(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
2021年12月9日(09.12.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/245926 A1

(51) 国際特許分類:
G16C 20/40 (2019.01)

(21) 国際出願番号： PCT/JP2020/022368

(22) 国際出願日： 2020年6月5日(05.06.2020)

(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).

古賀 裕美 (**KOGA, Hiromi**); 〒8120007 福岡県福岡市博多区東比恵一丁目５番１３号 株式会社富士通九州システムズ内 Fukuoka (JP).

多賀 万里子 (**TAGA, Mariko**); 〒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).

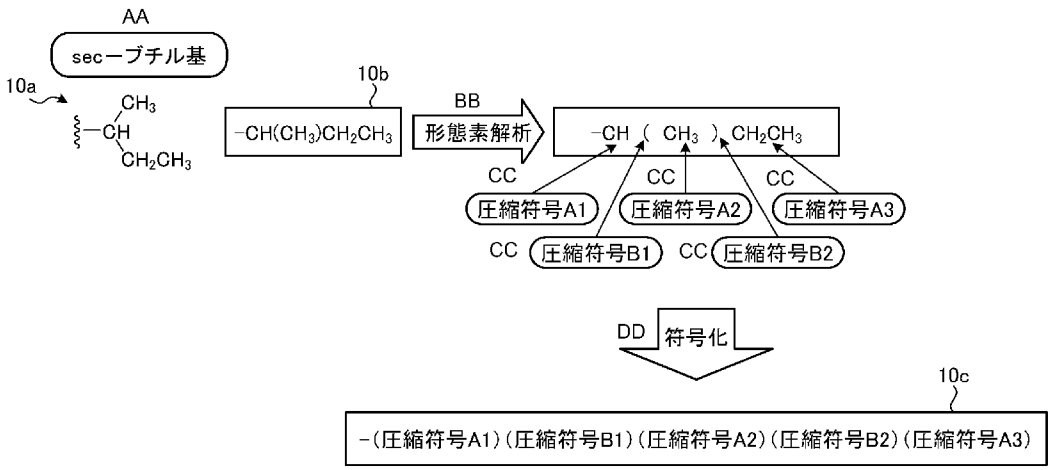
(74) 代理人: 特許業務法人酒井国際特許事務所 (**SAKAI INTERNATIONAL PATENT OFFICE**); 〒1000013 東京都千代田区霞が関３丁目８番１号 虎の門三井ビルディング Tokyo (JP).

(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,

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

(54) 発明の名称： 情報処理プログラム、情報処理方法および情報処理装置

[図1]



AA sec-butyl group
BB Morphological analysis
CC Compression code
DD Encoding
10c -(compression code A1)(compression code B1)(compression code A2)(compression code B2)(compression code A3)

(57) **Abstract:** This information processing device divides a sequence indicating the rational formula of a compound into smallest unit strings of the sequence and branching symbols indicating branchings of the compound. On the basis of a group dictionary that indicates the relation between compression codes and sequences in the rational formula of a compound, the information processing device, by assigning compression codes to the smallest unit strings and assigning compression codes that correspond to the type of branchings to the branch symbols, generates a first coded sequence in

WO 2021/245926 A1

[続葉有]

CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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))

which the sequence is encoded. On the basis of a group primary structure dictionary which indicates the relation between compression codes and group primary structures in the sequence of the rational formula of the compound, the information processing device generates a second coded sequence in which the compression codes included in the first coded sequence are encoded with the group primary structure units.

(57) 要約:情報処理装置は、化合物の示性式を示す配列を、配列の最小単位の文字列、および、化合物の分岐部分を示す分岐記号に分割する。情報処理装置は、化合物の示性式の配列と圧縮符号との関係を示す基辞書を基にして、最小単位の文字列に圧縮符号を割り当て、分岐記号に分岐部分の種別に応じた圧縮符号を割り当てることで、配列を符号化した第1符号配列を生成する。情報処理装置は、化合物の示性式の配列の基一次構造と圧縮符号との関係を示す基一次構造辞書を基にして、第1符号配列に含まれる圧縮符号を、基一次構造の単位で符号化した第2符号配列を生成する。

明　細　書

発明の名称 ：

情報処理プログラム、情報処理方法および情報処理装置

技術分野

[0001]　本発明は、情報処理プログラム等に関する。

背景技術

[0002]　デンプン、セルロース等の糖類や、タンパク質、核酸、天然ゴム等のように、分子量が１０００を超えるような物質が存在し、かかる物質は、高分子化合物とも呼ばれる。

[0003]　ここで、新薬や新素材の開発を行う場合、既に開発されている高分子化合物を検索し、類似度を評価することが求められている。

[0004]　たとえば、化合物の類似度を評価する技術として、次のような従来技術１、２がある。従来技術１では、アミノ酸記号で記述された、タンパク質などの化合物のアミノ酸配列を入力された文字列で検索するシステムである。また、従来技術２では、化合物の官能基や原子タイプの数を数えるなどの記述式に、ベクトルを割り当て、各化合物に予め割り当てたベクトルに対して、類似度を評価する。

先行技術文献

特許文献

[0005] 特許文献1：特開平１１－３０６１８８号公報

特許文献2：国際公開第２００７／１３９０３７号

特許文献3：特開２０１９－２０４３６２号公報

発明の概要

発明が解決しようとする課題

[0006]　しかしながら、上述した従来技術では、高分子化合物の類似度評価の精度が低いという問題がある。

[0007]　１つの側面では、本発明は、高分子化合物の類似度評価の精度の評価を向

上させることができる情報処理プログラム、情報処理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

[0008] 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、化合物の示性式を示す配列を、配列の最小単位の文字列、および、化合物の分岐部分を示す分岐記号に分割する。コンピュータは、化合物の示性式の配列と圧縮符号との関係を示す基辞書を基にして、最小単位の文字列に圧縮符号を割り当て、分岐記号に分岐部分の種別に応じた圧縮符号を割り当てることで、配列を符号化した第１符号配列を生成する。コンピュータは、化合物の示性式の配列の基一次構造と圧縮符号との関係を示す基一次構造辞書を基にして、第１符号配列に含まれる圧縮符号を、基一次構造の単位で符号化した第２符号配列を生成する。

発明の効果

[0009] 高分子化合物の類似度評価の精度を向上することができる。

図面の簡単な説明

[0010] [図1]図１は、本実施例にかかる情報処理装置が化合物を符号化する処理を説明するための図（１）である。

[図2]図２は、本実施例にかかる情報処理装置が化合物を符号化する処理を説明するための図（２）である。

[図3]図３は、本実施例にかかる情報処理装置が化合物を符号化する処理を説明するための図（３）である。

[図4]図４は、本実施例にかかる情報処理装置の構成を示す機能ブロック図である。

[図5]図５は、化学構造式ファイルのデータ構造の一例を示す図である。

[図6]図６は、分岐符号テーブルのデータ構造の一例を示す図である。

[図7]図７は、基辞書のデータ構造の一例を示す図である。

[図8]図８は、基ＨＭＭのデータ構造の一例を示す図である。

[図9]図９は、化学構造式圧縮ファイルのデータ構造の一例を示す図である。

[図10]図１０は、第１転置インデックスのデータ構造の一例を示す図である。

[図11]図１１は、基一次構造辞書のデータ構造の一例を示す図である。

[図12]図１２は、基一次構造ＨＭＭのデータ構造の一例を示す図である。

[図13]図１３は、基一次構造圧縮ファイルのデータ構造の一例を示す図である。

[図14]図１４は、第２転置インデックスのデータ構造の一例を示す図である。

[図15]図１５は、性質管理テーブルのデータ構造の一例を示す図である。

[図16A]図１６Ａは、基ベクトルテーブルのデータ構造の一例を示す図である。

[図16B]図１６Ｂは、基一次構造ベクトルテーブルのデータ構造の一例を示す図である。

[図16C]図１６Ｃは、遷移テーブルのデータ構造の一例を示す図である。

[図17]図１７は、第２符号化部の処理を説明するための図である。

[図18]図１８は、ポアンカレ空間の一例を示す図である。

[図19]図１９は、本実施例にかかる情報処理装置の処理手順を示すフローチャートである。

[図20]図２０は、本実施例にかかる情報処理装置のその他の処理を説明するための図である。

[図21]図２１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。

発明を実施するための形態

[0011] 以下に、本願の開示する情報処理プログラム、情報処理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

[0012] 図１は、本実施例にかかる情報処理装置が化合物を符号化する処理を説明

するための図（１）である。ここでは、化合物を符号化する場合について説明するが、高分子化合物を符号化する場合でも、同様に符号化することができる。図１では、ある化合物に含まれるｓｅｃ－ブチル基は、さらに、メチル基やエチル基に分解でき、それを符号化する場合について説明する。ｓｅｃ－ブチル基の化学構造式１０ａは、「ＣＨ」から、メチル基「CH_3」、エチル基「CH_2CH_3」に分岐している。ｓｅｃ－ブチル基の化学構造式１０ａの示性式１０ｂは、「－ＣＨ（CH_3）CH_2CH_3」のような配列で示される。

[0013] 情報処理装置は、示性式１０ｂに対して形態素解析を実行することで、示性式１０ｂに含まれる配列を、配列の最小単位（意味ある単位）の文字列、化合物の分岐部分を示す記号に分割する。以下の説明では、配列の最小単位の文字列であって意味ある単位の符号を「単位文字列」と表記し、化合物の分岐部分を示す記号を「分岐記号」と表記する。

[0014] 図１に示す例では、情報処理装置は、示性式１０ｂを、「ＣＨ」、「（」、「CH_3」、「）」、「CH_2CH_3」に分割する。「ＣＨ」、「CH_3」、「CH_2CH_3」は、基（または、官能基）の単位文字列である。「（」、「）」は、分岐記号である。

[0015] 情報処理装置は、化合物の示性式の配列の単位文字列と圧縮符号との関係を示す基辞書を基にして、各単位文字列に圧縮符号を割り当てる。たとえば、情報処理装置は、「ＣＨ」、「CH_3」、「CH_2CH_3」に、圧縮符号Ａ１、圧縮符号Ａ２、圧縮符号Ａ３をそれぞれ割り当てる。

[0016] 情報処理装置は、分岐記号と圧縮符号との関係を示す分岐符号テーブルを基にして、各分岐記号に圧縮符号を割り当てる。たとえば、情報処理装置は、「（」、「）」に、圧縮符号Ｂ１、圧縮符号Ｂ２を割り当てる。

[0017] 情報処理装置は、上記処理を実行することで、示性式１０ｂを符号化した圧縮符号配列１０ｃを生成する。情報処理装置は、図示しない化合物の残りの部分に対しても上記処理を実行することで、化合物を符号化する。また、ブチル基やｓｅｃ－ブチル基など、類似した基をまとめて、各基をポアンカ

レ空間にまとめて埋め込むことで、基のベクトルを生成する。

[0018] 情報処理装置は、圧縮符号配列１０ｃを、更に、分岐に着目し、基の一次構造の単位で符号化する。情報処理装置は、基の一次構造を構成する複数の基のベクトル加算することで、化合物の基の一次構造のベクトルを算出する。

[0019] 図２は、本実施例にかかる情報処理装置が化合物を符号化する処理を説明するための図（２）である。図２では、ある化合物に含まれるイソブチル基を符号化する場合について説明する。イソブチル基の化学構造式１１ａは、「ＣＨ」から、「CH_2」、２つの「CH_3」に分岐している。イソブチル基の化学構造式１１ａの示性式１１ｂは、「$-CH_2CH(CH_3)_2$」のような配列で示される。

[0020] 情報処理装置は、示性式１１ｂに対して形態素解析を実行することで、示性式１１ｂに含まれる配列を、単位文字列と、分岐記号に分割する。図２に示す例では、情報処理装置は、示性式１１ｂを、「CH_2」、「ＣＨ」、「（」、「CH_3」、「$)_2$」に分割する。「CH_2」、「ＣＨ」、「CH_3」は、単位文字列である。「（」、「$)_2$」は、分岐記号である。

[0021] 情報処理装置は、化合物の示性式の配列の単位文字列と圧縮符号との関係を示す基辞書を基にして、各単位文字列に圧縮符号を割り当てる。たとえば、情報処理装置は、「CH_2」、「ＣＨ」、「CH_3」に、圧縮符号Ａ４、圧縮符号Ａ１、圧縮符号Ａ２をそれぞれ割り当てる。

[0022] 情報処理装置は、分岐記号と圧縮符号との関係を示す分岐符号テーブルを基にして、各分岐記号に圧縮符号を割り当てる。たとえば、情報処理装置は、「（」、「$)_2$」に、圧縮符号Ｂ１、圧縮符号Ｂ３を割り当てる。

[0023] 情報処理装置は、上記処理を実行することで、示性式１１ｂを符号化した圧縮符号配列１１ｃを生成する。情報処理装置は、図示しない化合物の残りの部分に対しても上記処理を実行することで、化合物を符号化する。また、類似した基をまとめて、各基をポアンカレ空間にまとめて埋め込むことで、基のベクトルを生成する。

[0024] 情報処理装置は、圧縮符号配列１１ｃを、更に、基の一次構造の単位で符号化する。情報処理装置は、基の一次構造を構成する複数の基のベクトル加算することで、化合物の一次構造のベクトルを算出する。

