(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115668385 A
(43)申请公布日 2023.01.31

(21)申请号 202080101632.0
(22)申请日 2020.06.05
(85)PCT国际申请进入国家阶段日
2022.12.01
(86)PCT国际申请的申请数据
PCT/JP2020/022368 2020.06.05
(87)PCT国际申请的公布数据
WO2021/245926 JA 2021.12.09
(71)申请人 富士通株式会社
地址 日本神奈川县
(72)发明人 片冈正弘 古贺裕美 多贺万里子
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 舒艳君
(51)Int.Cl.
G16C 20/40(2006.01)

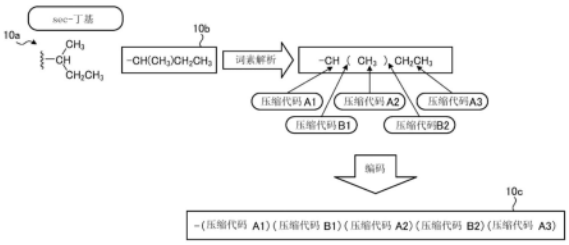
权利要求书2页 说明书11页 附图18页

(54)发明名称

信息处理程序、信息处理方法以及信息处理装置

(57)摘要

信息处理装置将表示化合物的示性式的排列分割为排列的最小单位的字符串以及表示化合物的分支部分的分支符号。信息处理装置通过基于表示化合物的示性式的排列与压缩代码之间的关系的基团词典，对最小单位的字符串分配压缩代码，并对分支符号分配与分支部分的种类对应的压缩代码，来生成对排列进行编码后的第一代码排列。信息处理装置基于表示化合物的示性式的排列的基团一次结构与压缩代码之间的关系的基团一次结构词典，生成以基团一次结构的单位对第一代码排列所包含的压缩代码进行编码后的第二代码排列。



CN 115668385 A

1.一种信息处理程序,其特征在于,使计算机执行如下处理:

将表示化合物的示性式的排列分割为上述排列的最小单位的字符串以及表示上述化合物的分支部分的分支符号,

通过基于表示化合物的示性式的排列与压缩代码之间的关系的基团词典,对上述最小单位的字符串分配压缩代码,并对上述分支符号分配与上述分支部分的种类对应的压缩代码,来生成对上述排列进行编码后的第一代码排列,

基于表示化合物的示性式的排列的基团一次结构与压缩代码之间的关系的基团一次结构词典,生成以上述基团一次结构的单位对上述第一代码排列所包含的压缩代码进行编码后的第二代码排列。

2.根据权利要求1所述的信息处理程序,其特征在于,

还执行如下处理:通过将上述第一代码排列所包含的每个上述基团的压缩代码埋入至庞加莱空间,来生成基团的向量,并且通过将构成上述基团一次结构的基团的向量相加,来对上述基团一次结构分配向量。

3.根据权利要求2所述的信息处理程序,其特征在于,

分配上述向量的处理将成为相似的性质的基团的压缩代码埋入至庞加莱空间的相似的位置,并将与上述庞加莱空间的位置对应的向量分配给所埋入的压缩代码。

4.根据权利要求2或者3所述的信息处理程序,其特征在于,

分配上述向量的处理还执行将上述第二代码排列所包含的多个基团一次结构的向量依次登记于迁移表格的处理。

5.根据权利要求4所述的信息处理程序,其特征在于,

还执行基于登记于上述迁移表格的向量评价各化合物的相似度的处理。

6.一种信息处理方法,是计算机执行的信息处理方法,其特征在于,执行如下处理:

将表示化合物的示性式的排列分割为上述排列的最小单位的字符串以及表示上述化合物的分支部分的分支符号,

通过基于表示化合物的示性式的排列与压缩代码之间的关系的基团词典,对上述最小单位的字符串分配压缩代码,并对上述分支符号分配与上述分支部分的种类对应的压缩代码,来生成对上述排列进行编码后的第一代码排列,

基于表示化合物的示性式的排列的基团一次结构与压缩代码之间的关系的基团一次结构词典,生成以上述基团一次结构的单位对上述第一代码排列所包含的压缩代码进行编码后的第二代码排列。

7.根据权利要求6所述的信息处理方法,其特征在于,

执行如下处理:通过将上述第一代码排列所包含的每个上述基团的压缩代码埋入至庞加莱空间,来生成基团的向量,并且通过将构成上述基团一次结构的基团的向量相加,来对上述基团一次结构分配向量。

8.根据权利要求7所述的信息处理方法,其特征在于,

分配上述向量的处理将成为相似的性质的基团的压缩代码埋入至庞加莱空间的相似的位置,并将与上述庞加莱空间的位置对应的向量分配给所埋入的压缩代码。

9.根据权利要求7或者8所述的信息处理方法,其特征在于,

分配上述向量的处理还执行将上述第二代码排列所包含的多个基团一次结构的向量

依次登记于迁移表格的处理。

10.根据权利要求9所述的信息处理方法,其特征在于,

还执行基于登记于上述迁移表格的向量评价各化合物的相似度的处理。

11.一种信息处理装置,其特征在于,具有:

