(19)日本国特許庁(JP)

# (12)特許公報(B2)

(11)特許番号
特許第7371779号
(P7371779)

(45)発行日　令和5年10月31日(2023.10.31)

(24)登録日　令和5年10月23日(2023.10.23)

(51)国際特許分類

| | | | |
|---|---|---|---|
| *Ｇ１６Ｃ　20/40　(2019.01)* | | Ｇ１６Ｃ | 20/40 |
| *Ｇ１６Ｃ　20/90　(2019.01)* | | Ｇ１６Ｃ | 20/90 |

ＦＩ

請求項の数　7　（全22頁）

(21)出願番号　特願2022-528386(P2022-528386)
(86)(22)出願日　令和2年6月5日(2020.6.5)
(86)国際出願番号　PCT/JP2020/022368
(87)国際公開番号　WO2021/245926
(87)国際公開日　令和3年12月9日(2021.12.9)
審査請求日　令和4年9月9日(2022.9.9)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　110002147
弁理士法人酒井国際特許事務所
(72)発明者　片岡 正弘
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
(72)発明者　古賀 裕美
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
(72)発明者　多賀 万里子
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

最終頁に続く

(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

(57)【特許請求の範囲】

【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】
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。
【発明の詳細な説明】
【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】
【文献】特開平１１－３０６１８８号公報

【文献】国際公開第２００７／１３９０３７号
【文献】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１４０ｌ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１４０ｌ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１４０ｌ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】

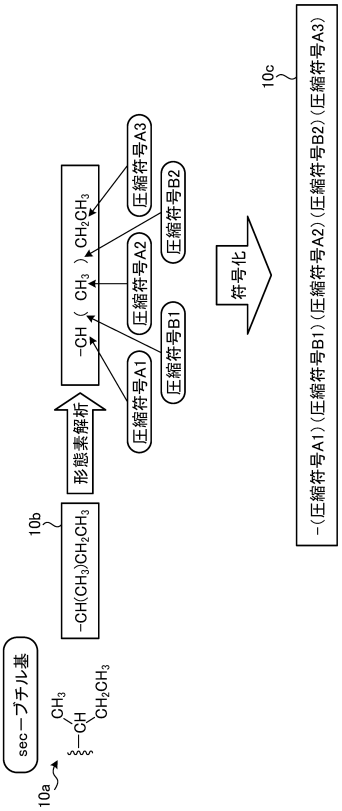
【０１０６】

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

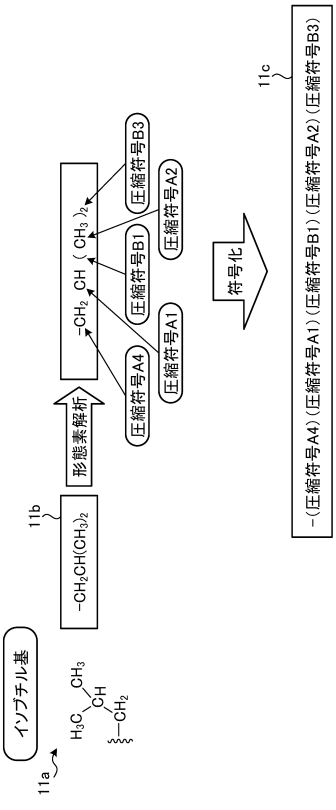
１５１　取得部
１５２　第１符号化部
１５３　第２符号化部
１５４　ベクトル算出部
１５５　類似度評価部