[0025] 図３は、本実施例にかかる情報処理装置が化合物を符号化する処理を説明するための図（３）である。図３では、ある化合物に含まれるｔｅｒｔ－ブチル基を符号化する場合について説明する。ｔｅｒｔ－ブチル基の化学構造式１２ａは、「Ｃ」から、３つの「CH_3」に分岐している。ｔｅｒｔ－ブチル基の化学構造式１２ａの示性式１２ｂは、「－Ｃ（CH_3）$_3$」のような配列で示される。

[0026] 情報処理装置は、示性式１２ｂに対して形態素解析を実行することで、示性式１２ｂに含まれる配列を、単位文字列と、分岐記号に分割する。図３に示す例では、情報処理装置は、示性式１２ｂを、「Ｃ」、「（」、「CH_3」、「）$_3$」に分割する。「CH_2」、「Ｃ」、「CH_3」は、単位文字列である。「（」、「）$_3$」は、分岐記号である。

[0027] 情報処理装置は、化合物の示性式の配列と圧縮符号との関係を示す基辞書を基にして、各単位文字列に圧縮符号を割り当てる。たとえば、情報処理装置は、「Ｃ」、「CH_3」に、圧縮符号Ａ５、圧縮符号Ａ２をそれぞれ割り当てる。

[0028] 情報処理装置は、分岐記号と圧縮符号との関係を示す分岐符号テーブルを基にして、各分岐記号に圧縮符号を割り当てる。たとえば、情報処理装置は、「（」、「）$_3$」に、圧縮符号Ｂ１、圧縮符号Ｂ４を割り当てる。

[0029] 情報処理装置は、上記処理を実行することで、示性式１２ｂを符号化した圧縮符号配列１２ｃを生成する。情報処理装置は、図示しない化合物の残りの部分に対しても上記処理を実行することで、化合物を符号化する。また、類似した基をまとめて、各基をポアンカレ空間にまとめて埋め込むことで、基のベクトルを生成する。

[0030] 情報処理装置は、圧縮符号配列１２ｃを、更に、基の一次構造の単位で符号化する。情報処理装置は、基の一次構造を構成する複数の基のベクトル加

算することで、化合物の一次構造のベクトルを算出する。

[0031] 上記のように、本実施例にかかる情報処理装置は、化合物の示性式を示す配列を、単位文字列および分岐記号に分割し、基辞書を基にして、化合物の示性式を符号化する。また、類似した基をまとめて、各基をポアンカレ空間にまとめて埋め込むことで、基のベクトルを生成する。情報処理装置は、符号化した化合物の示性式を、基の一次構造の単位で符号化することで、更に符号化を行い、一次構造を構成する複数の基のベクトル加算することで、化合物の一次構造のベクトルを算出する。これによって、化合物の基と、基の一次構造のベクトルを精度よく算出することができ、かかる基と基の一次構造のベクトルを用いて高分子化合物の類似度評価を実行することで、類似度評価の精度が向上する。なお、示性式は符号化対象の化学構造式の一例である。ＳＭＩＬＥＳ記法などで表現された化学構造式であっても良い。

[0032] 次に、本実施例にかかる情報処理装置の構成の一例について説明する。図４は、本実施例にかかる情報処理装置の構成を示す機能ブロック図である。図４に示すように、この情報処理装置１００は、通信部１１０と、入力部１２０と、表示部１３０と、記憶部１４０と、制御部１５０とを有する。

[0033] 通信部１１０は、有線又は無線で外部装置等に接続され、外部装置等との間で情報の送受信を行う。たとえば、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。

[0034] 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。

[0035] 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。

[0036] 記憶部１４０は、化学構造式ファイル１４０ａと、分岐符号テーブル１４０ｂと、基辞書１４０ｃと、基ＨＭＭ１４０ｄと、化学構造式圧縮ファイル

１４０ｅと、第１転置インデックス１４０ｆとを有する。記憶部１４０は、基一次構造辞書１４０ｇと、基一次構造ＨＭＭ１４０ｈと、基一次構造圧縮ファイル１４０ｉと、第２転置インデックス１４０ｊと、性質管理テーブル１４０ｋと、遷移テーブル１４０ｎとを有する。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。

[0037] 化学構造式ファイル１４０ａは、高分子化合物の示性式（化学構造式）の情報を含むファイルである。図５は、化学構造式ファイルのデータ構造の一例を示す図である。図５に示すように、この化学構造式ファイル１４０ａは、化合物識別情報と、示性式とを対応付ける。化合物識別情報は、高分子化合物を一意に識別する情報である。示性式は、高分子化合物の示性式（化学構造式）を示す配列の情報である。たとえば、高分子化合物は、デンプン、グリコーゲン、セルロース、アミロペクチン等に対応するが、これに限定されるものではなく、他の高分子化合物であってもよい。

[0038] 分岐符号テーブル１４０ｂは、分岐記号に割り当てる符号を定義するテーブルである。図６は、分岐符号テーブルのデータ構造の一例を示す図である。図６に示すように、この分岐符号テーブル１４０ｂは、分岐記号と、圧縮符号とを対応付ける。分岐記号は、高分子化合物の分岐部分を示す記号である。圧縮符号は、分岐記号に割り当てられる圧縮符号である。圧縮符号の「ｈ」は、１６進数であることを示す。

[0039] 基辞書１４０ｃは、化合物の示性式の配列と圧縮符号との関係を示す辞書情報である。図７は、基辞書のデータ構造の一例を示す図である。図７に示すように、この基辞書１４０ｃは、圧縮符号と、名称と、グループと、示性式とを対応付ける。圧縮符号は、該当する基（または官能基）の示性式に割り当てられる圧縮符号を示す。名称は、該当する名称の一例である。グループは、該当する基が属するグループを示す。名称例は、該当する基の名称の一例である。示性式は、該当する基の示性式となる配列を示す。

[0040] また、図７では図示を省略するが、基辞書１４０ｃは、図１～図３等で説明した、単位文字列と圧縮符号との関係を定義した情報を保持する。

[0041] 基ＨＭＭ（Hidden Markov Model）１４０ｄは、高分子化合物の配列において、ある基と、共起する可能性の高い他の基とを対応付ける情報である。以下の説明では適宜、ある基を、「第１基」と表記し、この第１基と共起する他の基を「第２基」と表記する。

[0042] 図８は、基ＨＭＭのデータ構造の一例を示す図である。図８に示すように、この基ＨＭＭ１４０ｄは、圧縮符号と、名称と、グループと、基の共起率とを対応付ける。圧縮符号は、第１基の示性式に割り当てられる圧縮符号を示す。名称は、該当する第１基の名称の一例である。グループは、該当する第１基が属するグループを示す。

[0043] 基の共起率では、第１基と共起する第２基について、第２基の圧縮符号と、共起率と示す。たとえば、１行目について説明すると、高分子化合物において、第１基（圧縮符号：８０００ｈ）と共起する第２基は圧縮符号「８０２８ｈ」の基と、圧縮符号「８１３２ｈ」の基と、圧縮符号「８０Ｆ５ｈ」の基等である。また、第１基「圧縮符号：８０００ｈ」と第２基「圧縮符号：８０２８ｈ」との共起率は「７８％」である。第１基「圧縮符号：８０００ｈ」と第２基「圧縮符号：８１３２ｈ」との共起率は「６３％」である。第１基「圧縮符号：８０００ｈ」と第２基「圧縮符号：８０Ｆ５ｈ」との共起率は「３３％」である。

[0044] 化学構造式圧縮ファイル１４０ｅは、符号化した高分子化合物の示性式（化学構造式）の情報を含むファイルである。図９は、化学構造式圧縮ファイルのデータ構造の一例を示す図である。図９に示すように、この化学構造式圧縮ファイル１４０ｅは、化合物識別情報と、圧縮符号配列とを対応付ける。化合物識別情報は、高分子化合物を一意に識別する情報である。圧縮符号配列は、示性式を単位文字列の単位で符号化した情報である。

[0045] 第１転置インデックス１４０ｆは、化学構造式圧縮ファイル１４０ｅに含まれる単位文字列または分岐記号の圧縮符号と、この圧縮符号の出現位置と

の関係を示すインデックス情報である。図１０は、第１転置インデックスのデータ構造の一例を示す図である。図１０に示すように、第１転置インデックス１４５は、横軸にオフセットをとり、縦軸に圧縮符号を取る。オフセットは、化学構造式圧縮ファイル１４０ｅの先頭の圧縮符号から、該当する圧縮符号までの出現位置を示すものである。先頭の圧縮符号のオフセットを「０」とする。

[0046] 第１転置インデックス１４０ｆは、図９に示した化合物識別情報毎の圧縮符号配列に対して、圧縮符号とオフセットとの関係を定義してもよい。たとえば、化合物識別情報「Ｃ１０１」の圧縮符号配列の転置インデックス、化合物識別情報「Ｃ１０２」の圧縮符号配列の転置インデックスをそれぞれ定義してもよい。

[0047] 基一次構造辞書１４０ｇは、高分子化合物の示性式の圧縮符号配列（連続する複数の圧縮符号）と、一つの基一次構造の圧縮符号との関係を示す辞書情報である。基一次構造は、複数の基が組み合わさった構造を示す。

[0048] 図１１は、基一次構造辞書のデータ構造の一例を示す図である。図１１に示すように、この基一次構造辞書１４０ｇは、圧縮符号と、グループと、名称と、圧縮符号配列とを対応付ける。圧縮符号は、基一次構造に対応する圧縮符号である。グループは、該当する基一次構造が属するグループを示す。名称は、基一次構造の名称の一例である。圧縮符号配列は、基一次構造に含まれる複数の基に対応する圧縮符号の配列であり、基の圧縮符号は、基辞書１４０ｃに定義された基の圧縮符号に対応する。

[0049] 基一次構造ＨＭＭ１４０ｈは、高分子化合物の配列において、ある基一次構造と、共起する可能性の高い他の基一次構造とを対応付ける情報である。以下の説明では適宜、ある基一次構造を「第１基一次構造」と表記し、この第１基一次構造と共起する他の基一次構造を「第２基一次構造」と表記する。

[0050] 図１２は、基一次構造ＨＭＭのデータ構造の一例を示す図である。図１２に示すように、この基一次構造ＨＭＭ１４０ｈは、圧縮符号と、グループと

、名称と、基一次構造の共起率とを対応付ける。圧縮符号は、第１基一次構造に割り当てられる圧縮符号を示す。グループは、該当する第１基一次構造の属するグループを示す。名称は、該当する第基一次構造の名称の一例である。

[0051] 基一次構造の共起率は、第１基一次構造と共起する第２基一次構造について、第２基一次構造の圧縮符号と、共起率とを示す。たとえば、１行目について説明すると、高分子化合物において、第１基一次構造（圧縮符号：Ｆ０００００ｈ）と共起する第２基一次構造の各圧縮符号は、「Ｆ０００１１ｈ」、「Ｆ００１０１ｈ」、「Ｆ００２２５ｈ」である。

[0052] また、第１基一次構造「圧縮符号：Ｆ０００００ｈ」と第２基「圧縮符号：Ｆ０００１１ｈ」との共起率は「７８％」である。第１基一次構造「圧縮符号：Ｆ０００００ｈ」と第２基「圧縮符号：Ｆ００１０１ｈ」との共起率は「６３％」である。第１基一次構造「圧縮符号：Ｆ０００００ｈ」と第２基「圧縮符号：Ｆ００２２５ｈ」との共起率は「５１％」である。

[0053] 基一次構造圧縮ファイル１４０ｉは、基一次構造の単位で符号化した高分子化合物の示性式の情報を含むファイルである。図１３は、基一次構造圧縮ファイルのデータ構造の一例を示す図である。図１３に示すように、この基一次構造圧縮ファイル１４０ｉは、化学物識別情報と、基一次構造の圧縮符号配列とを対応付ける。化合物識別情報は、高分子化合物を一意に識別する情報である。圧縮符号配列は、示性式を基一次構造の単位で符号化した情報である。

[0054] 第２転置インデックス１４０ｊは、基一次構造圧縮ファイル１４０ｉに含まれる基一次構造の圧縮符号と、この圧縮符号の出現位置との関係を示すインデックス情報である。図１４は、第２転置インデックスのデータ構造の一例を示す図である。図１４に示すように、第２転置インデックス１４０ｊは、横軸にオフセットをとり、縦軸に圧縮符号を取る。オフセットは、基一次構造圧縮ファイル１４０ｉの先頭の圧縮符号（基一次構造の単位の圧縮符号）から、該当する圧縮符号までの出現位置を示すものである。先頭の圧縮符

号のオフセットを「０」とする。

[0055] 第２転置インデックス１４０ｊは、図１３に示した化合物識別情報毎の圧縮符号配列に対して、圧縮符号とオフセットとの関係を定義してもよい。たとえば、化合物識別情報「Ｃ１０１」の圧縮符号配列の転置インデックス、化合物識別情報「Ｃ１０２」の圧縮符号配列の転置インデックスをそれぞれ定義してもよい。