第一编码部,将表示化合物的示性式的排列分割为上述排列的最小单位的字符串以及表示上述化合物的分支部分的分支符号,并且通过基于表示化合物的示性式的排列与压缩代码之间的关系的基团词典,对上述最小单位的字符串分配压缩代码,并对上述分支符号分配与上述分支部分的种类对应的压缩代码,来生成对上述排列进行编码后的第一代码排列;以及

第二编码部,基于表示化合物的示性式的排列的基团一次结构与压缩代码之间的关系的基团一次结构词典,生成以上述基团一次结构的单位对上述第一代码排列所包含的压缩代码进行编码后的第二代码排列。

12.根据权利要求11所述的信息处理装置,其特征在于,还具有:

向量计算部,通过将上述第一代码排列所包含的每个上述基团的压缩代码埋入至庞加莱空间,来生成基团的向量,并且通过将构成上述基团一次结构的基团的向量相加,来对上述基团一次结构分配向量。

13.根据权利要求12所述的信息处理装置,其特征在于,

上述向量计算部将成为相似的性质的基团的压缩代码埋入至庞加莱空间的相似的位置,并将与上述庞加莱空间的位置对应的向量分配给所埋入的压缩代码。

14.根据权利要求12或者13所述的信息处理装置,其特征在于,

上述向量计算部还执行将上述第二代码排列所包含的多个基团一次结构的向量依次登记于迁移表格的处理。

15.根据权利要求14所述的信息处理装置,其特征在于,还具有:

相似度评价部,基于登记于上述迁移表格的向量评价各化合物的相似度。

信息处理程序、信息处理方法以及信息处理装置

技术领域

[0001] 本发明涉及信息处理程序等。

背景技术

[0002] 存在如淀粉、纤维素等糖类、蛋白质、核酸、天然橡胶等那样分子量超过1000的物质，这样的物质也被称为高分子化合物。

[0003] 这里，在进行新药、新材料的开发的情况下，要求对已经开发的高分子化合物进行检索，并评价相似度。

[0004] 例如，作为评价化合物的相似度的技术，有以下那样的现有技术1、2。在现有技术1中，是利用输入的字符串检索以氨基酸符号记述的蛋白质等化合物的氨基酸排列的系统。另外，在现有技术2中，在对化合物的官能团或者原子类型的数目进行计数等的记述式分配向量，并对预先分配给各化合物的向量评价相似度。

[0005] 专利文献1：日本特开平11-306188号公报

[0006] 专利文献2：国际公开第2007/139037号

[0007] 专利文献3：日本特开2019-204362号公报

[0008] 然而，在上述的现有技术中，有高分子化合物的相似度评价的精度较低这样的问题。

发明内容

[0009] 在一个侧面，本发明的目的在于提供能够使高分子化合物的相似度评价的精度的评价提高的信息处理程序、信息处理方法以及信息处理装置。

[0010] 在第一方案中，使计算机执行以下处理。计算机将表示化合物的示性式的排列分割为排列的最小单位的字符串以及表示化合物的分支部分的分支符号。计算机通过基于表示化合物的示性式的排列与压缩代码之间的关系的基团词典，对最小单位的字符串分配压缩代码，并对分支符号分配与分支部分的种类对应的压缩代码，来生成对排列进行编码后的第一代码排列。计算机基于表示化合物的示性式的排列的基团一次结构与压缩代码之间的关系的基团一次结构词典，生成以基团一次结构的单位对第一代码排列所包含的压缩代码进行编码后的第二代码排列。

