した受容体が適切でないと推定する。
- [0097] 推論処理部 153 は、推定結果を、表示部 130 に出力して表示させる。
- [0098] 推論処理部 153 が実行するその他の処理に関する説明は、図 6 および図 7 で説明した推論フェーズの処理と同様である。
- [0099] 次に、本実施例に係る情報処理装置 100 の処理手順の一例について説明する。図 9 は、前処理フェーズの処理手順を示すフローチャートである。図 9 に示すように、情報処理装置 100 の前処理部 151 は、タンパク質構造データベース PDB 141 から、タンパク質構造データを取得する（ステップ S101）。前処理部 151 は、タンパク質構造データから、複数の一次構造と、各一次構造に対応する座標情報を取得する（ステップ S102）。
- [0100] 前処理部 151 は、一次構造を複数の基準単位の文字列に分解し、各文字列にベクトルを割り当てる（ステップ S103）。前処理部 151 は、文字列とベクトルとの関係を、第 1 ベクトル辞書 142 a に設定する（ステップ S104）。

- [0101] 前処理部151は、座標情報に含まれる第1原子、第2原子、第3原子の位置を結んだ線の形状を描画するPostscriptプログラムを生成する（ステップS105）。前処理部151は、Postscriptプログラムの文字列を複数のトークンに分解し、各トークンにベクトルを割り当てる（ステップS106）。
- [0102] 前処理部151は、トークンとベクトルとの関係を、第2ベクトル辞書142bに設定する（ステップS107）。
- [0103] 図10は、学習フェーズの処理手順を示すフローチャートである。図10に示すように、情報処理装置100の学習処理部152は、教師データテーブル143から教師データ（配列情報および座標情報）を取得する（ステップS201）。学習処理部152は、第1ベクトル辞書142aを基にして、配列情報に含まれる各一次構造の配列特性ベクトルを算出する（ステップS202）。
- [0104] 学習処理部152は、複数の座標情報から複数のPostscriptプログラムの文字列を生成する（ステップS203）。学習処理部152は、第2ベクトル辞書142bを基にして、各Postscriptプログラムから、立体座標ベクトルを算出する（ステップS204）。
- [0105] 学習処理部152は、配列特性ベクトルと立体座標ベクトルとの組を、機械学習モデルM1に入力する（ステップS205）。学習処理部152は、機械学習モデルM1の出力結果とラベルとの差分を算出する（ステップS206）。
- [0106] 学習処理部152は、差分が小さくなるように、機械学習モデルM1のパラメータを更新する（ステップS207）。学習処理部152は、処理を継続する場合には（ステップS208, Yes）、ステップS201に移行する。一方、学習処理部152は、処理を継続しない場合には（ステップS208, No）、処理を終了する。
- [0107] 図11は、推論フェーズの処理手順を示すフローチャートである。図11に示すように、情報処理装置100の推論処理部153は、入力部120か

ら、推論対象となる受容体とリガンドの配列情報および座標情報を取得する（ステップS301）。推論処理部153は、第1ベクトル辞書142aを基にして、配列情報に含まれる各一次構造の配列特性ベクトルを算出する（ステップS302）。

[0108] 推論処理部153は、複数の座標情報から複数のPostscriptプログラムを生成する（ステップS303）。推論処理部153は、第2ベクトル辞書142bを基にして、各Postscriptプログラムから、立体座標ベクトルを算出する（ステップS304）。

[0109] 推論処理部153は、一次構造の配列特性ベクトルと立体座標ベクトルとの組を、機械学習モデルM1に入力する（ステップS305）。推論処理部153は、機械学習モデルM1の出力結果を基にして、対象となる受容体（リガンドと合体した受容体）が適切であるか否かを判定する（ステップS306）。推論処理部153は、判定結果を表示部130に表示させる（ステップS307）。

[0110] 次に、本実施例に係る情報処理装置100の効果について説明する。情報処理装置100は、学習フェーズにおいて、リガンドと合体した受容体の高次構造に含まれる複数の一次構造の配列情報および一次構造の座標情報を有する入力データと、正解ラベルとを対応付けた教示データを用いて、機械学習モデルM1を訓練する。また、合体した状態で、リガンドの一次構造の前後に、受容体の高次構造を構成する、どの一次構造が隣接するか、識別することができる。これによって、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定するための機械学習モデルM1を生成することができる。

[0111] 情報処理装置100は、一次構造と該一次構造に対応する座標情報との組を順番に、機械学習モデルM1に入力し、機械学習モデルM1の出力結果と、ラベルとの差が小さくなるように、機械学習モデルM1の機械学習を実行する。これによって、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定するだけでなく、リガンドと合体した受容体の一次構造の配

列が適切であるか否かを推定するための機械学習モデルM1を生成することができる。

[0112] 情報処理装置100は、複数の一次構造に含まれる所定の原子の位置を基にして、座標情報を生成する。これによって、一次構造の配列と、原子の位置との双方を用いて、リガンドと合体した受容体の一次構造の配列が適切であるか否かを推定するための機械学習モデルM1を生成することができる。

[0113] 情報処理装置100は、複数の一次構造に含まれる所定の原子の位置を結んだ線分を描画するPostScriptプログラムの文字列を、ベクトルに変換した情報を、座標情報として生成する。これによって、所定の原子の位置をベクトルとして取り扱うことができ、機械学習モデルM1に対する機械学習を効率的に実行することができる。

[0114] 情報処理装置100は、タンパク質の基本単位の文字列とベクトルとを対応付けた辞書を基にして、一次構造をベクトルに変換した情報を、配列情報として生成する。これによって、一次構造をベクトルとして取り扱うことができ、機械学習モデルM1に対する機械学習を効率的に実行することができる。

[0115] 情報処理装置100は、推論フェーズにおいて、対象の受容体（リガンドと合体した受容体）の高次構造に含まれる複数の一次構造の配列情報および一次構造の座標情報を有する入力データと、訓練済みの機械学習モデルM1に入力する。これによって、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定することができる。

[0116] 情報処理装置100は、一次構造と該一次構造に対応する座標情報との組を順番に、訓練済みの機械学習モデルM1に入力し、機械学習モデルM1の出力結果を得る。これによって、対象となる受容体とリガンドとが合体可能であるか否かを適切に推定するだけでなく、リガンドと合体した受容体の一次構造の配列が適切であるか否かを推定することができる。

[0117] 次に、上記実施例に示した情報処理装置100と同様の機能を実現するコンピュータのハードウェア構成の一例について説明する。図12は、実施例

の情報処理装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。

[0118] 図12に示すように、コンピュータ300は、各種演算処理を実行するCPU301と、ユーザからのデータの入力を受け付ける入力装置302と、ディスプレイ303とを有する。また、コンピュータ300は、有線または無線ネットワークを介して、外部装置等との間でデータの授受を行う通信装置304と、インタフェース装置305とを有する。また、コンピュータ300は、各種情報を一時記憶するRAM306と、ハードディスク装置307とを有する。そして、各装置301～307は、バス308に接続される。

[0119] ハードディスク装置307は、前処理プログラム307a、学習処理プログラム307b、推論処理プログラム307cを有する。また、CPU301は、各プログラム307a～307cを読み出してRAM306に展開する。

[0120] 前処理プログラム307aは、前処理プロセス306aとして機能する。学習処理プログラム307bは、学習処理プロセス306bとして機能する。推論処理プログラム307cは、推論処理プロセス306cとして機能する。

[0121] 前処理プロセス306aの処理は、前処理部151の処理に対応する。学習処理プロセス306bの処理は、学習処理部152の処理に対応する。推論処理プロセス306cの処理は、推論処理部153の処理に対応する。

[0122] なお、各プログラム307a～307cについては、必ずしも最初からハードディスク装置307に記憶させておかなくても良い。例えば、コンピュータ300に挿入されるフレキシブルディスク(FD)、CD-ROM、DVD、光磁気ディスク、ICカードなどの「可搬用の物理媒体」に各プログラムを記憶させておく。そして、コンピュータ300が各プログラム307a～307cを読み出して実行するようにしてもよい。

符号の説明

[0123]	1 0 0	情報処理装置
	1 1 0	通信部
	1 2 0	入力部
	1 3 0	表示部
	1 4 0	記憶部
	1 4 1	タンパク質構造ＤＢ
	1 4 2 a	第１ベクトル辞書
	1 4 2 b	第２ベクトル辞書
	1 4 3	教師データテーブル
	1 5 0	制御部
	1 5 1	前処理部
	1 5 2	学習処理部
	1 5 3	推論処理部