[0056] 性質管理テーブル１４０ｋは、性質が類似する基を定義するテーブルである。図１５は、性質管理テーブルのデータ構造の一例を示す図である。図１５に示すように、この性質管理テーブル１４０ｋは、性質番号と、圧縮符号と、名称とを対応付ける。性質番号は、基一次構造の性質を識別する番号であり、類似の性質をもつ基の各圧縮符号には、同一の性質番号が割り振られる。圧縮符号は、基の圧縮符号である。名称は、基の名称である。

[0057] 図１５に示す例では、圧縮符号「８００４ｈ」、圧縮符号「８００５ｈ」、圧縮符号「８００６ｈ」、圧縮符号「８００７ｈ」に同一の性質符号「Ｉ１０２」が割り振られている。このため、圧縮符号「８００４ｈ」、圧縮符号「８００５ｈ」、圧縮符号「８００６ｈ」、圧縮符号「８００７ｈ」に対応する各基は、類似の性質を持つことを意味する。

[0058] 図１５では説明を省略するが、この性質管理テーブル１４０ｋは、性質番号と、類似の性質をもつ基の圧縮符号とを対応付けた情報を保持していてもよい。

[0059] 基ベクトルテーブル１４０Ｉは、高分子化合物に含まれる各基に対して割り当てられる圧縮符号のベクトルを保持するテーブルである。図１６Ａは、基ベクトルテーブルのデータ構造の一例を示す図である。図１６Ａに示すように、基ベクトルテーブル１４０Ｉは、基の圧縮符号と、ベクトルとを対応付ける。なお、基ベクトルテーブル１４０Ｉは、分岐符号テーブル１４０ｂに定義された分岐記号に対応するベクトルを保持してもよい。

[0060] 基一次構造ベクトルテーブル１４０ｍは、高分子化合物に含まれる各基一次構造に対して割り当てられる圧縮符号のベクトルを保持するテーブルであ

る。図１６Ｂは、基一次構造ベクトルテーブルのデータ構造の一例を示す図である。図１６Ｂに示すように、基一次構造ベクトルテーブル１４０ｍは、基一次構造の圧縮符号と、ベクトルとを対応付ける。

[0061] 遷移テーブル１４０ｎは、高分子化合物のベクトルの情報を保持するテーブルである。図１６Ｃは、遷移テーブルのデータ構造の一例を示す図である。図１６Ｃに示すように、遷移テーブル１４０ｎは、化合物識別情報と、複数のベクトルとを有する。化合物識別情報は、高分子化合物を一意に識別する情報である。複数のベクトルは、高分子化合物に含まれる各基一次構造のベクトルを示す。たとえば、ベクトル（ｎ）は、高分子化合物の先頭からｎ番目に位置する基一次構造のベクトルである。

[0062] 図４の説明に戻る。制御部１５０は、取得部１５１と、第１符号化部１５２と、第２符号化部１５３と、ベクトル算出部１５４と、類似度評価部１５５とを有する。制御部１５０は、例えば、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。

[0063] 取得部１５１は、ネットワークを介して、外部装置等から各種の情報を取得する処理部である。たとえば、取得部１５１は、化学構造式ファイル１４０ａを取得し、取得した化学構造式ファイル１４０ａを記憶部１４０に格納する。取得部１５１は、基辞書１４０ｃ、基ＨＭＭ１４０ｄ、基一次構造辞書１４０ｇ、基一次構造ＨＭＭ１４０ｈを取得して、記憶部１４０に格納してもよい。

[0064] 第１符号化部１５２は、化学構造式ファイル１４０ａから、高分子化合物の示性式を取得し、取得した示性式を符号化する処理部である。以下において、第１符号化部１５２の処理の一例について説明する。

[0065] 第１符号化部１５２は、高分子化合物の示性式に対して形態素解析を実行することで、示性式に含まれる配列を単位文字列、または、分岐記号に分割

する。第１符号化部１５２が、形態素解析を実行して、単位文字列、または、分岐記号に分割する処理は、図１、図２、図３で説明した処理に対応する。

[0066] 第１符号化部１５２は、分割候補となる単位文字列が複数通り存在する場合には、基ＨＭＭ１４０ｄを基にして、共起率の高い方の単位文字列を選択して、分割してもよい。

[0067] 第１符号化部１５２は、高分子化合物の示性式を単位文字列、または、分岐記号に分割した後に、分岐符号テーブル１４０ｂおよび基辞書１４０ｃを用いて、示性式を符号化することで、圧縮符号配列を生成する。第１符号化部１５２が生成する圧縮符号配列は、「第１符号配列」に対応する。第１符号化部１５２は、化合物識別情報と、圧縮符号配列とを対応付けて、化学構造式圧縮ファイル１４０ｅに登録する。

[0068] 第１符号化部１５２は、高分子化合物の示性式の単位文字列、または、分岐記号に圧縮符号を割り当てる場合に、圧縮符号の種別と、オフセットとの関係を第１転置インデックス１４０ｆに登録する。たとえば、第１符号化部１５２は、高分子化合物の示性式の圧縮符号配列（第１符号配列）において、圧縮符号「８０００ｈ」のオフセットが「１」である場合には、第１転置インデックス１４０ｆのオフセット「１」の列と、圧縮符号「８０００ｈ」の行とが交差する部分に「１」を設定する。

[0069] 第１符号化部１５２は、化学構造式ファイル１４０ａから、未選択の化合物識別情報に対応する示性式の情報を取得し、上記処理を繰り返し実行することで、化学構造式圧縮ファイル１４０ｅを生成する。

[0070] 第２符号化部１５３は、化学構造式圧縮ファイル１４０ｅから、高分子化合物の圧縮符号配列（第１符号配列）を取得し、取得した圧縮符号配列を、基一次構造の単位で符号化する処理部である。高分子化合物の圧縮符号配列と、基一次構造の圧縮符号との関係は、基一次構造辞書１４０ｇに定義されている。

[0071] 図１７は、第２符号化部の処理を説明するための図である。図１７におい

て、圧縮符号配列１５は、化学構造式圧縮ファイル１４０ｅに登録された化合物識別情報「Ｃ１０１」に対応する圧縮符号配列１５ａとする。圧縮符号配列１５ａに含まれる各圧縮符号は、基辞書１４０ｃに定義された単位符号の圧縮符号、または、分岐記号の圧縮符号である。

[0072] たとえば、基一次構造辞書１４０ｇにおいて、圧縮符号配列１５ａの各圧縮符号Ｘ１、Ｘ２、・・・、Ｘ１９と、基一次構造の圧縮符号「Ｆ００１１２ｈ」とが対応づけられているものとする。この場合、第２符号化部１５３は、各圧縮符号Ｘ１、Ｘ２、・・・、Ｘ１９に、基一次構造の圧縮符号「Ｆ００１１２ｈ」を割り当てる。

[0073] 基一次構造辞書１４０ｇにおいて、圧縮符号配列１５ａの各圧縮符号Ｙ１、Ｙ２、・・・、Ｘ３８と、基一次構造の圧縮符号「Ｆ００３３２ｈ」とが対応づけられているものとする。この場合、第２符号化部１５３は、各圧縮符号Ｙ１、Ｙ２、・・・、Ｙ３８に、基一次構造の圧縮符号「Ｆ００３３２ｈ」を割り当てる。

[0074] 基一次構造辞書１４０ｇにおいて、圧縮符号配列１５ａの各圧縮符号Ｚ１、Ｚ２、・・・、Ｚ２０と、基一次構造の圧縮符号「Ｆ００１１０ｈ」とが対応づけられているものとする。この場合、第２符号化部１５３は、各圧縮符号Ｚ１、Ｚ２、・・・、Ｚ２０に、基一次構造の圧縮符号「Ｆ００１１０ｈ」を割り当てる。

[0075] 第２符号化部１５３は、基一次構造辞書１４０ｇを基にして、上記処理を繰り返し実行することで、圧縮符号配列１５ａを、基一次構造の単位で符号化した圧縮符号配列１５ｂを生成する。圧縮符号配列１５ｂは「第２符号配列」に対応する。第２符号化部１５３は、化合物識別情報と、圧縮符号配列とを対応付けて、基一次構造圧縮ファイル１４０ｉに登録する。

[0076] 第２符号化部１５３は、圧縮符号配列１５ａと、基一次構造辞書１４０ｇとを比較して、符号化していく過程において、基一次構造の単位の複数の圧縮符号が、符号化の候補となる場合には、基一次構造ＨＭＭ１４０ｈを基にして、共起率の高い方の圧縮符号を選択して、割り当ててもよい。

[0077] 第２符号化部１５３は、基一次構造の単位で符号化する場合に、圧縮符号の種別と、オフセットとの関係を第２転置インデックス１４０ｊに登録する。たとえば、高分子化合物の示性式の圧縮符号配列（第２符号配列）において、圧縮符号「Ｆ００１１２ｈ」のオフセットが「０」となる。この場合、第２符号化部１５３は、第１転置インデックス１４０ｆのオフセット「０」の列と、圧縮符号「Ｆ００１１２ｈ」の行とが交差する部分に「１」を設定する。

[0078] 第２符号化部１５３は、化学構造式圧縮ファイル１４０ｅから、未選択の化合物識別情報に対応する示性式の圧縮符号配列を取得し、上記処理を繰り返し実行することで、基一次構造圧縮ファイル１４０ｉを生成する。

[0079] また、第２符号化部１５３は、基一次構造の圧縮符号と、この基一次構造の圧縮符号に対応する複数の基の圧縮符号、分岐記号の圧縮符号とを関係を、ベクトル算出部１５４に出力する。

[0080] 図４の説明に戻る。ベクトル算出部１５４は、基一次構造圧縮ファイル１４０ｉから、高分子化合物の圧縮符号配列（第２符号配列）を取得し、高分子化合物のベクトルを算出する処理部である。ベクトル算出部１５４は、基一次構造圧縮ファイル１４０ｉから、基一次構造の圧縮符号を取得し、取得した圧縮符号のベクトルを、次の処理を実行することで算出する。ベクトル算出部１５４は、上記の第２符号化部１５３から、基一次構造の圧縮符号と、複数の基の圧縮符号および分岐記号の圧縮符号との関係を取得しておく。

[0081] ベクトル算出部１５４は、基毎の圧縮符号を、ポアンカレ空間等のベクトル空間に埋め込むことで、圧縮符号に対応するベクトルをそれぞれ算出する。ベクトル算出部１５４は、分岐記号の圧縮符号についても、ポアンカレ空間に埋め込むことで、ベクトルを算出してもよいし、予めベクトルを割り当てておいてもよい。ベクトル算出部１５４は、基の圧縮符号と、ベクトルとを対応付けて、基ベクトルテーブル１４０ｌに登録する。

[0082] ベクトル算出部１５４は、基一次構造に含まれる各基の圧縮符号、分岐記号の圧縮符号に割り当てたベクトルを加算することで、基一次構造のベクト

ルを算出する。ベクトル算出部１５４は、基一次構造の圧縮符号と、ベクトルとを対応付けて、基一次構造ベクトルテーブル１４０ｍに登録する。

[0083] 高分子化合物には、複数の基一次構造が含まれている。ベクトル算出部１５４は、化合物識別情報と、基一次構造毎のベクトルとを対応付けて、遷移テーブル１４０ｎに登録する。ベクトル算出部１５４は、基一次構造が出現する順番に、基一次構造のベクトルを、遷移テーブル１４０ｎに登録する。

[0084] ここで、ベクトル算出部１５４が、基の圧縮符号を、ポアンカレ空間に埋め込んで、ベクトルを算出する処理は、ポアンカレエンベッディング（Poincare Embeddings）と呼ばれる技術である。ポアンカレエンベッディングは、たとえば、非特許文献「Valentin Khrulkov1 et al.「Hyperbolic Image Embeddings」Cornell University,2019 April 3」等に記載された技術を用いればよい。

[0085] ポアンカレエンベディングでは、ポアンカレ空間に埋め込まれた位置に応じて、ベクトルが割り当てられるものであり、また、類似する情報ほど、近い位置に埋め込まれるという特徴がある。ベクトル算出部１５４は、性質管理テーブル１４０ｋを基にして、類似の性質をもつ各基の圧縮符号を特定する。

[0086] 図１８は、ポアンカレ空間の一例を示す図である。図１５で説明したように、圧縮符号「８００４ｈ」、圧縮符号「８００５ｈ」、圧縮符号「８００６ｈ」、圧縮符号「８００７ｈ」に対応する基は、類似の性質を持つ。このため、ベクトル算出部１５４は、圧縮符号「８００４ｈ」、圧縮符号「８００５ｈ」、圧縮符号「８００６ｈ」、圧縮符号「８００７ｈ」を、ポアンカレ空間Ｐ上の相互に近い位置に埋め込み、位置に応じたベクトルを付与する。

[0087] ベクトル算出部１５４は、高分子化合物の圧縮符号配列に含まれる各基一次構造の圧縮符号のベクトルを、先頭から順に算出し、遷移テーブル１４０ｎに順に登録する。ベクトル算出部１５４は、基一次構造圧縮ファイル１４０ｉから、未選択の化合物識別情報に対応する示性式の圧縮符号配列を取得

し、上記処理を繰り返し実行することで、遷移テーブル１４０ｎを生成する。

[0088] 類似度評価部１５５は、遷移テーブル１４０ｎに登録された各化合物識別情報に対応するベクトルを比較することで、高分子化合物の類似度を評価する処理部である。たとえば、類似度評価部１５５は、各高分子化合物のベクトルの距離を算出し、距離が閾値未満となる高分子化合物の組を、相互に類似する高分子化合物として特定する。