[0011] 能够提高高分子化合物的相似度评价的精度。

附图说明

[0012] 图1是用于说明本实施例所涉及的信息处理装置对化合物进行编码的处理的图(1)。

[0013] 图2是用于说明本实施例所涉及的信息处理装置对化合物进行编码的处理的图(2)。

[0014] 图3是用于说明本实施例所涉及的信息处理装置对化合物进行编码的处理的图

(3)。

[0015] 图4是表示本实施例所涉及的信息处理装置的构成的功能框图。

[0016] 图5是表示化学结构式文件的数据结构的一个例子的图。

[0017] 图6是表示分支代码表格的数据结构的一个例子的图。

[0018] 图7是表示基团词典的数据结构的一个例子的图。

[0019] 图8是表示基团HMM的数据结构的一个例子的图。

[0020] 图9是表示化学结构式压缩文件的数据结构的一个例子的图。

[0021] 图10是表示第一转置索引的数据结构的一个例子的图。

[0022] 图11是表示基团一次结构词典的数据结构的一个例子的图。

[0023] 图12是表示基团一次结构HMM的数据结构的一个例子的图。

[0024] 图13是表示基团一次结构压缩文件的数据结构的一个例子的图。

[0025] 图14是表示第二转置索引的数据结构的一个例子的图。

[0026] 图15是表示性质管理表格的数据结构的一个例子的图。

[0027] 图16A是表示基团向量表格的数据结构的一个例子的图。

[0028] 图16B是表示基团一次结构向量表格的数据结构的一个例子的图。

[0029] 图16C是表示迁移表格的数据结构的一个例子的图。

[0030] 图17是用于说明第二编码部的处理的图。

[0031] 图18是表示庞加莱空间的一个例子的图。

[0032] 图19是表示本实施例所涉及的信息处理装置的处理顺序的流程图。

[0033] 图20是用于说明本实施例所涉及的信息处理装置的其它的处理的图。

[0034] 图21是表示实现与实施例的信息处理装置相同的功能的计算机的硬件构成的一个例子的图。

具体实施方式

[0035] 以下,基于附图对本申请公开的信息处理程序、信息处理方法以及信息处理装置的实施例进行详细的说明。此外,并不通过该实施例限定该发明。

[0036] 实施例

[0037] 图1是用于说明本实施例所涉及的信息处理装置对化合物进行编码处理的图(1)。这里,对对化合物进行编码的情况进行说明,但在对高分子化合物进行编码的情况下,也能够相同地进行编码。在图1中,某一化合物所包含的sec-丁基能够进一步分解为甲基、乙基,对对其进行编码的情况进行说明。sec-丁基的化学结构式10a从“CH”分支为甲基“CH_3”、乙基“CH_2CH_3”。以“-CH(CH_3)CH_2CH_3”那样的排列表示sec-丁基的化学结构式10a的示性式10b。

[0038] 信息处理装置通过对示性式10b执行词素解析,将示性式10b所包含的排列分割为排列的最小单位(有意义的单位)的字符串、表示化合物的分支部分的符号。在以下的说明中,将是排列的最小单位的字符串且有意义的单位的代码表述为“单位字符串”,将表示化合物的分支部分的符号表述为“分支符号”。

[0039] 在图1所示的例子中,信息处理装置将示性式10b分割为“CH”、“(”、“CH_3”、“)”以及“CH_2CH_3”。“CH”、“CH_3”以及“CH_2CH_3”是基团(或者,官能团)的单位字符串。“(”、“)”是分支符号。

[0040] 信息处理装置基于表示化合物的示性式的排列的单位字符串与压缩代码之间的关系的基团词典，对各单位字符串分配压缩代码。例如，信息处理装置分别对“CH”、“CH_3”、“CH_2CH_3”分配压缩代码A1、压缩代码A2、压缩代码A3。

[0041] 信息处理装置基于表示分支符号与压缩代码之间的关系的分支代码表格，对各分支符号分配压缩代码。例如，信息处理装置对“(”、“)”分配压缩代码B1、压缩代码B2。

[0042] 信息处理装置通过执行上述处理，生成对示性式10b进行编码后的压缩代码排列10c。信息处理装置也对未图示的化合物的剩余的部分执行上述处理，从而对化合物进行编码。另外，通过将丁基、sec-丁基等相似的基团汇集，并将各基团汇集地埋入至庞加莱空间，来生成基团的向量。

[0043] 信息处理装置还着眼于分支，以基团的一次结构的单位对压缩代码排列10c进行编码。信息处理装置通过将构成基团的一次结构的多个基团的向量相加，计算化合物的基团的一次结构的向量。

[0044] 图2是用于说明本实施例所涉及的信息处理装置对化合物进行编码的处理的图(2)。在图2中，对对某一化合物所包含的异丁基进行编码的情况进行说明。异丁基的化学结构式11a从“CH”分支为“CH_2”、两个“CH_3”。以“-$CH_2CH(CH_3)_2$”那样的排列表示异丁基的化学结构式11a的示性式11b。

[0045] 信息处理装置通过对示性式11b执行词素解析，将示性式11b所包含的排列分割为单位字符串和分支符号。在图2所示的例子中，信息处理装置将示性式11b分割为“CH_2”、“CH”、“(”、“CH_3”以及“$)_2$”。“CH_2”、“CH”、“CH_3”是单位字符串。“(”、“$)_2$”是分支符号。

[0046] 信息处理装置基于表示化合物的示性式的排列的单位字符串与压缩代码之间的关系的基团词典，对各单位字符串分配压缩代码。例如，信息处理装置分别对“CH_2”、“CH”、“CH_3”分配压缩代码A4、压缩代码A1、压缩代码A2。