請求の範囲

- [請求項1] リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、前記複数の一次構造の構造情報を有する入力データと、前記受容体と前記リガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを取得し、
- 前記教師データを基にして、機械学習モデルの機械学習を実行する処理をコンピュータに実行させることを特徴とする学習プログラム。
- [請求項2] 前記入力データの前記高次構造は、前記リガンドの一次構造と、前記リガンドの一次構造以外の複数の一次構造とを含み、
- 前記機械学習を実行する処理は、前記一次構造と構造情報との組を順番に、前記機械学習モデルに入力し、前記機械学習モデルの出力結果と、前記ラベルとの差が小さくなるように、前記機械学習モデルの機械学習を実行することを特徴とする請求項1に記載の学習プログラム。
- [請求項3] 前記一次構造に含まれる所定の原子の位置を基にして、前記構造情報を生成する処理を更にコンピュータに実行させることを特徴とする請求項1に記載の学習プログラム。
- [請求項4] 前記一次構造に含まれる所定の原子の位置を結んだ線分を描画するPostScriptの文字列を、ベクトルに変換する処理を更にコンピュータに実行させることを特徴とする請求項2に記載の学習プログラム。
- [請求項5] 前記一次構造を、タンパク質のアミノ酸配列や有機化合物の官能基配列の文字列に分割し、各文字や各文字列にベクトルを割り当てることで、前記一次構造をベクトルに変換する処理を更にコンピュータに実行させることを特徴とする請求項1に記載の学習プログラム。
- [請求項6] 推論対象となる対象受容体の対象高次構造であって、対象リガンドと合体した前記対象受容体の前記対象高次構造に含まれる複数の対象一次構造、および、前記複数の対象一次構造に対応する複数の対象構

造情報を取得し、

リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、前記複数の一次構造の構造情報を有する入力データと、前記受容体と前記リガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを基にして機械学習が実行された機械学習モデルに対して、前記複数の対象一次構造、前記複数の対象構造情報を入力することで、前記対象受容体が適切であるか否かを推論する

処理をコンピュータに実行させることを特徴とする推論プログラム

。

[請求項7] 前記対象高次構造は、前記対象リガンドの対象一次構造と、前記対象リガンドの対象一次構造以外の複数の対象一次構造とを含み、

前記推論する処理は、前記対象一次構造と前記対象構造情報との組を順番に、前記機械学習モデルに入力し、前記機械学習モデルの出力結果と基にして、前記対象受容体が適切であるか否かを推論することを特徴とする請求項6に記載の推論プログラム。

[請求項8] リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、前記複数の一次構造の構造情報を有する入力データと、前記受容体と前記リガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを取得し、

前記教師データを基にして、機械学習モデルの機械学習を実行する処理をコンピュータが実行することを特徴とする学習方法。

[請求項9] 前記入力データの前記高次構造は、前記リガンドの一次構造と、前記リガンドの一次構造以外の複数の一次構造とを含み、

前記機械学習を実行する処理は、前記一次構造と構造情報との組を順番に、前記機械学習モデルに入力し、前記機械学習モデルの出力結果と、前記ラベルとの差が小さくなるように、前記機械学習モデルの機械学習を実行することを特徴とする請求項8に記載の学習方法。

[請求項10] 前記一次構造に含まれる所定の原子の位置を基にして、前記構造情

報を生成する処理を更にコンピュータが実行することを特徴とする請求項8に記載の学習方法。

[請求項11] 前記一次構造に含まれる所定の原子の位置を結んだ線分を描画するPostScriptの文字列を、ベクトルに変換する処理を更にコンピュータが実行することを特徴とする請求項9に記載の学習方法。

[請求項12] 前記一次構造を、タンパク質のアミノ酸配列や有機化合物の官能基配列の文字列に分割し、各文字や各文字列にベクトルを割り当てることで、前記一次構造をベクトルに変換する処理を更にコンピュータが実行することを特徴とする請求項8に記載の学習方法。

[請求項13] 推論対象となる対象受容体の対象高次構造であって、対象リガンドと合体した前記対象受容体の前記対象高次構造に含まれる複数の対象一次構造、および、前記複数の対象一次構造に対応する複数の対象構造情報を取得し、

リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、前記複数の一次構造の構造情報を有する入力データと、前記受容体と前記リガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを基にして機械学習が実行された機械学習モデルに対して、前記複数の対象一次構造、前記複数の対象構造情報を入力することで、前記対象受容体が適切であるか否かを推論する

処理をコンピュータが実行することを特徴とする推論方法。

[請求項14] 前記対象高次構造は、前記対象リガンドの対象一次構造と、前記対象リガンドの対象一次構造以外の複数の対象一次構造とを含み、

前記推論する処理は、前記対象一次構造と前記対象構造情報との組を順番に、前記機械学習モデルに入力し、前記機械学習モデルの出力結果と基にして、前記対象受容体が適切であるか否かを推論することを特徴とする請求項13に記載の推論方法。

[請求項15] リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、前記複数の一次構造の構造情報を有する入力データと、前記

受容体と前記リガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを取得し、

前記教師データを基にして、機械学習モデルの機械学習を実行する処理を実行する制御部を有する情報処理装置。

[請求項16] 前記入力データの前記高次構造は、前記リガンドの一次構造と、前記リガンドの一次構造以外の複数の一次構造とを含み、

前記制御部は、前記一次構造と構造情報との組を順番に、前記機械学習モデルに入力し、前記機械学習モデルの出力結果と、前記ラベルとの差が小さくなるように、前記機械学習モデルの機械学習を実行することを特徴とする請求項15に記載の情報処理装置。

[請求項17] 前記制御部は、前記一次構造に含まれる所定の原子の位置を基にして、前記構造情報を生成する処理を更に実行することを特徴とする請求項15に記載の情報処理装置。

[請求項18] 前記制御部は、前記一次構造に含まれる所定の原子の位置を結んだ線分を描画するPostScriptの文字列を、ベクトルに変換する処理を更に実行することを特徴とする請求項17に記載の情報処理装置。

[請求項19] 前記制御部は、前記一次構造を、タンパク質のアミノ酸配列や有機化合物の官能基配列の文字列に分割し、各文字列にベクトルを割り当てることで、前記一次構造をベクトルに変換する処理を更に実行することを特徴とする請求項15に記載の情報処理装置。

[請求項20] 推論対象となる対象受容体の対象高次構造であって、対象リガンドと合体した前記対象受容体の前記対象高次構造に含まれる複数の対象一次構造、および、前記複数の対象一次構造に対応する複数の対象構造情報を取得し、

リガンドと合体した受容体の高次構造に含まれる複数の一次構造、および、前記複数の一次構造の構造情報を有する入力データと、前記受容体と前記リガンドとが合体可能であるか否かを示すラベルとが対応づけられた教師データを基にして機械学習が実行された機械学習モ

デルに対して、前記複数の対象一次構造、前記複数の対象構造情報を入力することで、前記対象受容体が適切であるか否かを推論する処理を実行する制御部を有する情報処理装置。

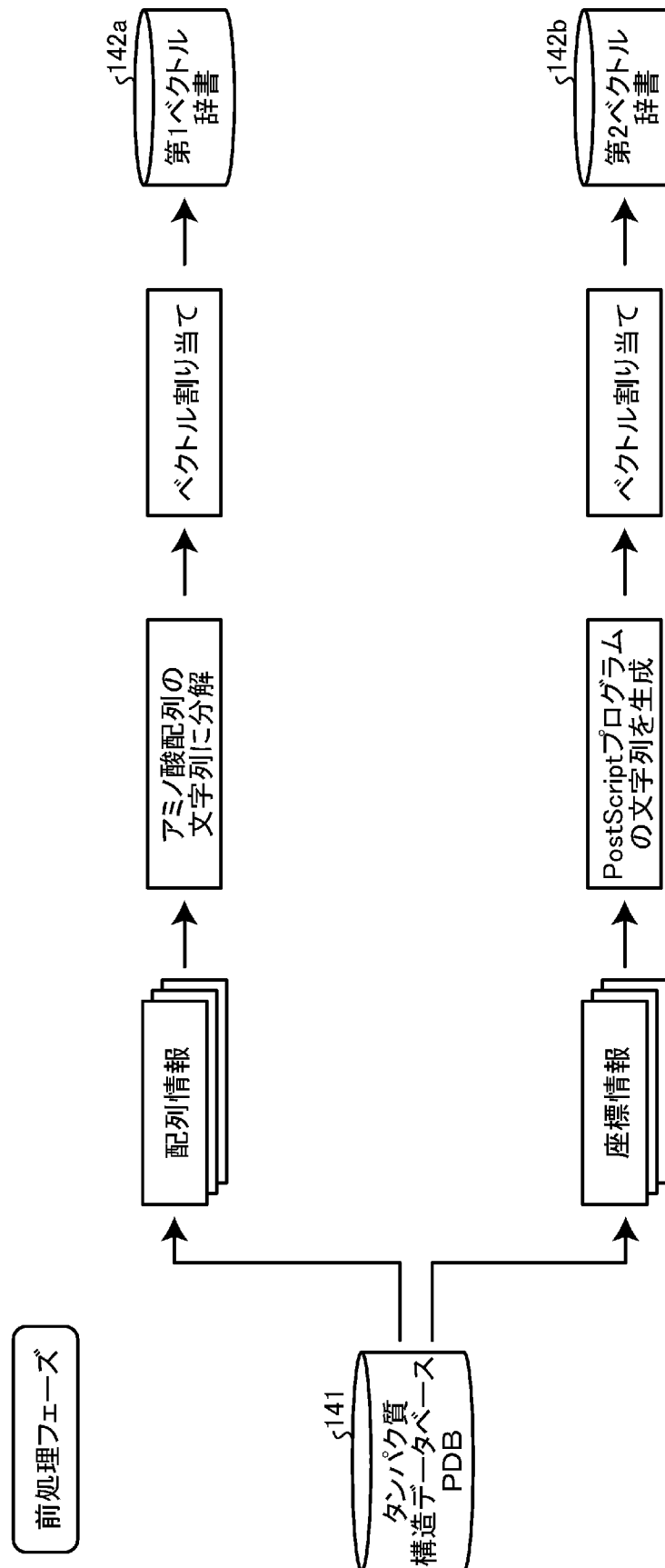
[請求項21] 前記対象高次構造は、前記対象リガンドの対象一次構造と、前記対象リガンドの対象一次構造以外の複数の対象一次構造とを含み、前記制御部は、前記対象一次構造と前記対象構造情報との組を順番に、前記機械学習モデルに入力し、前記機械学習モデルの出力結果と基にして、前記対象受容体が適切であるか否かを推論することを特徴とする請求項20に記載の情報処理装置。