[0089] たとえば、類似度評価部１５５は、遷移テーブル１４０ｎに登録された高分子化合物に対応する各基一次構造のベクトルを集積したベクトルを、高分子化合物のベクトルし、各高分子化合物のベクトルの距離を算出してもよい。類似度評価部１５５は、評価結果を表示部１３０に出力して表示させてもよいし、外部装置等に通知してもよい。

[0090] 続いて、本実施例にかかる情報処理装置１００の処理手順の一例について説明する。図１９は、本実施例にかかる情報処理装置の処理手順を示すフローチャートである。図１９に示すように、情報処理装置１００の第１符号化部１５２は、化学構造式ファイル１４０ａから、高分子化合物の示性式の配列を取得する（ステップＳ１０１）。

[0091] 第１符号化部１５２は、形態素解析を実行して、示性式の配列を、単位文字列、または、分岐記号に分割する（ステップＳ１０２）。第１符号化部１５２は、分岐符号テーブル１４０ｂ、基辞書１４０ｃ、基ＨＭＭ１４０ｄを基にして、単位文字列、分岐記号に圧縮符号を割り当て、第１転置インデックス１４０ｆ、化学構造式圧縮ファイル１４０ｅを生成する。また、情報処理装置１００のベクトル算出部１５４は、ポアンカレ空間に埋め込むことで基のベクトルを生成する（ステップＳ１０３）。

[0092] 情報処理装置１００の第２符号化部１５３は、化学構造式圧縮ファイル１４０ｅから、高分子化合物の圧縮符号配列を取得する（ステップＳ１０４）。第２符号化部１５３は、基一次構造辞書１４０ｇ、基一次構造ＨＭＭ１４０ｈを基にして、基一次構造単位で、圧縮符号化配列を生成し、第２転置イ

ンデックス１４０ｊ、基一次構造圧縮ファイル１４０ｉを生成する（ステップＳ１０５）。

[0093] 情報処理装置１００のベクトル算出部１５４は、基一次構造圧縮ファイル１４０ｉに格納された基一次構造単位を構成する基のベクトルを加算することで、基一次構造のベクトルを算出し、遷移テーブル１４０ｎを生成する（ステップＳ１０６）。情報処理装置１００の類似度評価部１５５は、遷移テーブル１４０ｎを基にして、類似度評価を実行する（ステップＳ１０７）。

[0094] 次に、本実施例にかかる情報処理装置１００の効果について説明する。情報処理装置１００は、高分子化合物の示性式を、単位文字列、分岐記号に分割して圧縮符号を割り当てることで、圧縮符号配列を生成し、さらに、基一次構造毎の各圧縮符号に変換する。情報処理装置は、基毎の圧縮符号をもとに、類似の性質となる基をまとめてポアンカレ空間に埋め込むことで、化合物の性質を適切に表現するための基のベクトルを生成する。かかる情報処理装置１００の処理によって生成したベクトルを用いて、類似度の評価を実行することで、類似度評価の精度を向上させることができる。

[0095] 情報処理装置１００は、基一次構造を構成する圧縮符号を用いて、各基のベクトルを加算し基の一次構造のベクトルを算出し、遷移テーブルを生成する。このため、性質の類似する各基一次構造に、高精度なベクトルを割り当てることができる。

[0096] 情報処理装置１００は、遷移テーブル１４０ｎに格納された各高分子化合物に対応する基一次構造のベクトルを比較することで、類似の特性を有する高分子化合物を精度よく評価することができる。

[0097] ところで、上述した本実施例にかかる情報処理装置１００の処理は一例であり、その他の処理を実行してもよい。図１～図３で説明したように、情報処理装置１００は、高分子化合物の示性式を符号化する場合に、基に含まれる単位文字列、分岐記号の粒度で、圧縮符号を割り当てていたが、これに限定されるものではない。

[0098] 情報処理装置１００は、基を最小単位とみなし、分岐部分に位置する基に

対して、分岐の基固有の圧縮符号を割り当てることで、符号化を行ってもよい。図２０は、本実施例にかかる情報処理装置のその他の処理を説明するための図である。図２０では、高分子化合物の一例として、グリコーゲンを用いてする。

[0099] 図２０に示す高分子化合物３０は、複数の基が接続されて構成される。高分子化合物３０の領域３０ａに着目して説明を行う。領域３０ａには、複数の基３１と、分岐部分の基３２とが含まれる。情報処理装置１００の第１符号化部１５２は、基辞書１４０ｃを基にして、複数の基３１に、圧縮符号を割り当てる。また、第１符号化部１５２は、分岐部分の基３２に対して、分岐部分の固有の圧縮符号を割り当てる。

[0100] 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２１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。

[0101] 図２１に示すように、コンピュータ２００は、各種演算処理を実行するＣＰＵ２０１と、ユーザからのデータの入力を受け付ける入力装置２０２と、ディスプレイ２０３とを有する。また、コンピュータ２００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２０４と、インタフェース装置２０５とを有する。また、コンピュータ２００は、各種情報を一時記憶するＲＡＭ２０６と、ハードディスク装置２０７とを有する。そして、各装置２０１～２０７は、バス２０８に接続される。

[0102] ハードディスク装置２０７は、取得プログラム２０７ａ、第１符号化プログラム２０７ｂ、第２符号化プログラム２０７ｃ、ベクトル算出プログラム２０７ｄ、類似度評価プログラム２０７ｅを有する。また、ＣＰＵ２０１は、各プログラム２０７ａ～２０７ｅを読み出してＲＡＭ２０６に展開する。

[0103] 取得プログラム２０７ａは、取得プロセス２０６ａとして機能する。第１符号化プログラム２０７ｂは、第１符号化プロセス２０６ｂとして機能する

。第２符号化プログラム２０７ｃは、第２符号化プロセス２０６ｃとして機能する。ベクトル算出プログラム２０７ｄは、ベクトル算出プロセス２０６ｄとして機能する。類似度評価プログラム２０７ｅは、類似度評価プロセス２０６ｅとして機能する。

[0104] 取得プロセス２０６ａの処理は、取得部１５１の処理に対応する。第１符号化プロセス２０６ｂの処理は、第１符号化部１５２の処理に対応する。第２符号化プロセス２０６ｃの処理は、第２符号化部１５３の処理に対応する。ベクトル算出プロセス２０６ｄの処理は、ベクトル算出部１５４の処理に対応する。類似度評価プロセス２０６ｅの処理は、類似度評価部１５５の処理に対応する。

[0105] なお、各プログラム２０７ａ～２０７ｅ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２００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２００が各プログラム２０７ａ～２０７ｅを読み出して実行するようにしてもよい。

符号の説明

[0106]
１００　情報処理装置
１１０　通信部
１２０　入力部
１３０　表示部
１４０　記憶部
１４０ａ　化学構造式ファイル
１４０ｂ　分岐符号テーブル
１４０ｃ　基辞書
１４０ｄ　基ＨＭＭ
１４０ｅ　化学構造式圧縮ファイル
１４０ｆ　第１転置インデックス

１４０ｇ　基一次構造辞書

１４０ｈ　基一次構造ＨＭＭ

１４０ｉ　基一次構造圧縮ファイル

１４０ｊ　第２転置インデックス

１４０ｋ　性質管理テーブル

１４０ｌ　基ベクトルテーブル

１４０ｍ　基一次構造ベクトルテーブル

１４０ｎ　遷移テーブル

１５０　制御部

１５１　取得部

１５２　第１符号化部

１５３　第２符号化部

１５４　ベクトル算出部

１５５　類似度評価部

請求の範囲

[請求項1] コンピュータに、

化合物の示性式を示す配列を、前記配列の最小単位の文字列、および、前記化合物の分岐部分を示す分岐記号に分割し、

化合物の示性式の配列と圧縮符号との関係を示す基辞書を基にして、前記最小単位の文字列に圧縮符号を割り当て、前記分岐記号に前記分岐部分の種別に応じた圧縮符号を割り当てることで、前記配列を符号化した第１符号配列を生成し、

化合物の示性式の配列の基一次構造と圧縮符号との関係を示す基一次構造辞書を基にして、前記第１符号配列に含まれる圧縮符号を、前記基一次構造の単位で符号化した第２符号配列を生成する

処理を実行させることを特徴とする情報処理プログラム。

[請求項2] 前記第１符号配列に含まれる前記基毎の圧縮符号を、ポアンカレ空間に埋め込むことで、基のベクトルを生成し、前記基一次構造を構成する基のベクトルを加算することで、前記基一次構造にベクトルを割り当てる処理を更に実行することを特徴とする請求項１に記載の情報処理プログラム。

[請求項3] 前記ベクトルを割り当てる処理は、類似の性質となる基の圧縮符号を、ポアンカレ空間の類似の位置に埋め込み、前記ポアンカレ空間の位置に応じたベクトルを埋め込んだ圧縮符号に割り当てることを特徴とする請求項２に記載の情報処理プログラム。

[請求項4] 前記ベクトルを割り当てる処理は、前記第２符号配列に含まれる複数の基一次構造のベクトルを順に遷移テーブルに登録する処理を更に実行することを特徴とする請求項２または３に記載の情報処理プログラム。

[請求項5] 前記遷移テーブルに登録されたベクトルを基にして、各化合物の類似度を評価する処理を更に実行することを特徴とする請求項４に記載の情報処理プログラム。

[請求項6] コンピュータが実行する情報処理方法であって、

化合物の示性式を示す配列を、前記配列の最小単位の文字列、および、前記化合物の分岐部分を示す分岐記号に分割し、

化合物の示性式の配列と圧縮符号との関係を示す基辞書を基にして、前記最小単位の文字列に圧縮符号を割り当て、前記分岐記号に前記分岐部分の種別に応じた圧縮符号を割り当てることで、前記配列を符号化した第１符号配列を生成し、

化合物の示性式の配列の基一次構造と圧縮符号との関係を示す基一次構造辞書を基にして、前記第１符号配列に含まれる圧縮符号を、前記基一次構造の単位で符号化した第２符号配列を生成する

処理を実行することを特徴とする情報処理方法。

[請求項7] 前記第１符号配列に含まれる前記基毎の圧縮符号を、ポアンカレ空間に埋め込むことで、基のベクトルを生成し、前記基一次構造を構成する基のベクトルを加算することで、前記基一次構造にベクトルを割り当てる

処理を実行させることを特徴とする請求項６に記載の情報処理方法。

[請求項8] 前記ベクトルを割り当てる処理は、類似の性質となる基の圧縮符号を、ポアンカレ空間の類似の位置に埋め込み、前記ポアンカレ空間の位置に応じたベクトルを埋め込んだ圧縮符号に割り当てることを特徴とする請求項７に記載の情報処理方法。

[請求項9] 前記ベクトルを割り当てる処理は、前記第２符号配列に含まれる複数の基一次構造のベクトルを順に遷移テーブルに登録する処理を更に実行することを特徴とする請求項７または８に記載の情報処理方法。

[請求項10] 前記遷移テーブルに登録されたベクトルを基にして、各化合物の類似度を評価する処理を更に実行することを特徴とする請求項９に記載の情報処理方法。

[請求項11] 化合物の示性式を示す配列を、前記配列の最小単位の文字列、およ

び、前記化合物の分岐部分を示す分岐記号に分割し、化合物の示性式の配列と圧縮符号との関係を示す基辞書を基にして、前記最小単位の文字列に圧縮符号を割り当て、前記分岐記号に前記分岐部分の種別に応じた圧縮符号を割り当てることで、前記配列を符号化した第１符号配列を生成する第１符号化部と、

化合物の示性式の配列の基一次構造と圧縮符号との関係を示す基一次構造辞書を基にして、前記第１符号配列に含まれる圧縮符号を、前記基一次構造の単位で符号化した第２符号配列を生成する第２符号化部と、