【図面】

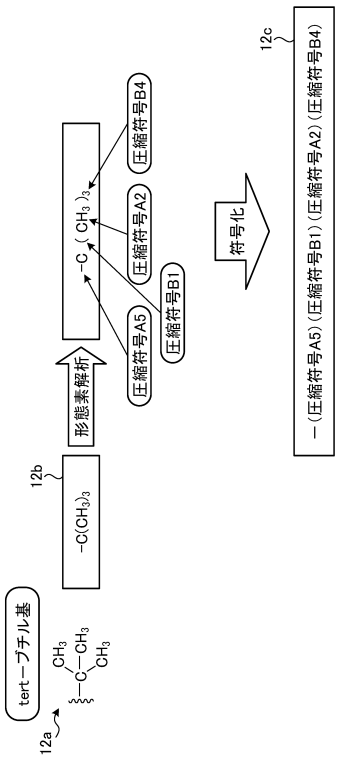
【図１】



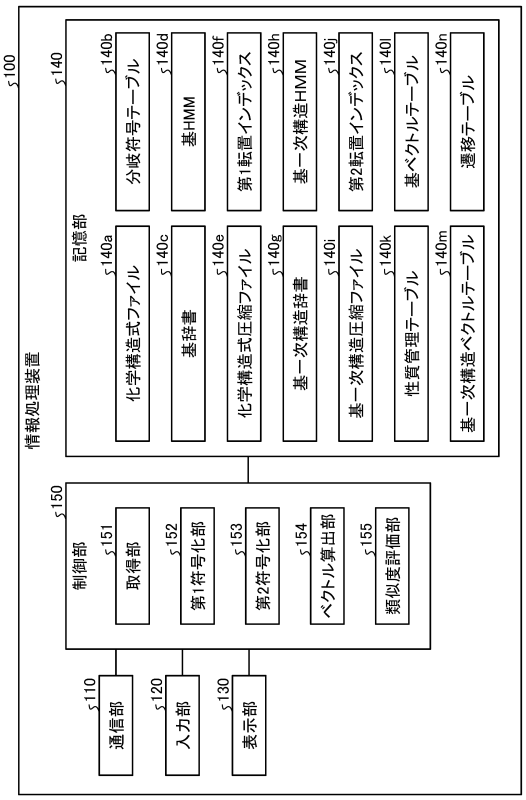

【図２】



【図３】




【図４】




【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

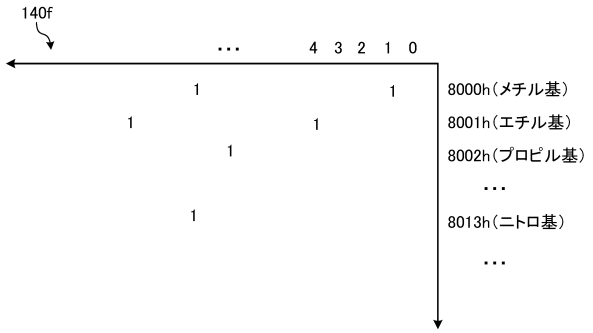
| 圧縮符号 | グループ | 名称例 | 基の共起率 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 基（圧縮符号） | 共起率 | 基（圧縮符号） | 共起率 | 基（圧縮符号） | 共起率 | … | | | |
| 8000h | メチル基 | メタン | 8028h | 78% | 8132h | 63% | 80F5h | 33% | … | | | |
| 8001h | エチル基 | エタン | 81E3h | 81% | 8021h | 71% | 8241h | 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

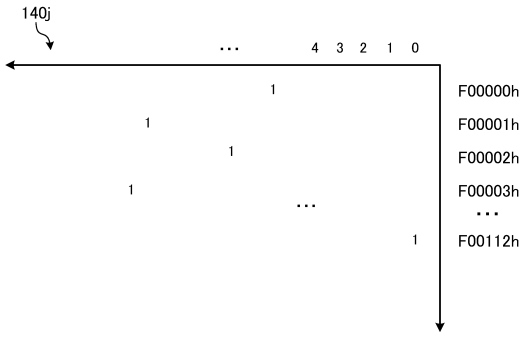
| 圧縮符号 | グループ | 名称 | 基一次構造の共起率 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 基一次構造（圧縮符号） | 共起率 | 基一次構造（圧縮符号） | 共起率 | 基一次構造（圧縮符号） | 共起率 | ・・・ |
| F00000h | G1 | α一次構造 | F00011h | 78% | F00101h | 63% | F00225h | 51% | ・・・ |
| F00001h | G1 | β一次構造 | F00645h | 81% | F00108h | 71% | F00555h | 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

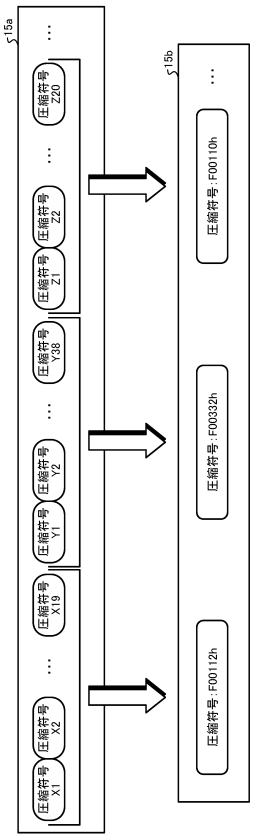
| 圧縮符号 | ベクトル |
|---|---|
| F00000h | Vec2－1 |
| F00001h | Vec2－2 |
| F00002h | Vec2－3 |
| F00003h | Vec2－4 |
| ・・・ | ・・・ |