[図1]

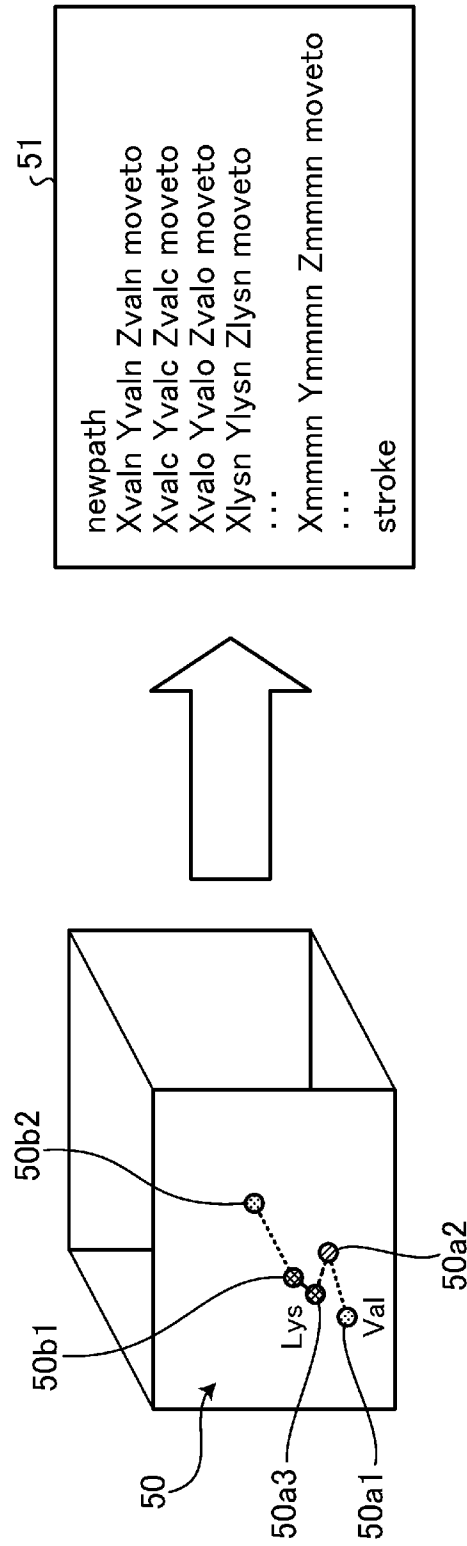
30

HEADER	HYDROLASE	04-MAY-99	1CLV
TITLE	YELLOW MEAL WORM ALPHA-AMYLASE IN COMPLEX WITH THE AMARANTH		
TITLE	2 ALPHA-AMYLASE INHIBITOR		
COMPND	MOL_ID: 1;		
COMPND	2 MOLECULE: PROTEIN (ALPHA-AMYLASE);		
COMPND	3 CHAIN: A;		
COMPND	4 SYNONYM: ALPHA-1,4-GLUCAN-4-GLUCANOHYDROLASE;		
COMPND	5 EC: 3.2.1.1;		
COMPND	6 MOL_ID: 2;		
COMPND	7 MOLECULE: PROTEIN (ALPHA-AMYLASE INHIBITOR);		
COMPND	8 CHAIN: I;		
COMPND	9 ENGINEERED: YES		
SOURCE	MOL_ID: 1;		
SOURCE	ORGANISM_SCIENTIFIC: TENEBRIO CRUCIATUM COMMON: YELLOW MEALWORM TAXID: 7067		
REMARK 999 IS	(BLANK LINE)		
DBREF	1CLV	A	1 P56634 AMY_TENMO
DBREF	1CLV	I	501 P80403 IAAI_AMAHP
SEQADV	1CLV	PCA	A 1 UNP P56634 GLN 1 SEE REMARK 999
SEQRES	1	A	471 VAL LYS ASP ALA ASN PHE ALA SER GLY ARG ASN SER ILE
SEQRES	2	A	471 VAL LYS LEU PHE GLU TRP LYS TRP ASN ASP ILE ALA ASP
SEQRES	3	A	471 GLU LYS GLU ARG PHE LEU GLN PRO GLN GLY PHE GLY GLY
SEQRES	4	A	471 VAL GLU ILE SER PRO PRO ASN GLU TYR LEU VAL ALA ASP
SEQRES	5	A	471 GLY ARG PRO TRP TRP GLU ARG TYR GLN PRO VAL SER TYR
SEQRES	6	A	471 ILE ILE ASN THR ARG SER GLY ASP GLU SER ALA PHE THR
SEQRES	7	A	471 ASP MET THR ARG ARG CYS ASN ASP ALA GLY VAL ARG ILE
SEQRES	8	A	471 TYR VAL ASP ALA VAL MET THR GLY MET ASN
SCALE3	0.000	0.000	0.000
HETATM	1	N	VAL 39.070 7.713 8.323
HETATM	2	CA	VAL A 39.380 8.263 8.128 1.00 16.19
HETATM	3	CB	VAL A 1 31.390 7.193 8.612 1.00 16.93
HETATM	4	CG	VAL A 1 30.495 5.943 8.987 1.00 19.39
HETATM	5	CD	VAL A 1 29.101 6.476 8.787 1.00 22.92
HETATM	6	OE	VAL A 1 28.089 5.796 9.037 1.00 13.70
HETATM	7	O	VAL A 1 30.667 8.643 6.676 1.00 14.12
HETATM	8	U	LYS A 1 31.514 9.493 6.417 1.00 14.51
ATOM	9	N	LYS A 2 29.983 7.994 5.735 1.00 13.28
ATOM	10	CA	LYS A 2 30.178 8.269 4.313 1.00 16.12
ATOM	11	C	LYS A 2 28.999 8.963 3.640 1.00 16.12