を有することを特徴とする情報処理装置。

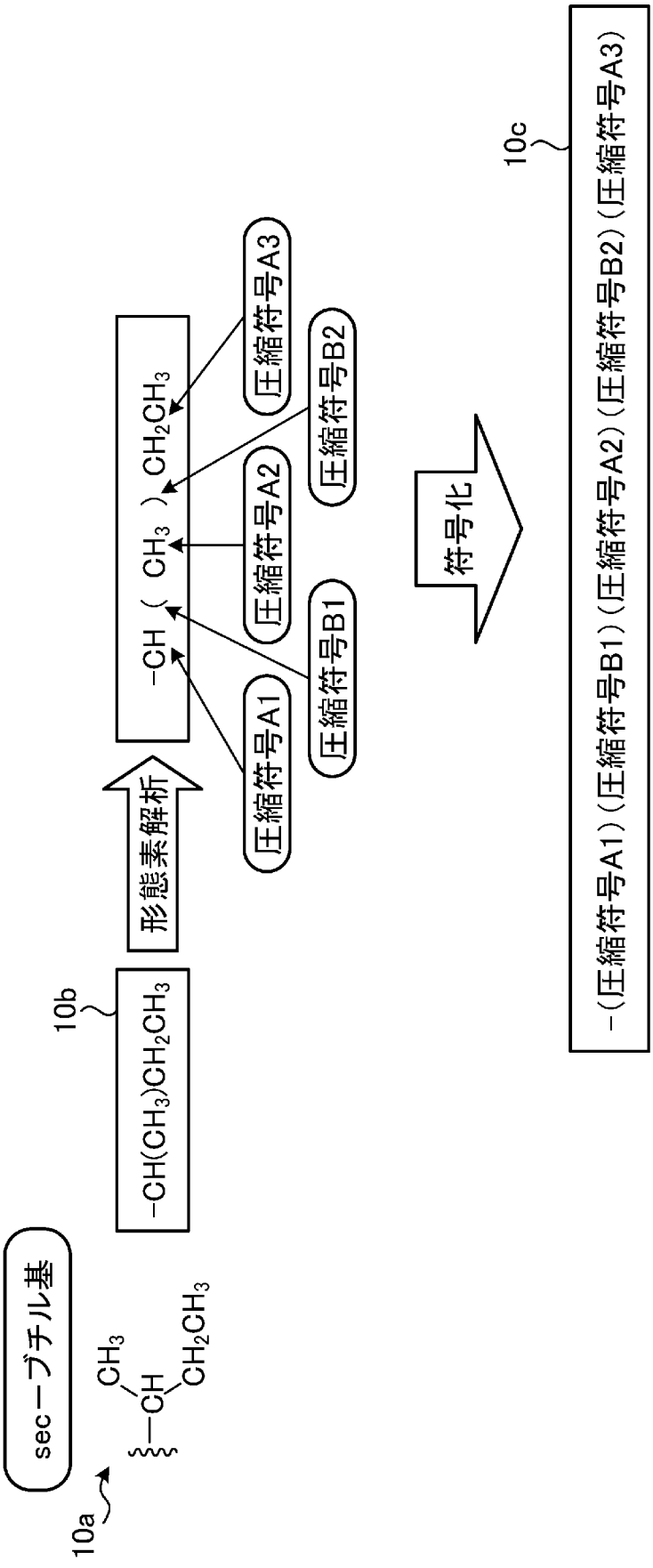
[請求項12] 前記第１符号配列に含まれる前記基毎の圧縮符号を、ポアンカレ空間に埋め込むことで、基のベクトルを生成し、前記基一次構造を構成する基のベクトルを加算することで、前記基一次構造にベクトルを割り当てるベクトル算出部を更に有することを特徴とする請求項１１に記載の情報処理装置。

[請求項13] 前記ベクトル算出部は、類似の性質となる基の圧縮符号を、ポアンカレ空間の類似の位置に埋め込み、前記ポアンカレ空間の位置に応じたベクトルを埋め込んだ圧縮符号に割り当てることを特徴とする請求項１２に記載の情報処理装置。

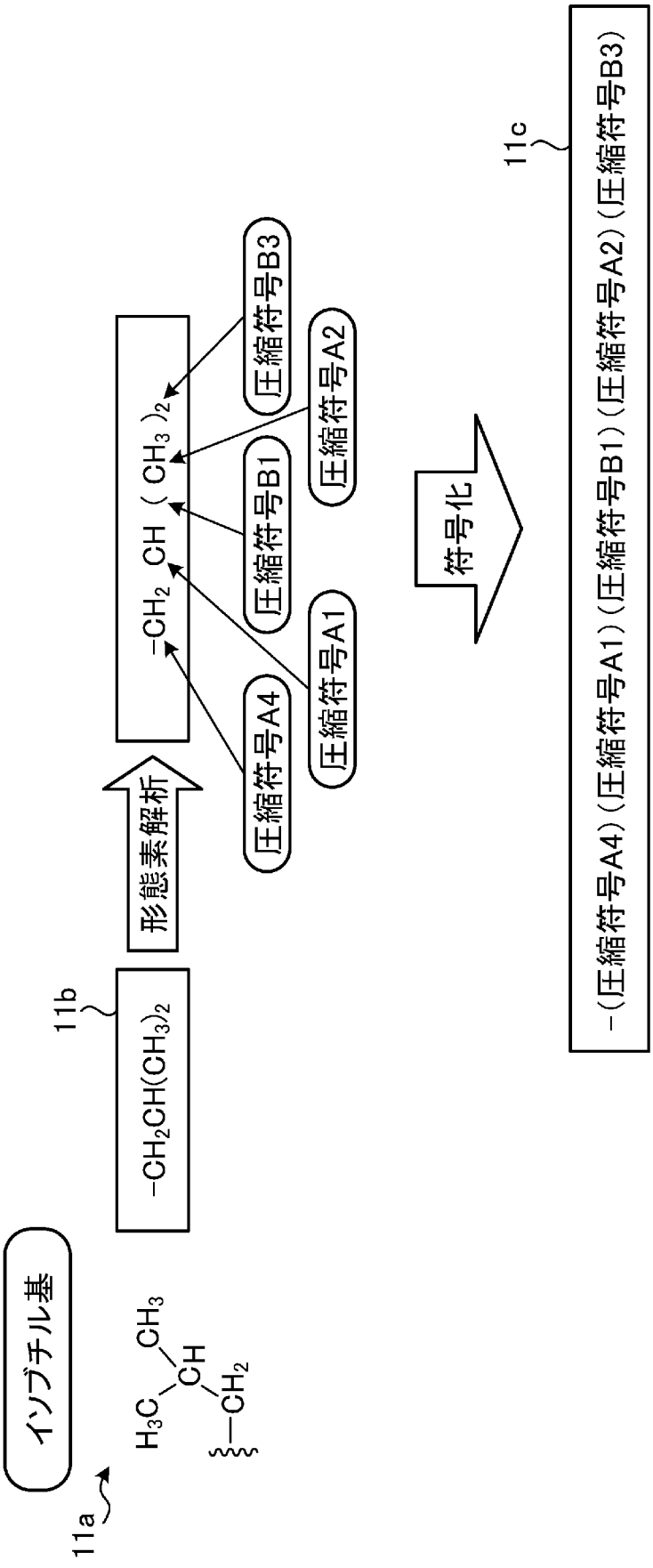
[請求項14] 前記ベクトル算出部は、前記第２符号配列に含まれる複数の基一次構造のベクトルを順に遷移テーブルに登録する処理を更に実行することを特徴とする請求項１２または１３に記載の情報処理装置。

[請求項15] 前記遷移テーブルに登録されたベクトルを基にして、各化合物の類似度を評価する類似度評価部を更に有することを特徴とする請求項１４に記載の情報処理装置。

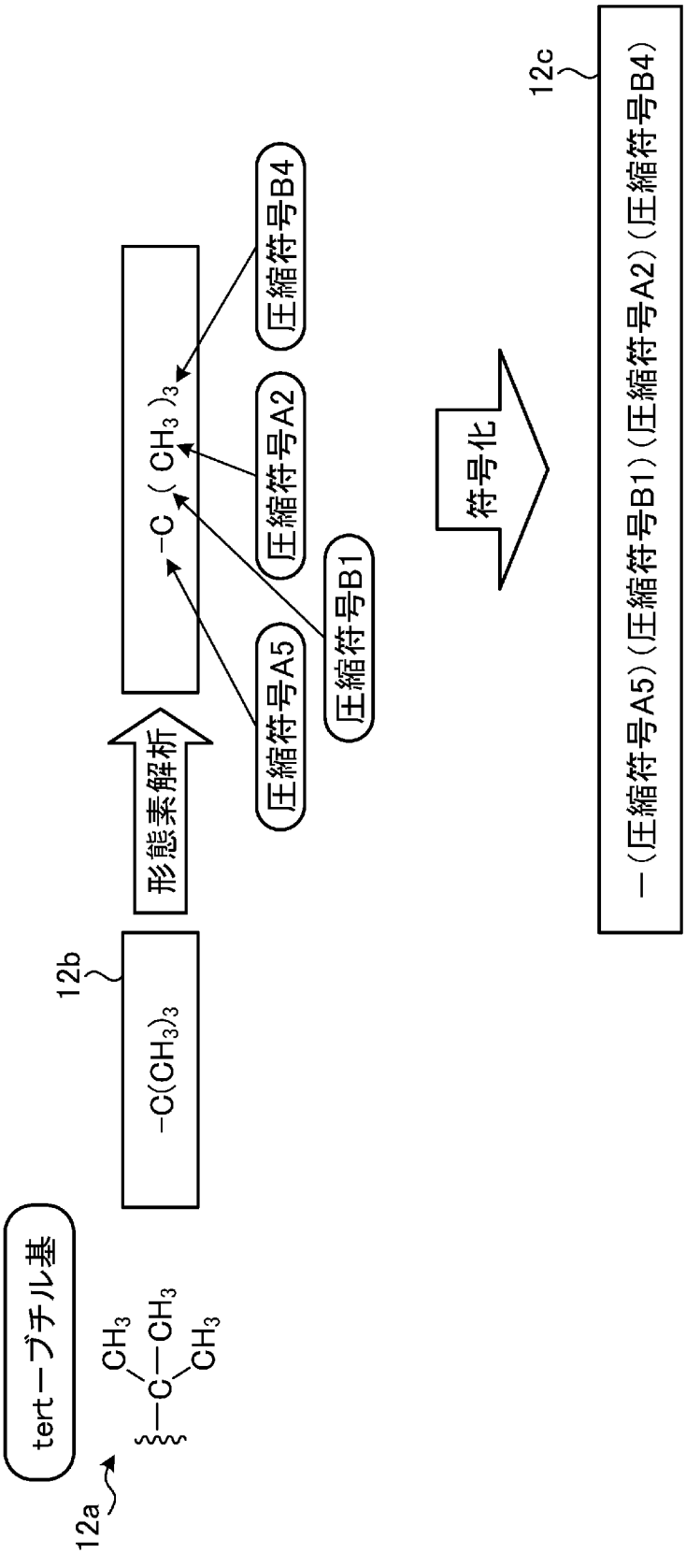
[図1]


sec－ブチル基
10a
CH3
CH
CH2CH3
10b
－CH(CH3)CH2CH3
形態素解析
－CH（ CH3 ）CH2CH3
圧縮符号A1
圧縮符号A2
圧縮符号A3
圧縮符号B1
圧縮符号B2
符号化
10c
－（圧縮符号A1）（圧縮符号B1）（圧縮符号A2）（圧縮符号B2）（圧縮符号A3）

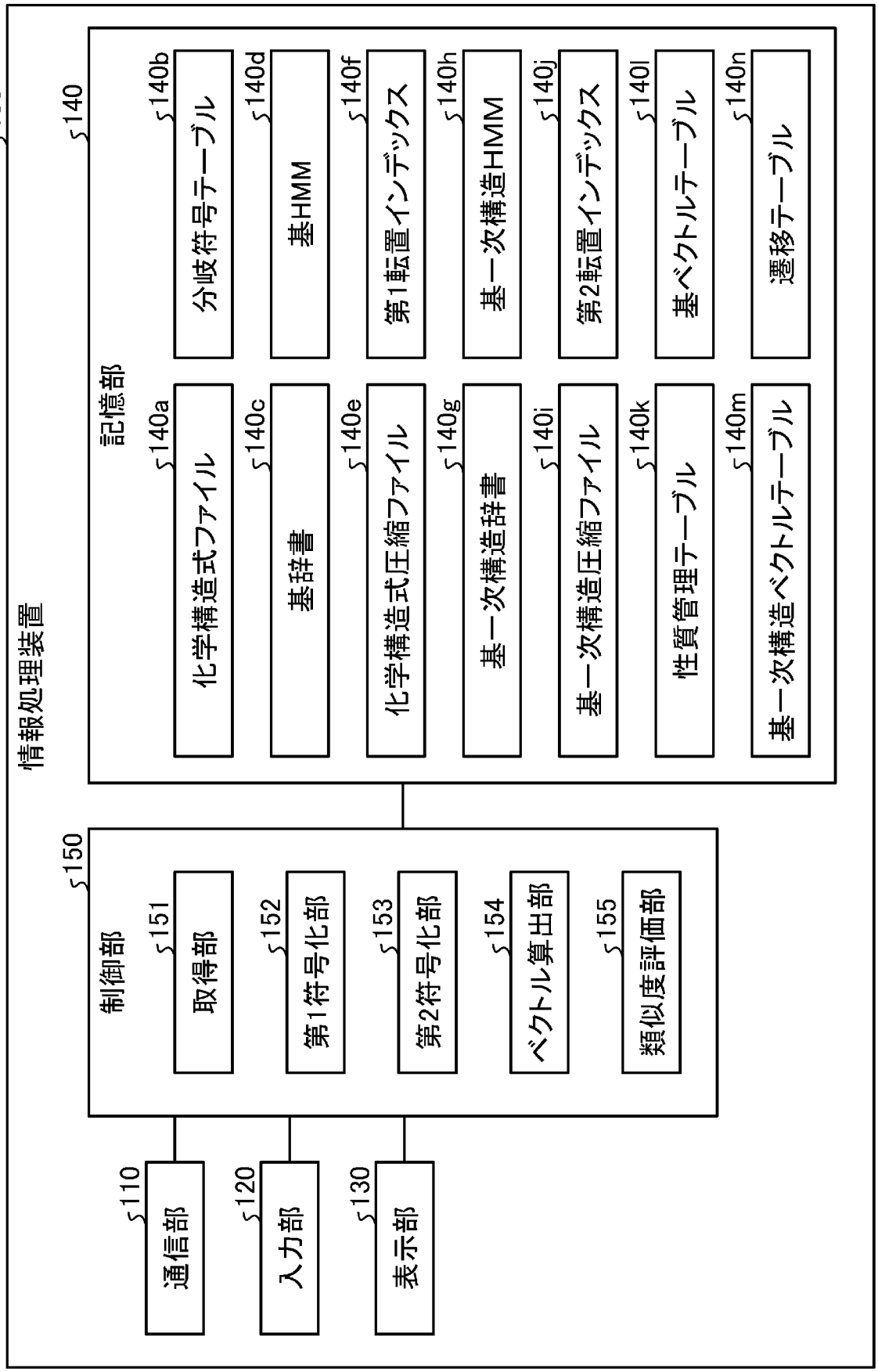
[図2]