【図１６Ｃ】

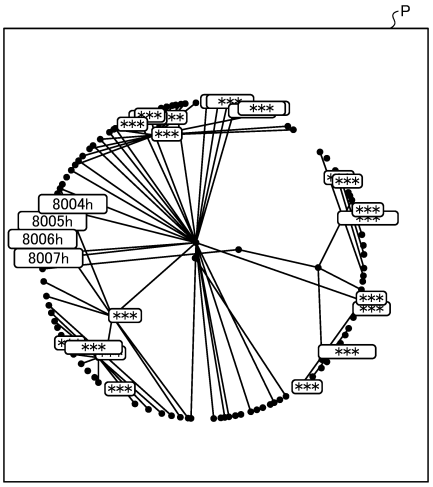
140n

| 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 | ・・・ |
| ・・・ | ・・・ | ・・・ | ・・・ | ・・・ |

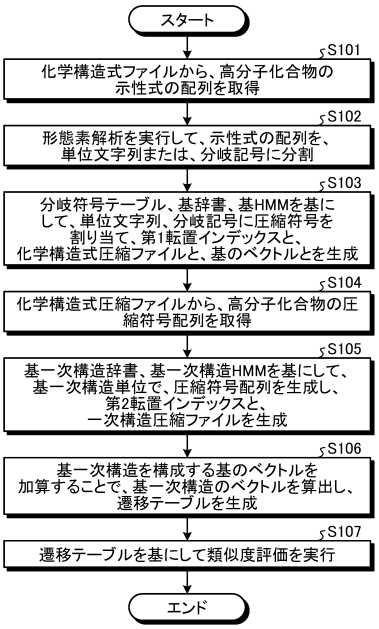
【図１７】



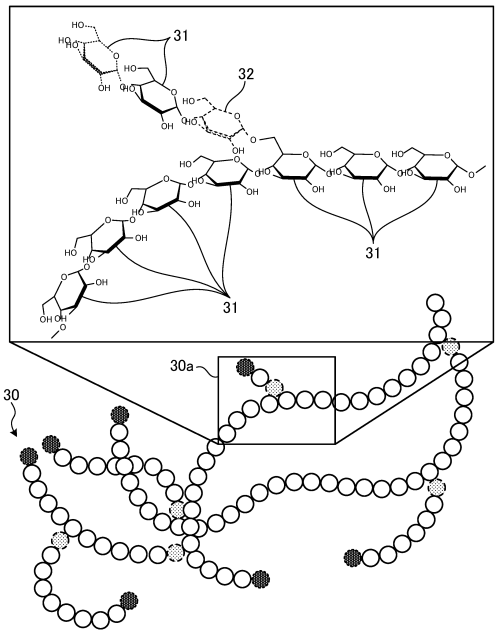

【図１８】



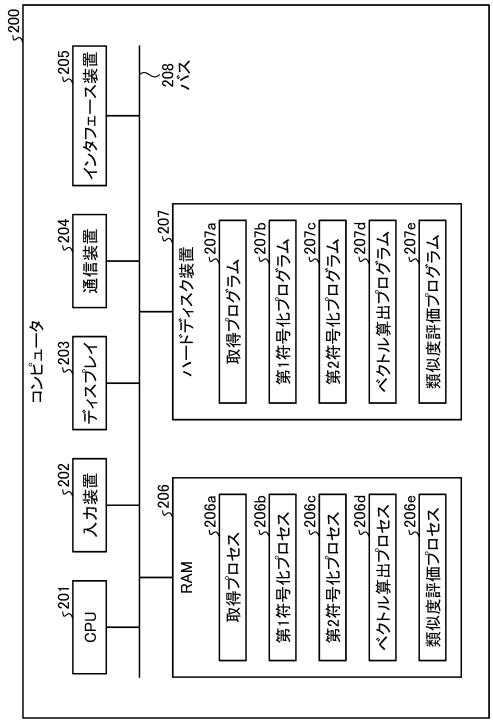

【図１９】




【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

フロントページの続き

審査官　松野 広一

(56)参考文献　特開２０１８－１４７３７４（ＪＰ，Ａ）
国際公開第２０１７／０１７７３８（WO，Ａ１）
特開２０１９－２０４３６２（ＪＰ，Ａ）
米国特許出願公開第２０１２／０１０９９７２（ＵＳ，Ａ１）
米国特許出願公開第２０１３／０１２４１５２（ＵＳ，Ａ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１６Ｃ　１０／００－９９／００
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩＩ）
ＰｕｂＭｅｄ