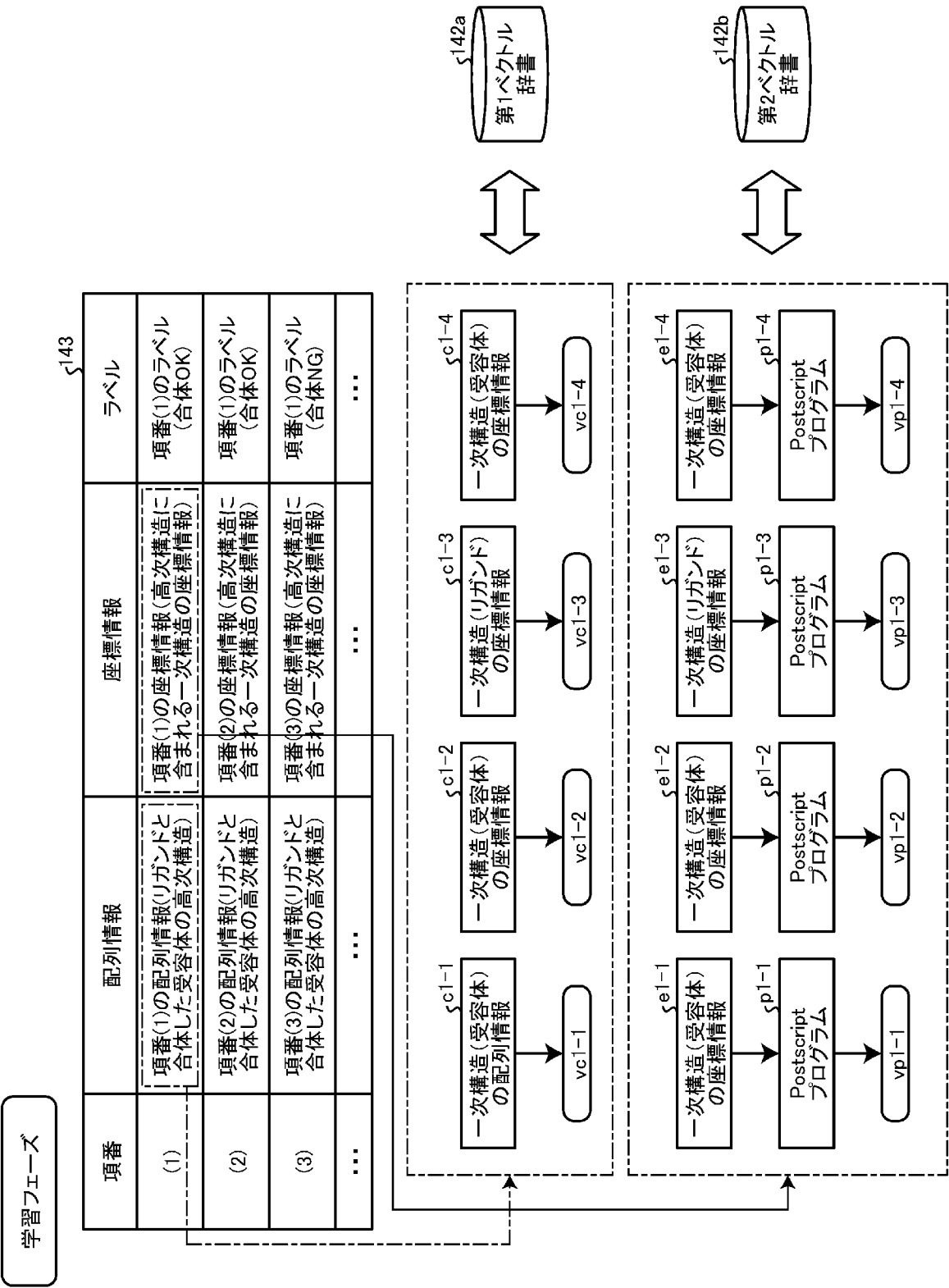
[図2]



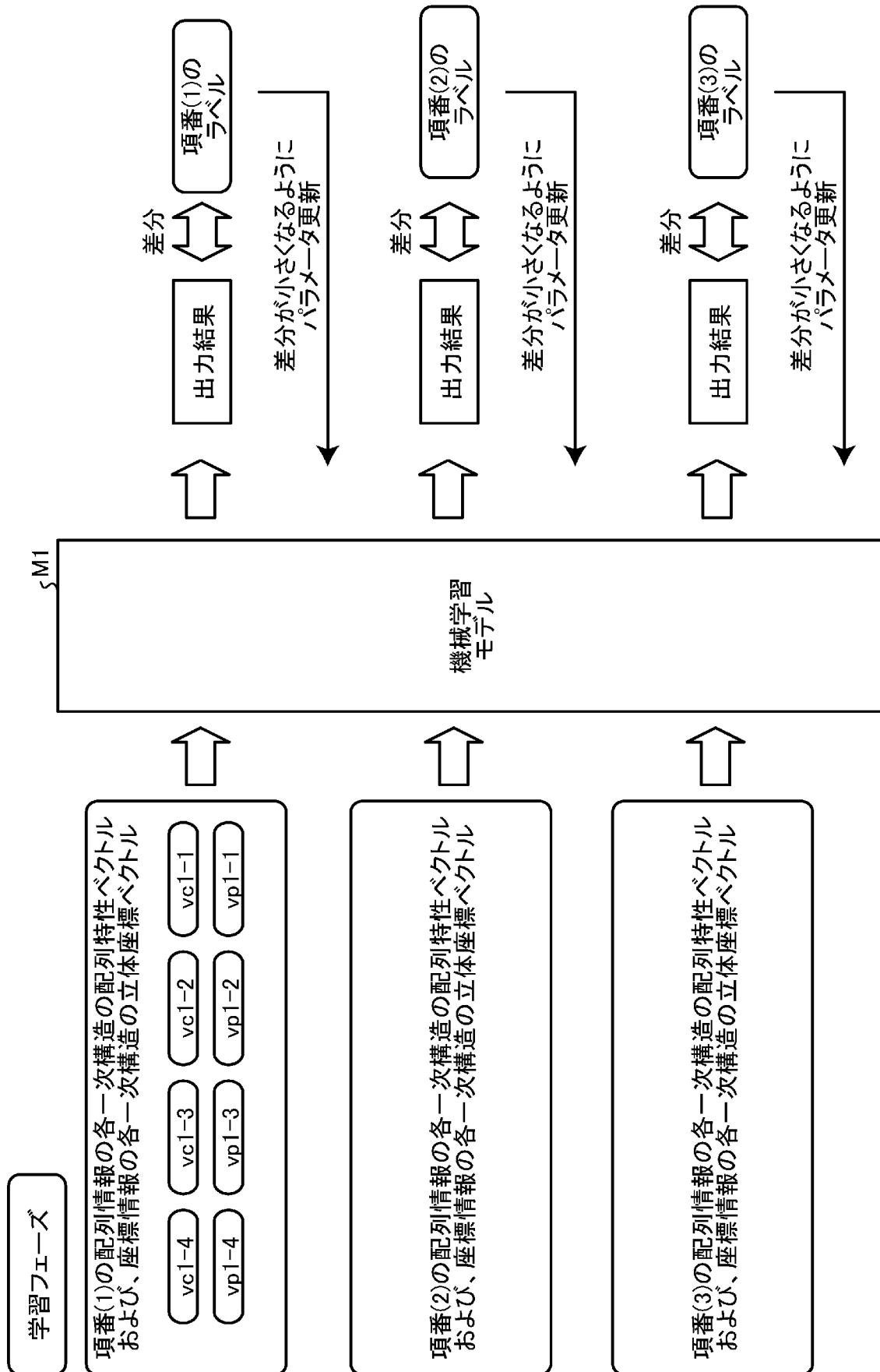
[図3]



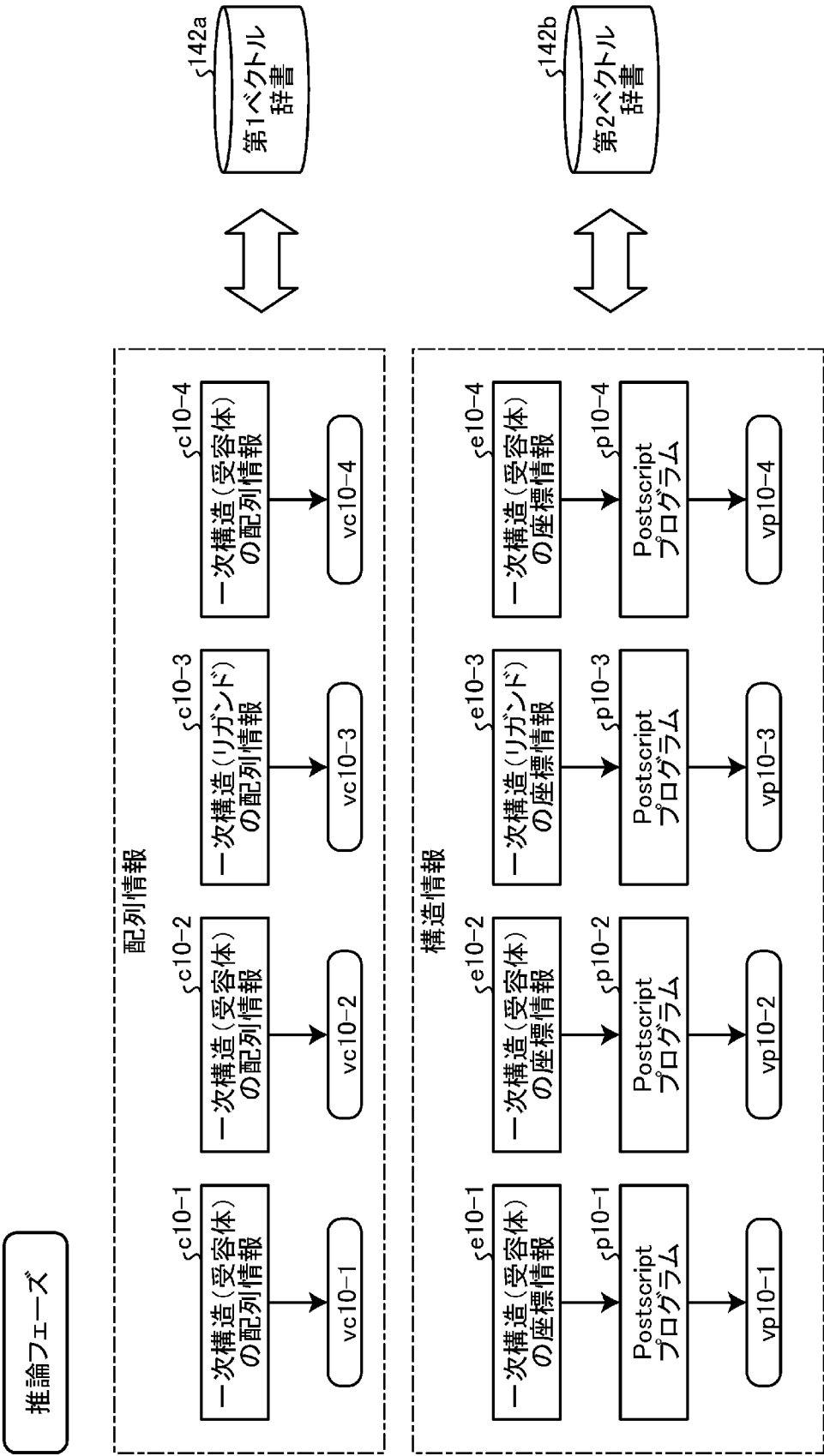
[図4]