イソブチル基
11a
H3C
CH3
CH
CH2
11b
-CH2CH(CH3)2
形態素解析
-CH2 CH (CH3)2
圧縮符号A4
圧縮符号B1
圧縮符号A1
圧縮符号A2
圧縮符号B3
符号化
11c
-(圧縮符号A4)(圧縮符号A1)(圧縮符号B1)(圧縮符号A2)(圧縮符号B3)

[図3]


tert－ブチル基
12a
12b
形態素解析
圧縮符号A5
圧縮符号A2
圧縮符号B1
圧縮符号B4
符号化
12c
－（圧縮符号A5）（圧縮符号B1）（圧縮符号A2）（圧縮符号B4）

[図4]


100
情報処理装置
140
記憶部
140a
化学構造式ファイル
140b
分岐符号テーブル
140c
基辞書
140d
基HMM
140e
化学構造式圧縮ファイル
140f
第1転置インデックス
140g
基一次構造辞書
140h
基一次構造HMM
140i
基一次構造圧縮ファイル
140j
第2転置インデックス
140k
性質管理テーブル
140l
基ベクトルテーブル
140m
基一次構造ベクトルテーブル
140n
遷移テーブル
150
制御部
151
取得部
152
第1符号化部
153
第2符号化部
154
ベクトル算出部
155
類似度評価部
110
通信部
120
入力部
130
表示部

[図5]

140a

化合物識別情報	示性式
C101	化合物識別情報「C101」の示性式を示す記号配列の情報
C102	化合物識別情報「C102」の示性式を示す記号配列の情報
C103	化合物識別情報「C103」の示性式を示す記号配列の情報
・・・	・・・

[図6]

140b

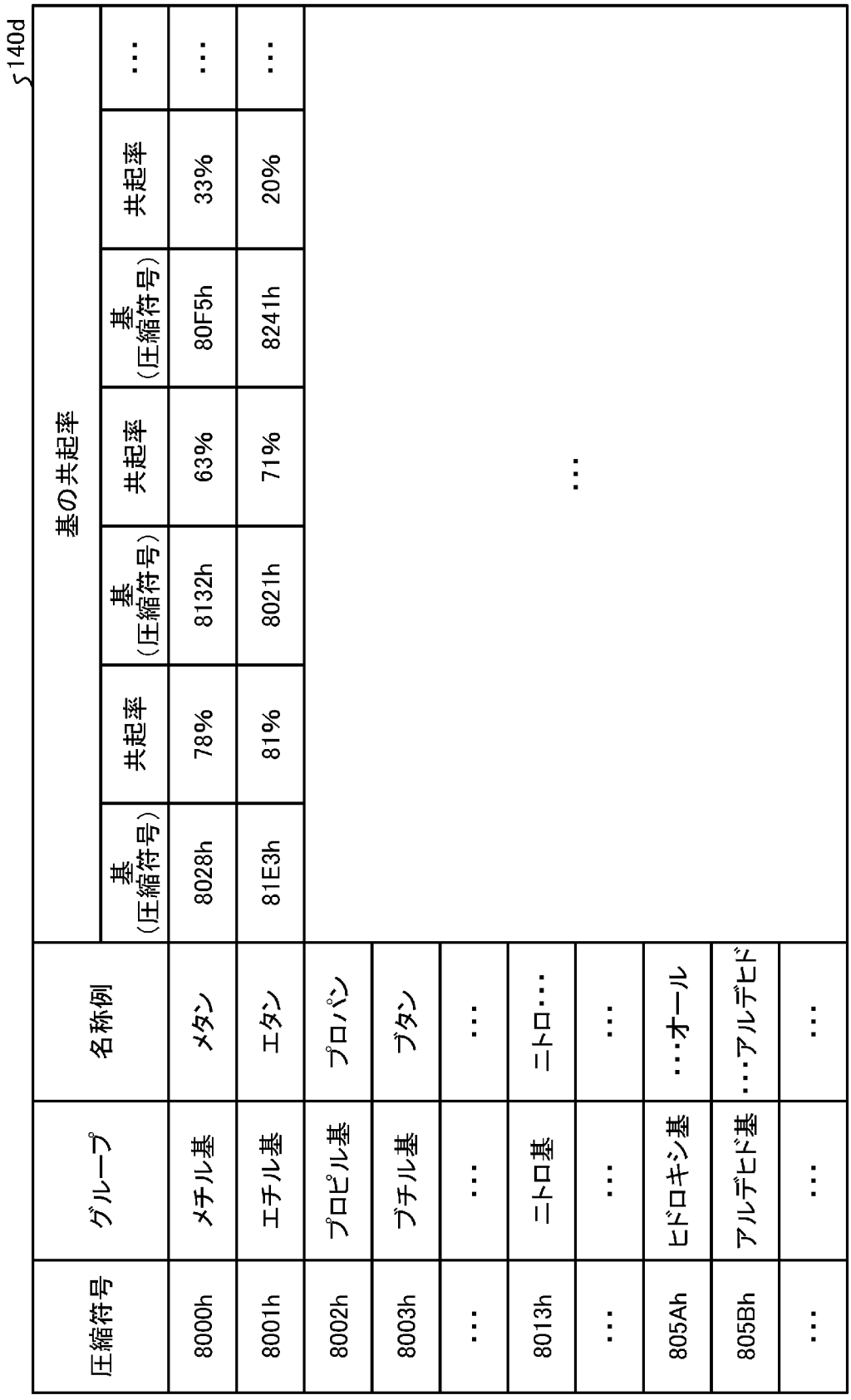
分岐記号	圧縮符号
(	30h
)	31h
$)_2$	32h
$)_3$	33h

[図7]

140c

圧縮符号	グループ	名称例	示性式
8000h	メチル基	メタン	CH3
8001h	エチル基	エタン	CH3CH2
8002h	プロピル基	プロパン	CH3CH2CH2
8003h	ブチル基	ブタン	CH3CH2CH2CH2
・・・	・・・	・・・	・・・
8013h	ニトロ基	ニトロ・・・	NO2
・・・	・・・	・・・	・・・
805Ah	ヒドロキシ基	・・・オール	OH
805Bh	アルデヒド基	・・・アルデヒド	OHO
・・・	・・・	・・・	・・・

[図8]

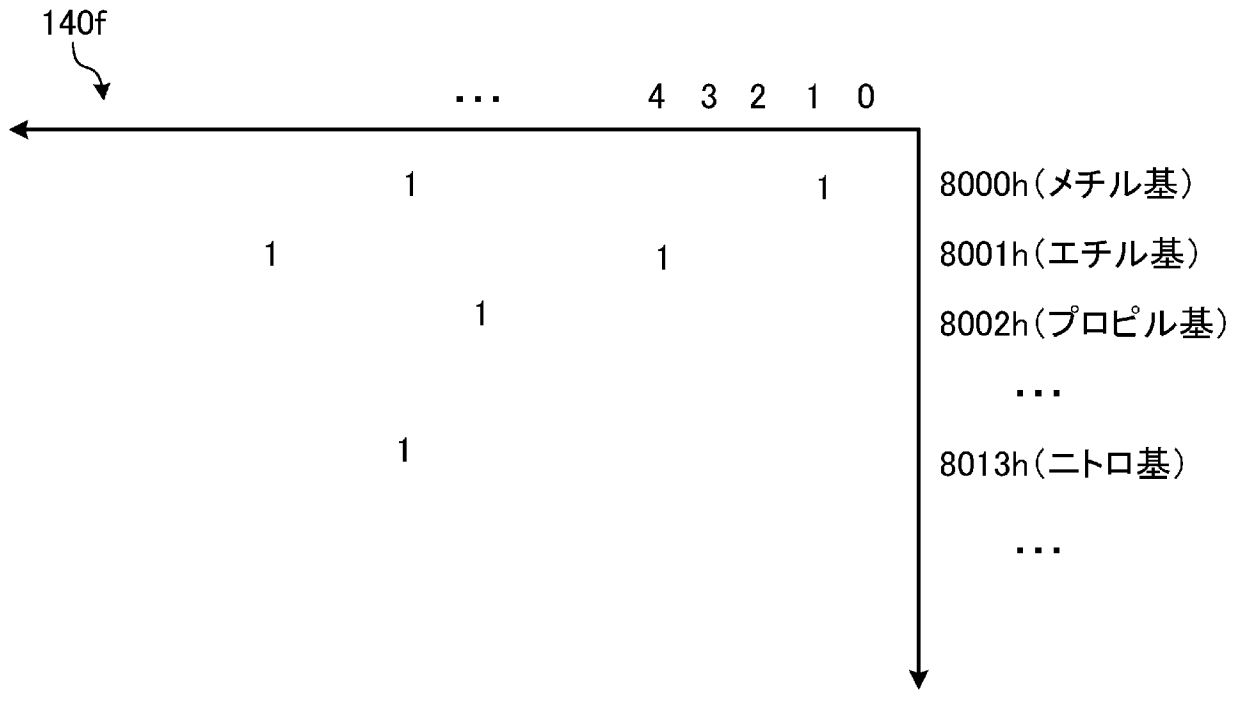

140d
圧縮符号
グループ
名称例
基の共起率
基（圧縮符号）
共起率
基（圧縮符号）
共起率
基（圧縮符号）
共起率
8000h
メチル基
メタン
8028h
78%
8132h
63%
80F5h
33%
8001h
エチル基
エタン
81E3h
81%
8021h
71%
8241h
20%
8002h
プロピル基
プロパン
8003h
ブチル基
ブタン
8013h
ニトロ基
ニトロ…
805Ah
ヒドロキシ基
…オール
805Bh
アルデヒド基
…アルデヒド

[図9]

140e

化合物識別情報	圧縮符号配列
C101	化合物識別情報「C101」の示性式を符号化した圧縮符号配列
C102	化合物識別情報「C102」の示性式を符号化した圧縮符号配列
C103	化合物識別情報「C103」の示性式を符号化した圧縮符号配列
・・・	・・・

[図10]

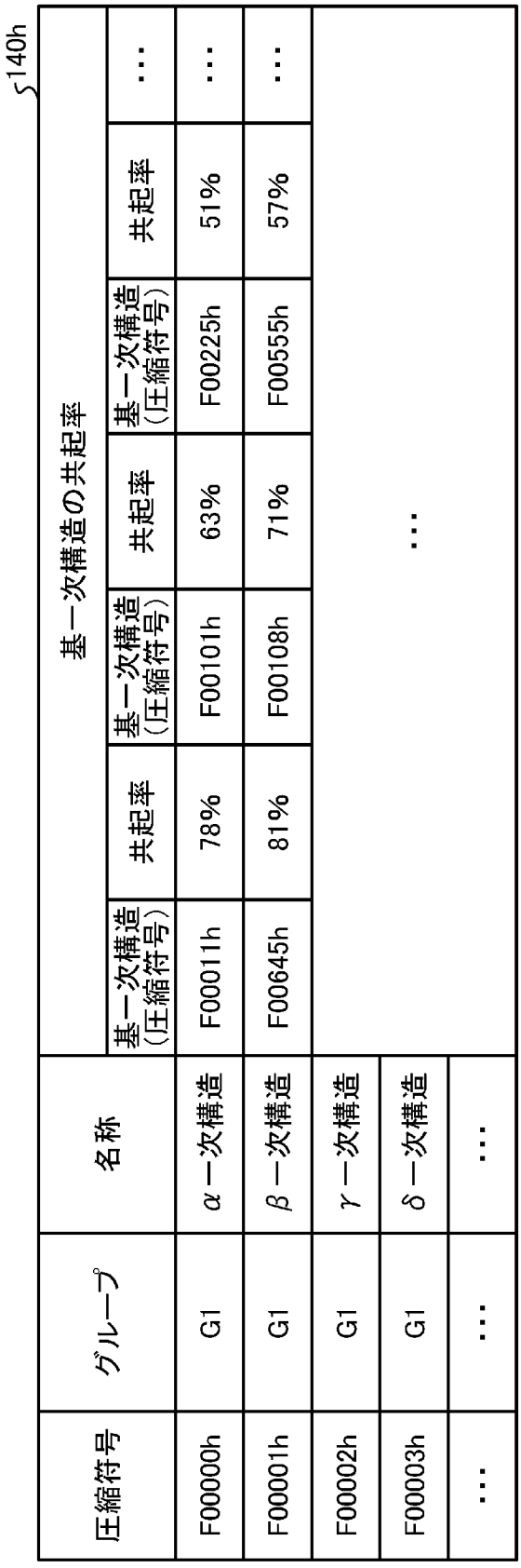


[図11]

140g

圧縮符号	グループ	名称	圧縮符号配列
F0O000h	G1	α一次構造	02h8028h・・・・・・・03h
F0O001h	G1	β一次構造	02h80F5h・・・・03h
F0O002h	G1	γ一次構造	02h81E3h・・・・・・・・・03h
F0O003h	G1	δ一次構造	・・・・・
・・・	・・・	・・・	・・・・・

[図12]

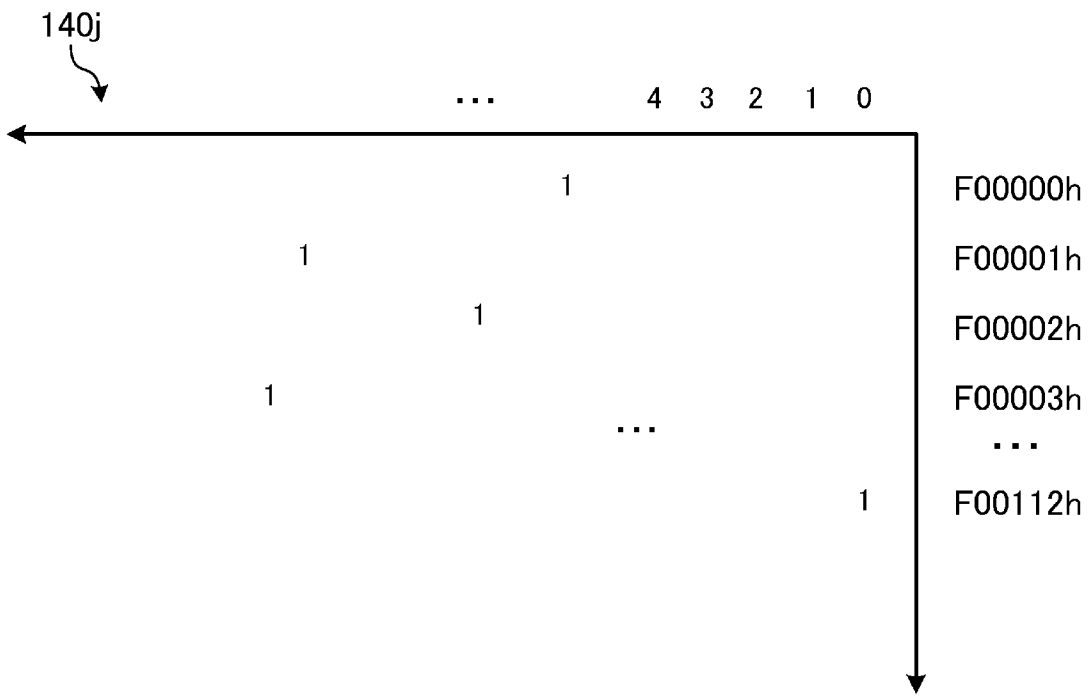

140h
圧縮符号
グループ
名称
基一次構造の共起率
基一次構造（圧縮符号）
共起率
基一次構造（圧縮符号）
共起率
基一次構造（圧縮符号）
共起率
F00000h
G1
α一次構造
F00011h
78%
F00101h
63%
F00225h
51%
F00001h
G1
β一次構造
F00645h
81%
F00108h
71%
F00555h
57%
F00002h
G1
γ一次構造
F00003h
G1
δ一次構造

[図13]

140i

化合物識別情報	基一次構造の圧縮符号配列				
C101	F00112h	F00332h	F00110h	F00460h	・・・
C102	F00232h	F00582h	F00874h	F00209h	・・・
C103	F00102h	F00082h	F00874h	F00009h	・・・
・・・	・・・				

[図14]



[図15]

140k

性質番号	圧縮符号	名称
I101	8000h	メタン
	8120h	・・・
	8234h	・・・
I102	8004h	エタン
	8005h	・・・
	8006h	・・・
	8007h	・・・
・・・	・・・	・・・

[図16A]

140l

圧縮符号	ベクトル
8000h	Vec1－1
8001h	Vec1－2
8002h	Vec1－3
・・・	・・・
8013h	Vec1－3
・・・	・・・
805Ah	Vec1－10
805Bh	Vec1－11
・・・	・・・

[図16B]

140m

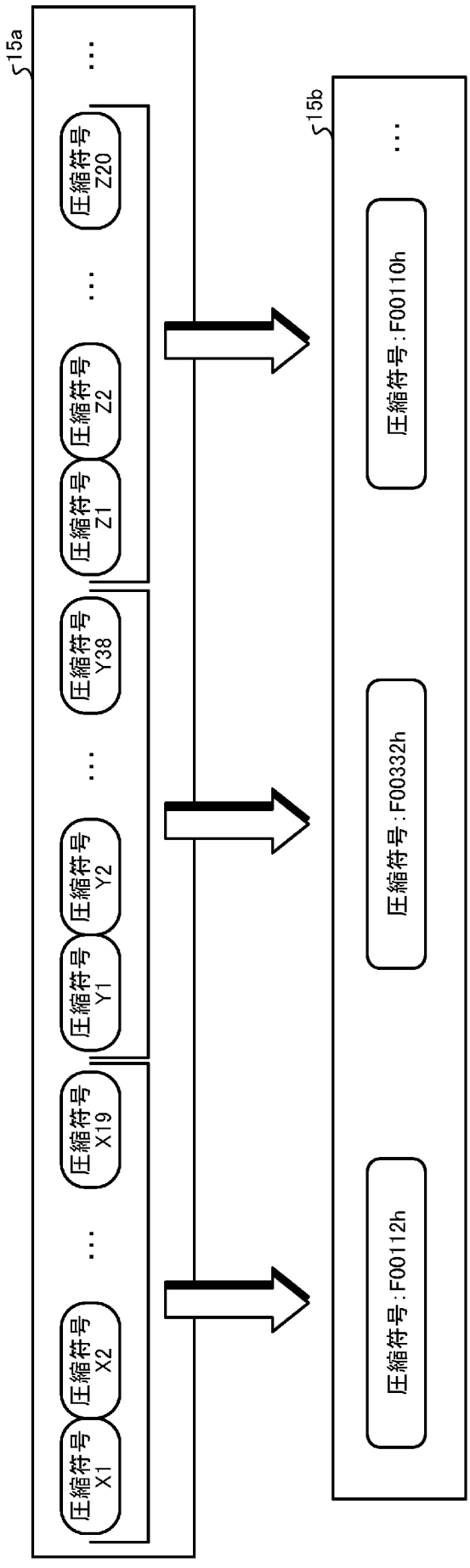
圧縮符号	ベクトル
F00000h	Vec2－1
F00001h	Vec2－2
F00002h	Vec2－3
F00003h	Vec2－4
・・・	・・・

[図16C]

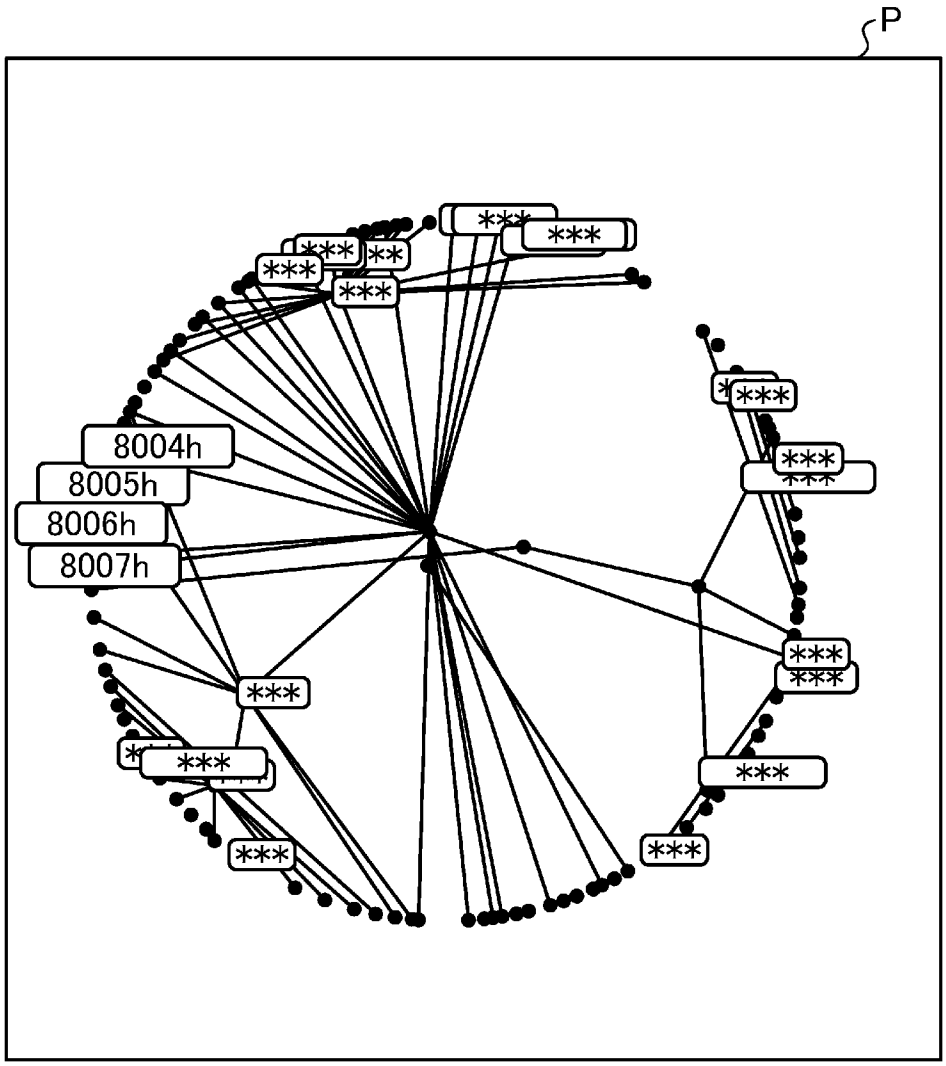
140n

化合物識別情報	ベクトル(1)	ベクトル(2)	ベクトル(3)	・・・
C101	0.28213	0.14975	0.18975	・・・
C102	0.97325	0.23800	0.15575	・・・
C103	0.309085	0.13800	0.14925	・・・
・・・	・・・	・・・	・・・	・・・

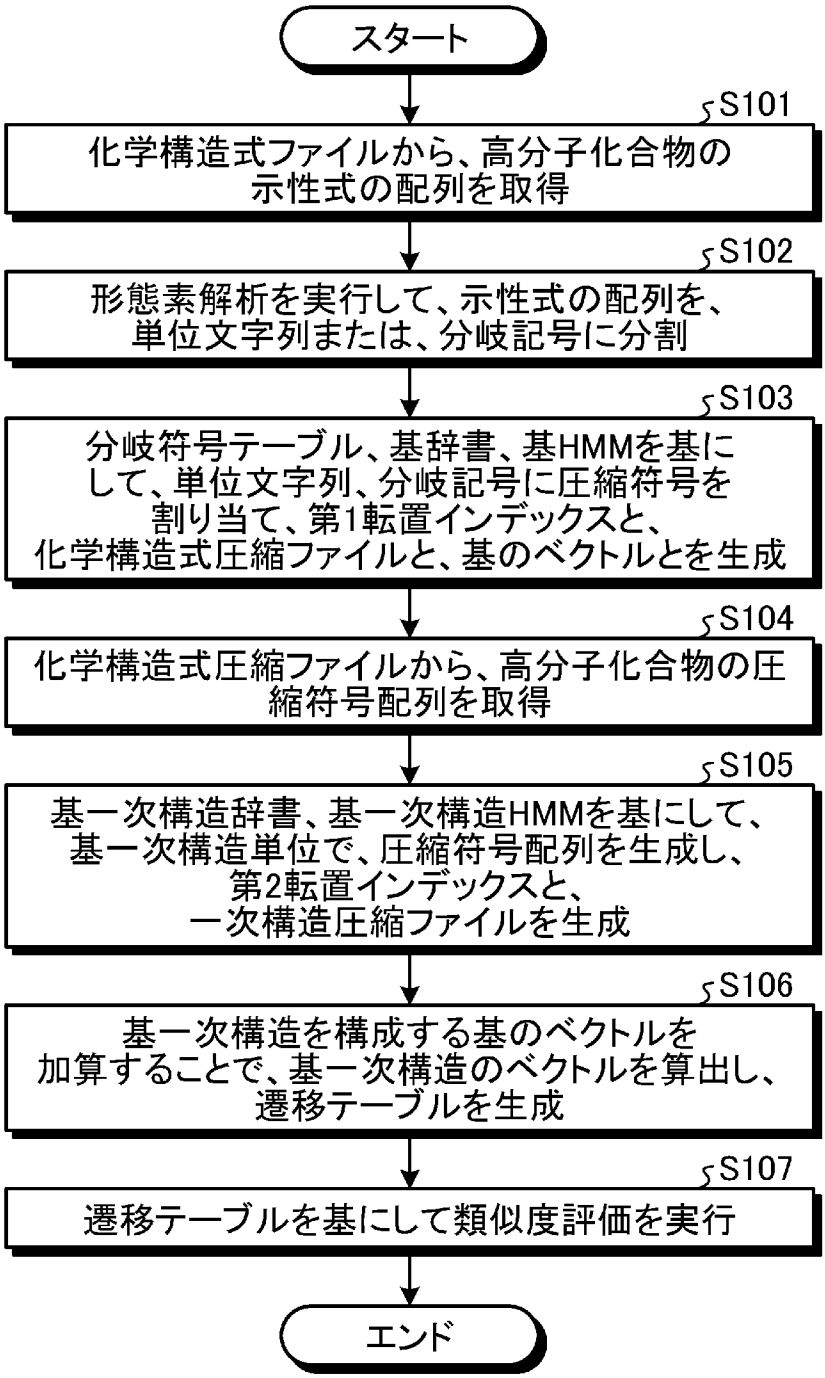
[図17]