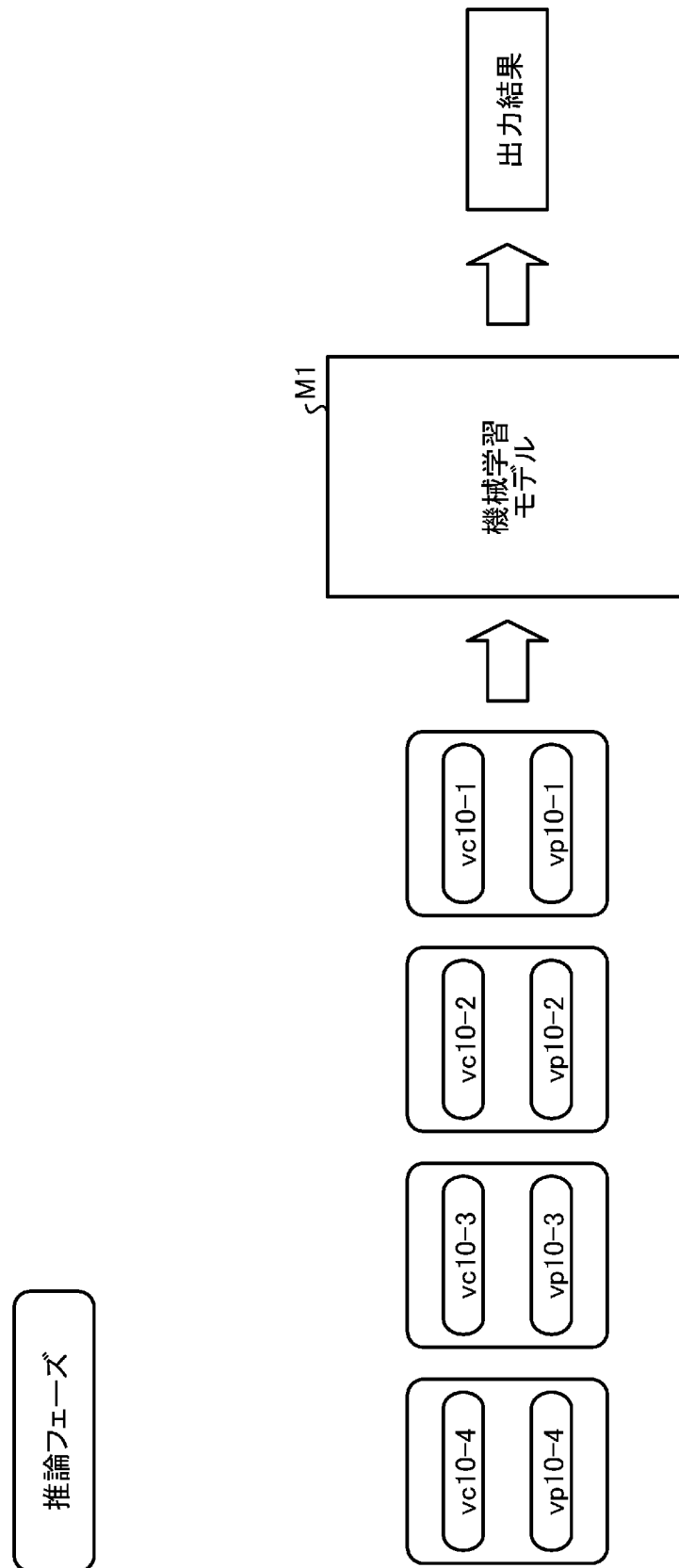
[図5]



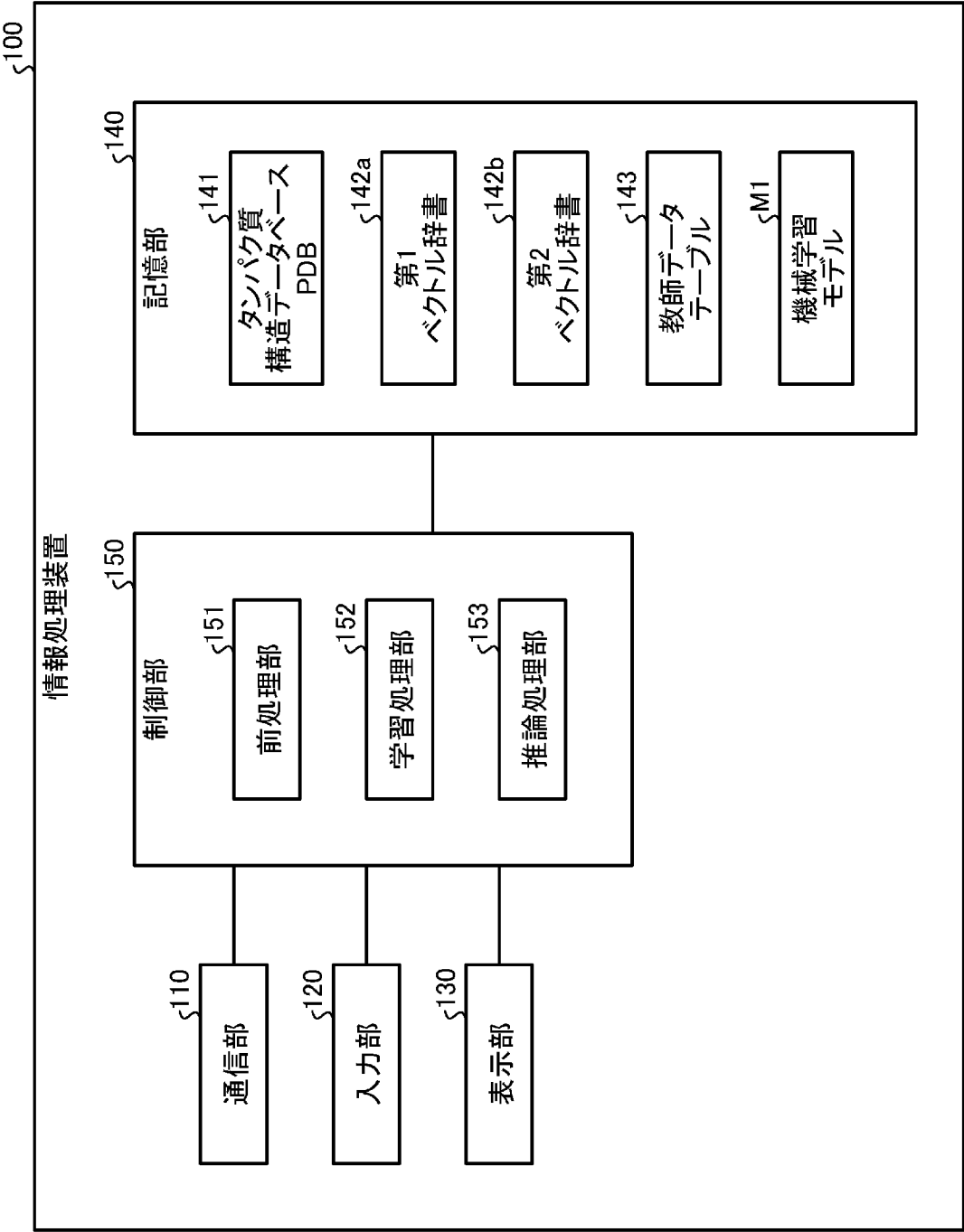
[図6]



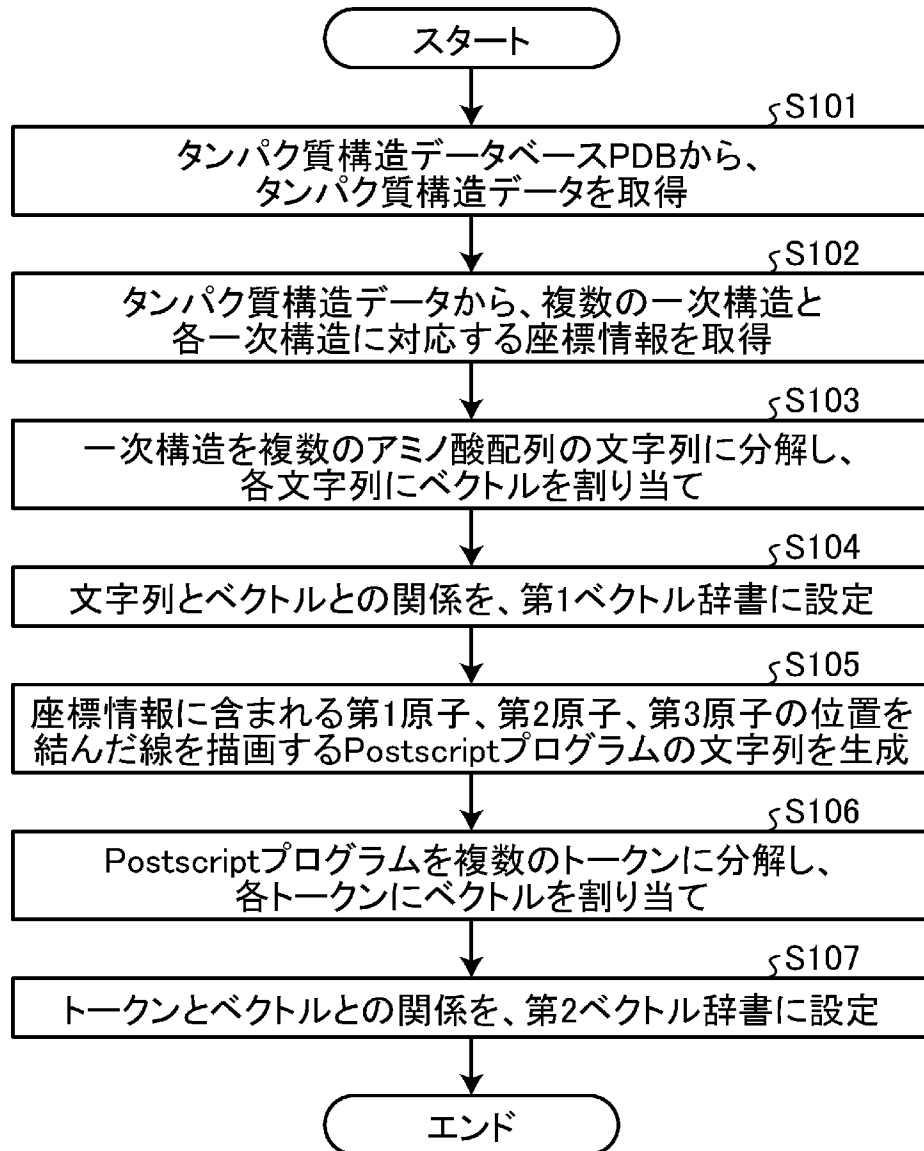
[図7]



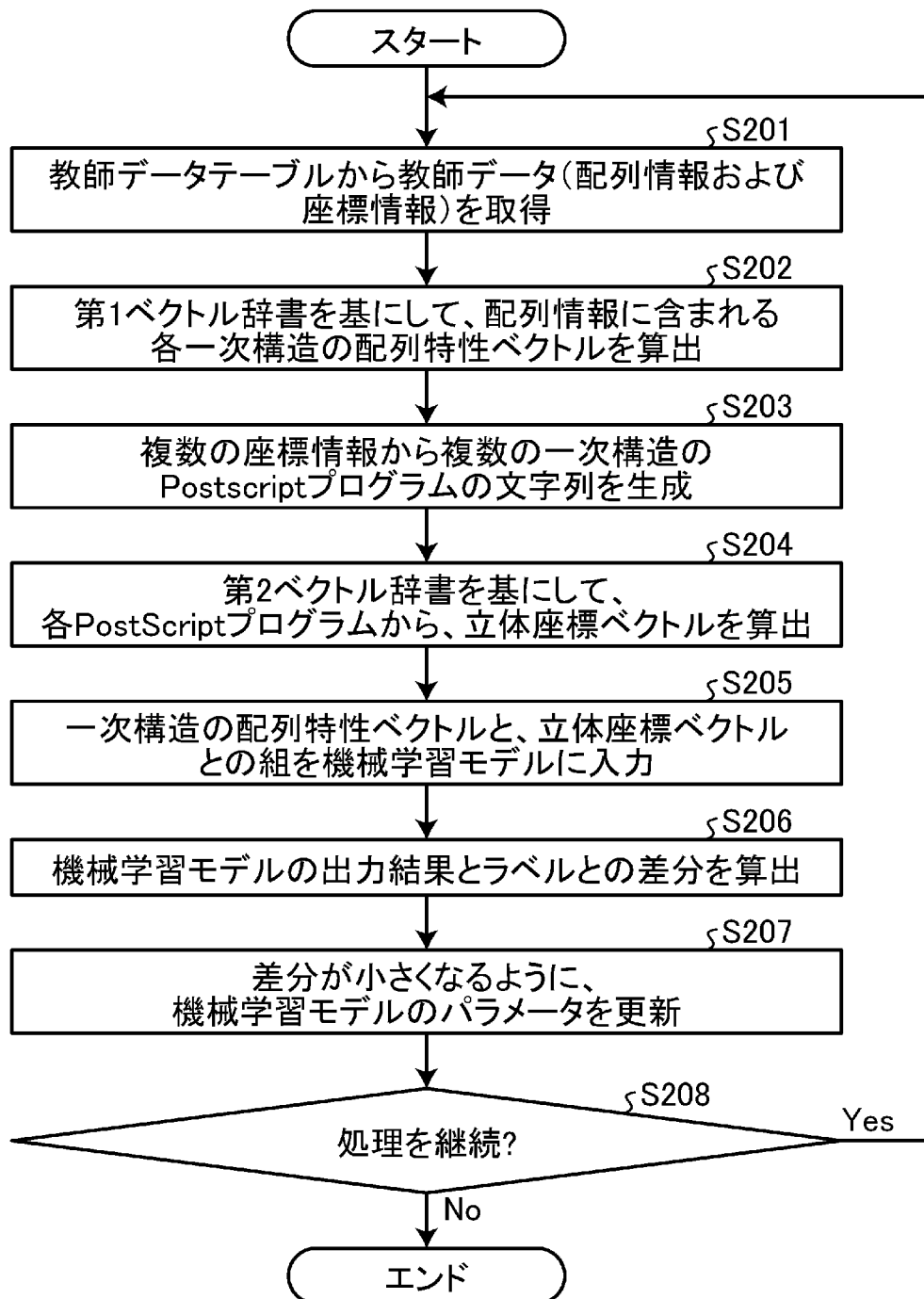
[図8]



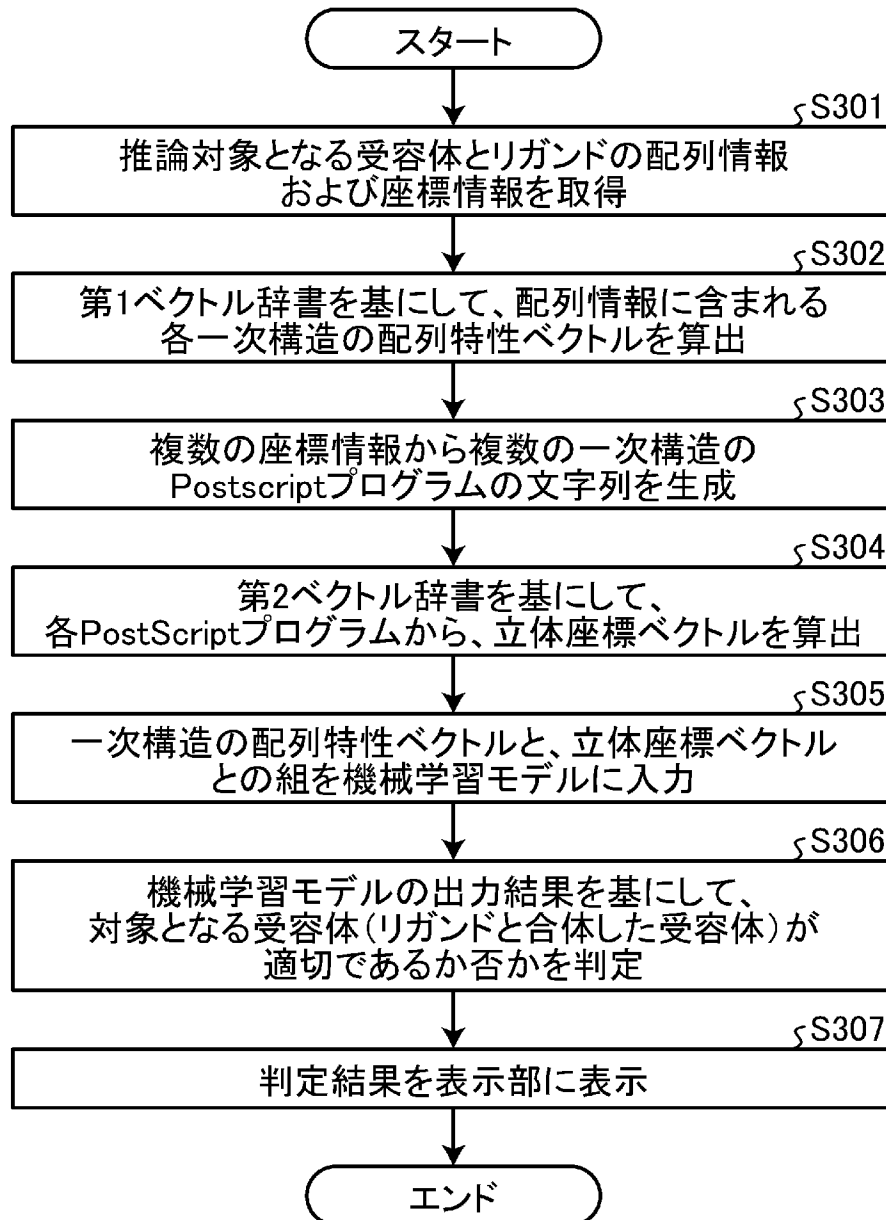
[図9]