15a
15b
圧縮符号 X1
圧縮符号 X2
圧縮符号 X19
圧縮符号 Y1
圧縮符号 Y2
圧縮符号 Y38
圧縮符号 Z1
圧縮符号 Z2
圧縮符号 Z20
圧縮符号:F00112h
圧縮符号:F00332h
圧縮符号:F00110h

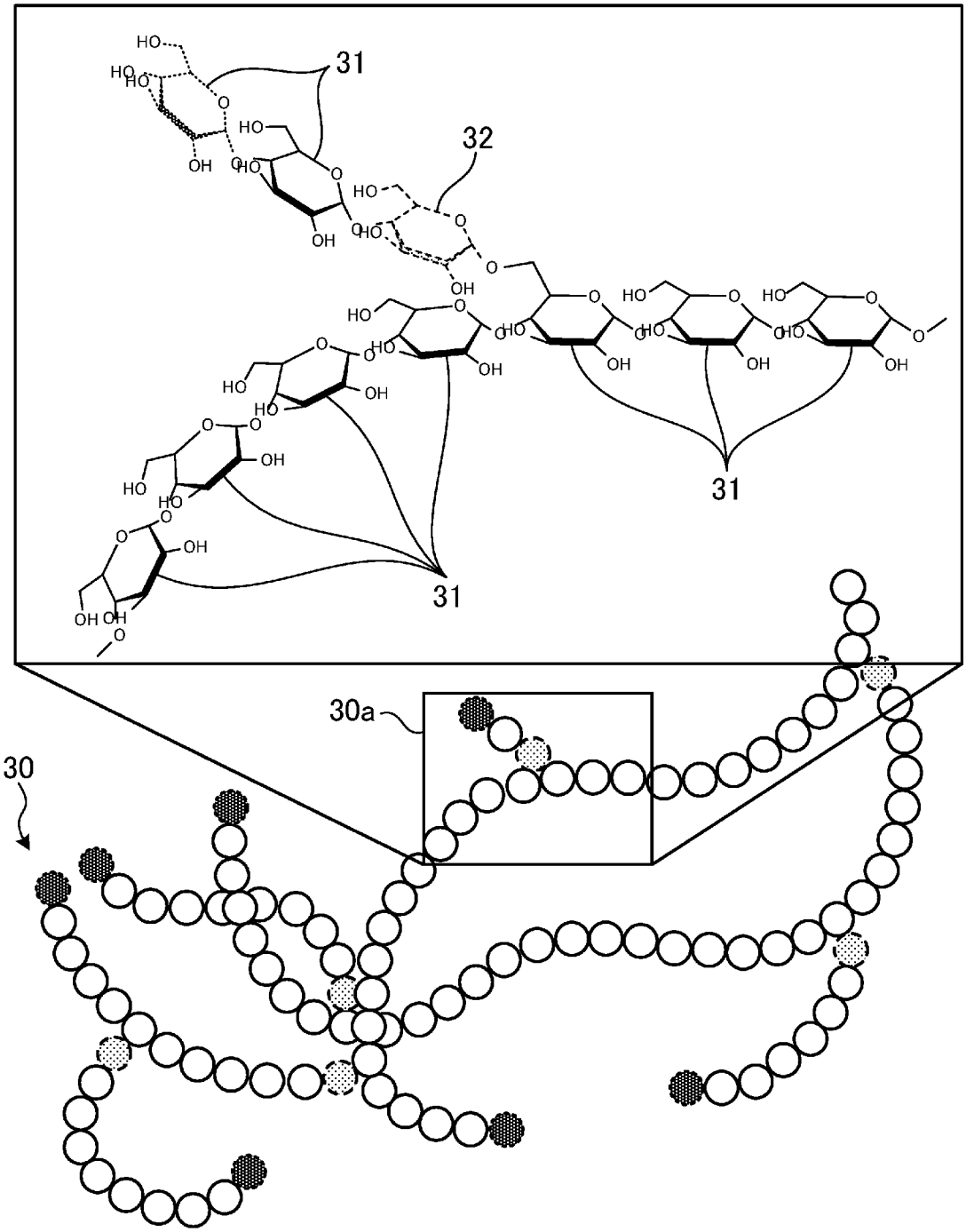
[図18]



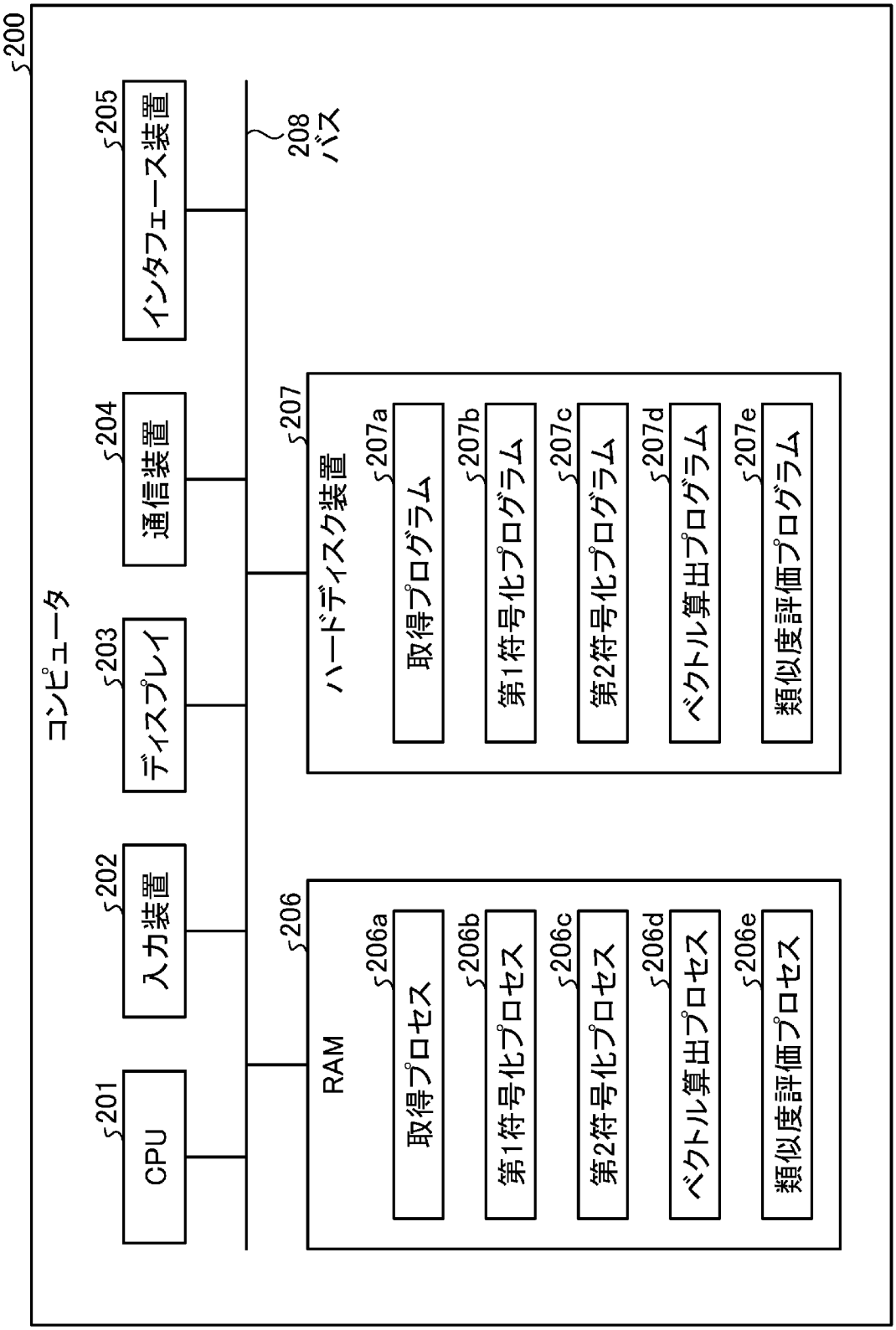
[図19]


スタート
S101
化学構造式ファイルから、高分子化合物の示性式の配列を取得
S102
形態素解析を実行して、示性式の配列を、単位文字列または、分岐記号に分割
S103
分岐符号テーブル、基辞書、基HMMを基にして、単位文字列、分岐記号に圧縮符号を割り当て、第1転置インデックスと、化学構造式圧縮ファイルと、基のベクトルとを生成
S104
化学構造式圧縮ファイルから、高分子化合物の圧縮符号配列を取得
S105
基一次構造辞書、基一次構造HMMを基にして、基一次構造単位で、圧縮符号配列を生成し、第2転置インデックスと、一次構造圧縮ファイルを生成
S106
基一次構造を構成する基のベクトルを加算することで、基一次構造のベクトルを算出し、遷移テーブルを生成
S107
遷移テーブルを基にして類似度評価を実行
エンド

[図20]



[図21]


200
コンピュータ
201
CPU
202
入力装置
203
ディスプレイ
204
通信装置
205
インタフェース装置
208
バス
206
RAM
206a
取得プロセス
206b
第1符号化プロセス
206c
第2符号化プロセス
206d
ベクトル算出プロセス
206e
類似度評価プロセス
207
ハードディスク装置
207a
取得プログラム
207b
第1符号化プログラム
207c
第2符号化プログラム
207d
ベクトル算出プログラム
207e
類似度評価プログラム

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/022368

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G16C20/40(2019.01)i
FI: G16C20/40

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)
Int. Cl. G16C20/40

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 1971-2020
Registered utility model specifications of Japan 1996-2020
Published registered utility model applications of Japan 1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus/JMEDPlus/JST7580 (JDreamIII);PubMed

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2018-147374 A (FUJITSU LTD.) 20 September 2018, entire text, all drawings	1-15
A	WO 2017/017738 A1 (FUJITSU LTD.) 02 February 2017, entire text, all drawings	1-15
A	JP 2019-204362 A (Z HOLDINGS CORP.) 28 November 2019, entire text, all drawings	1-15
A	US 2012/0109972 A1 (BOYER, Stephen Kane) 03 May 2012, entire text, all drawings	1-15
A	US 2013/0124152 A1 (CHO, Kwang-Hwi) 16 May 2013, entire text, all drawings	1-15

☐ 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
11.09.2020	24.09.2020
Name and mailing address of the ISA/ Japan Patent Office 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/JP2020/022368

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
JP 2018-147374 A	20.09.2018	(Family: none)	
WO 2017/017738 A1	02.02.2017	US 2018/0143954 A1	
JP 2019-204362 A	28.11.2019	(Family: none)	
US 2012/0109972 A1	03.05.2012	US 2008/0004810 A1	
US 2013/0124152 A1	16.05.2013	KR 10-1236966 B1	

国際調査報告

国際出願番号
PCT/JP2020/022368

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G16C 20/40(2019.01)i
FI: G16C20/40

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））

G16C20/40

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	１９７１－２０２０年
日本国実用新案登録公報	１９９６－２０２０年
日本国登録実用新案公報	１９９４－２０２０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）

JSTPlus/JMEDPlus/JST7580 (JDreamIII); PubMed

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 2018-147374 A (富士通株式会社) 20.09.2018 (2018-09-20) 全文,全図	1-15
A	WO 2017/017738 A1 (富士通株式会社) 02.02.2017 (2017-02-02) 全文,全図	1-15
A	JP 2019-204362 A (Ｚホールディングス株式会社) 28.11.2019 (2019-11-28) 全文,全図	1-15
A	US 2012/0109972 A1 (BOYER STEPHEN KANE) 03.05.2012 (2012-05-03) 全文,全図	1-15
A	US 2013/0124152 A1 (CHO KWANG-HWI) 16.05.2013 (2013-05-16) 全文,全図	1-15

☐ 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
11.09.2020	24.09.2020

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）

松野　広一　5R 3658

電話番号 03-3581-1101 内線 3502

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/JP2020/022368

引用文献	公表日	パテントファミリー文献	公表日
JP 2018-147374 A	20.09.2018	（ファミリーなし）	
WO 2017/017738 A1	02.02.2017	US 2018/0143954 A1	
JP 2019-204362 A	28.11.2019	（ファミリーなし）	
US 2012/0109972 A1	03.05.2012	US 2008/0004810 A1	
US 2013/0124152 A1	16.05.2013	KR 10-1236966 B1	