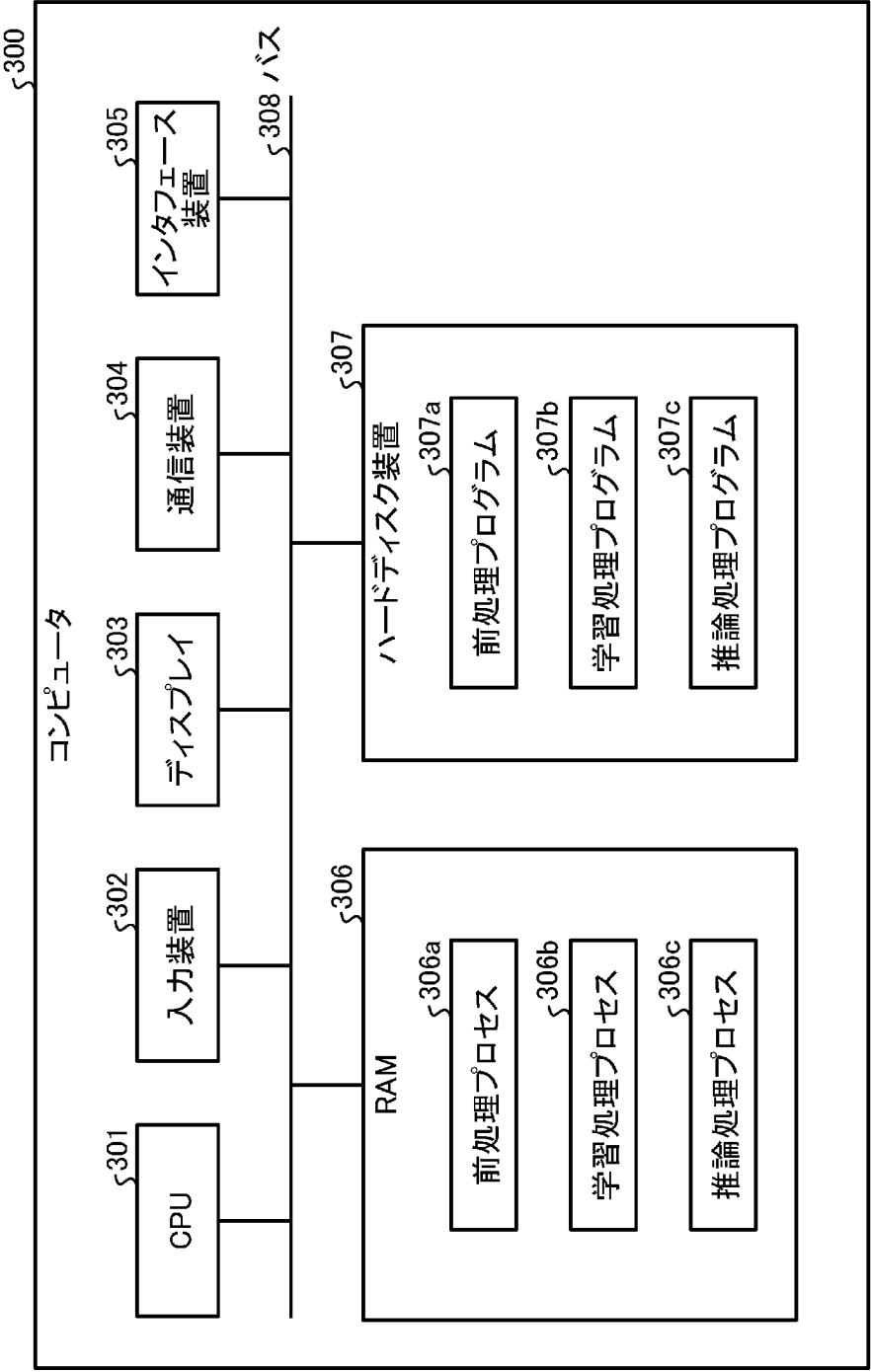
[図10]



[図11]



[図12]

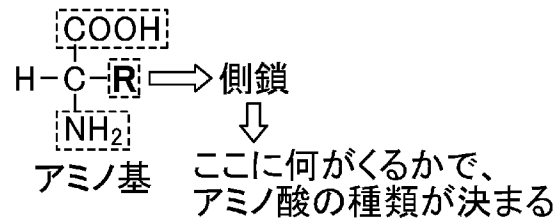


[図13A]

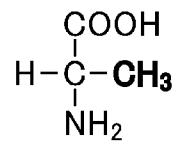
アミノ酸		
名称	略称	記号
アラニン	Ala	A
システイン	Cys	C
アスパラギン酸	Asp	D
グルタミン酸	Glu	E
フェニルアラニン	Phe	F
グリシン	Gly	G
ヒスチジン	His	H
イソロイシン	Ile	I
リシン	Lys	K
ロイシン	Leu	L
メチオニン(開始)	Met	M
アスパラギン	Asn	N
プロリン	Pro	P
グルタミン	Gln	Q
アルギニン	Arg	R
トレオニン	Thr	T
バリン	Val	V
トリプトファン	Trp	W
チロシン	Tyr	Y
(終止)		
(終止)		
決定されていないアミノ酸	Xxx	X
グルタミン酸か グルタミンのどちらか	Glx	Z

[図13B]

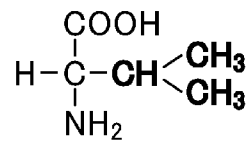
アミノ酸の一般化学構造式



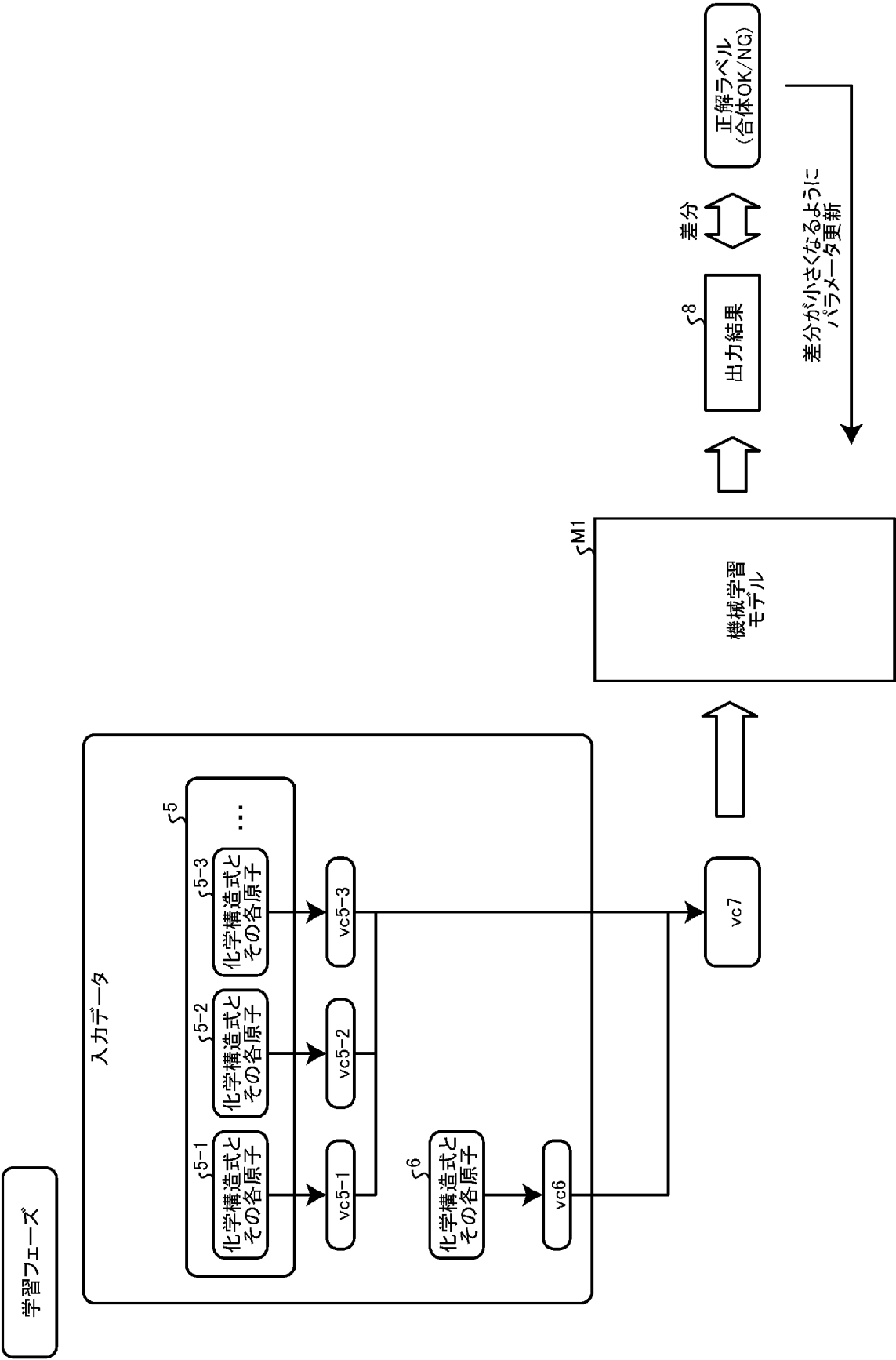
アラニン「Ala/A」の化学構造式



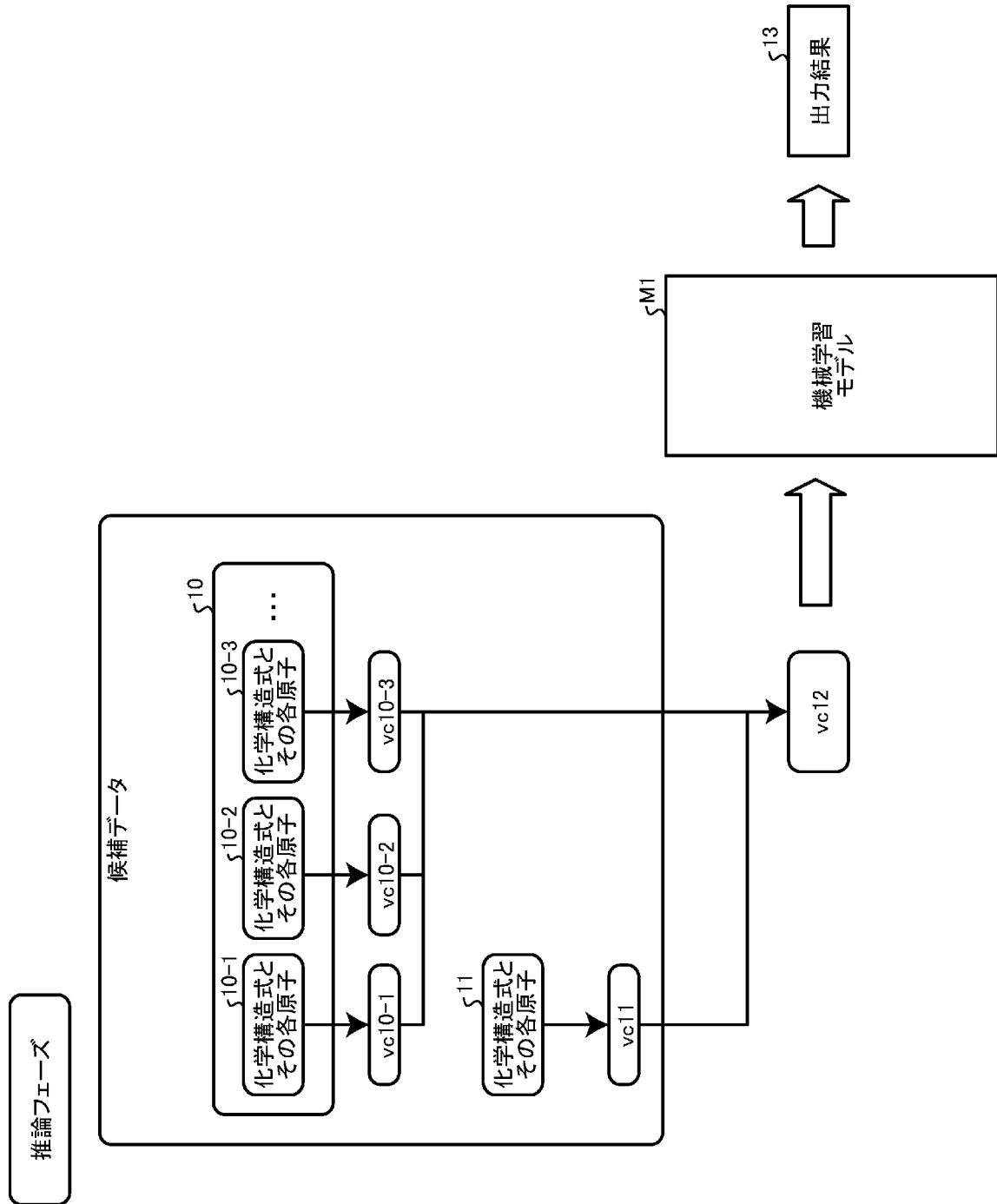
バリン「Val/V」の化学構造式



[図14]



[図15]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2023/015686

A. CLASSIFICATION OF SUBJECT MATTER*G16B 15/00*(2019.01)i; *G16B 40/00*(2019.01)i

FI: G16B15/00; G16B40/00

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)

G06Q10/00-99/00; G16B5/00-99/00

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.
X	JONES, Derek et al. Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. Journal of Chemical Information and Modeling [online]. AMERICAN CHEMICAL SOCIETY. 23 March 2021, [retrieved on 25 May 2023], internet: <URL: https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.0c01306 >, pp. 1583-1592 in particular, pp. 1584-1585, "METHODS", "Data.", "3D-CNN."	1-3, 6-10, 13-17, 20-21
A	entire text, all drawings	4-5, 11-12, 18-19
A	OFER, Dan et al. The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal [online]. ELSEVIER. 25 March 2021, [retrieved on 25 May 2023], internet: <URL: https://www.sciencedirect.com/science/article/pii/S2001037021000945 >, pp. 1750-1758 entire text, all drawings	1-21
A	OZCELIK, Riza et al. Structure-based drug discovery with deep learning. arXiv [online]. 26 December 2022, [retrieved on 25 May 2023], internet: <URL: https://arxiv.org/pdf/2212.13295.pdf >, pp. 1-18 entire text, all drawings	1-21



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

26 May 2023

Date of mailing of the international search report

06 June 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.

A. 発明の属する分野の分類（国際特許分類（IPC）） G16B 15/00(2019.01)i; G16B 40/00(2019.01)i FI: G16B15/00; G16B40/00		
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）） G06Q10/00-99/00; G16B5/00-99/00		
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
X	JONES, Derek et al., Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference, Journal of Chemical Information and Modeling [online], AMERICAN CHEMICAL SOCIETY, 2021.03.23, [検索日 2023.5.25], インターネット: <URL: https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.0c01306 >, p.1583-1592 特に, p.1584-1585 “METHODS” の “Data.” 及び “3D-CNN.”	1-3, 6-10, 13-17, 20-21
A	全文、全図	4-5, 11-12, 18-19
A	OFER, Dan et al., The language of proteins: NLP, machine learning & protein sequences, Computational and Structural Biotechnology Journal [online], ELSEVIER, 2021.03.25, [検索日 2023.5.25], インターネット: <URL: https://www.sciencedirect.com/science/article/pii/S2001037021000945 >, p.1750-1758 全文、全図	1-21
A	OZCELIK, Riza et al., Structure-based drug discovery with deep learning, arXiv [online], 2022.12.26, [検索日 2023.5.25], インターネット: <URL: https://arxiv.org/pdf/2212.13295.pdf >, p.1-18 全文、全図	1-21
☐ 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 26.05.2023		国際調査報告の発送日 06.06.2023
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） 岡北 有平 5R 2569 電話番号 03-3581-1101 内線 3562