[0047] 信息处理装置基于表示分支符号与压缩代码之间的关系的分支代码表格，对各分支符号分配压缩代码。例如，信息处理装置对“(”、“$)_2$”分配压缩代码B1、压缩代码B3。

[0048] 信息处理装置通过执行上述处理，生成对示性式11b进行编码后的压缩代码排列11c。信息处理装置也对未图示的化合物的剩余的部分执行上述处理，从而对化合物进行编码。另外，通过将相似的基团汇集，并将各基团汇集地埋入至庞加莱空间，来生成基团的向量。

[0049] 信息处理装置还以基团的一次结构的单位对压缩代码排列11c进行编码。信息处理装置通过将构成基团的一次结构的多个基团的向量相加，计算化合物的一次结构的向量。

[0050] 图3是用于说明本实施例所涉及的信息处理装置对化合物进行编码的处理的图(3)。在图3中，对对某一化合物所包含的tert-丁基进行编码的情况进行说明。tert-丁基的化学结构式12a从“C”分支为三个“CH_3”。以“-$C(CH_3)_3$”那样的排列表示tert-丁基的化学结构式12a的示性式12b。

[0051] 信息处理装置通过对示性式12b执行词素解析，将示性式12b所包含的排列分割为单位字符串和分支符号。在图3所示的例子中，信息处理装置将示性式12b分割为“C”、“(”、“CH_3”以及“$)_3$”。“CH_2”、“C”、“CH_3”是单位字符串。“(”、“$)_3$”是分支符号。

[0052] 信息处理装置基于表示化合物的示性式的排列与压缩代码之间的关系的基团词

典，对各单位字符串分配压缩代码。例如，信息处理装置分别对“C”、“CH_3”分配压缩代码A5、压缩代码A2。

[0053] 信息处理装置基于表示分支符号与压缩代码之间的关系的分支代码表格，对各分支符号分配压缩代码。例如，信息处理装置对“(”、“)$_3$”分配压缩代码B1、压缩代码B4。

[0054] 信息处理装置通过执行上述处理，生成对示性式12b进行编码后的压缩代码排列12c。信息处理装置也对未图示的化合物的剩余的部分执行上述处理，从而对化合物进行编码。另外，通过将相似的基团汇集，并将各基团汇集地埋入至庞加莱空间，来生成基团的向量。

[0055] 信息处理装置还以基团的一次结构的单位对压缩代码排列12c进行编码。信息处理装置通过将构成基团的一次结构的多个基团的向量相加，计算化合物的一次结构的向量。

[0056] 如上述那样，本实施例所涉及的信息处理装置将表示化合物的示性式的排列分割为单位字符串以及分支符号，并基于基团词典，对化合物的示性式进行编码。另外，通过将相似的基团汇集，并将各基团汇集地埋入至庞加莱空间，来生成基团的向量。信息处理装置通过以基团的一次结构的单位对编码后的化合物的示性式进行编码，来进一步进行编码，并且通过将构成一次结构的多个基团的向量相加，来计算化合物的一次结构的向量。由此，能够精度良好地计算化合物的基团和基团的一次结构的向量，通过使用这样的基团和基团的一次结构的向量执行高分子化合物的相似度评价，相似度评价的精度提高。此外，示性式是编码对象的化学结构式的一个例子。也可以是利用SMILES记法等表现的化学结构式。

[0057] 接下来，对本实施例所涉及的信息处理装置的构成的一个例子进行说明。图4是表示本实施例所涉及的信息处理装置的构成的功能框图。如图4所示，该信息处理装置100具有通信部110、输入部120、显示部130、存储部140以及控制部150。

[0058] 通信部110利用有线或者无线与外部装置等连接，与外部装置等之间进行信息的收发。例如，通信部110由NIC(Network Interface Card:网卡)等实现。通信部110也可以与未图示的网络连接。

[0059] 输入部120是将各种信息输入到信息处理装置100的输入装置。输入部120与键盘、鼠标、触摸面板等对应。

[0060] 显示部130是显示从控制部150输出的信息的显示装置。显示部130与液晶显示器、有机EL(Electro Luminescence:电致发光)显示器、触摸面板等对应。

[0061] 存储部140具有化学结构式文件140a、分支代码表格140b、基团词典140c、基团HMM140d、化学结构式压缩文件140e以及第一转置索引140f。存储部140具有基团一次结构词典140g、基团一次结构HMM140h、基团一次结构压缩文件140i、第二转置索引140j、性质管理表格140k以及迁移表格140n。存储部140例如由RAM(Random Access Memory:随机存取存储器)、闪存(Flash Memory)等半导体存储器元件，或者硬盘、光盘等存储装置实现。

[0062] 化学结构式文件140a是包含高分子化合物的示性式(化学结构式)的信息的文件。图5是表示化学结构式文件的数据结构的一个例子的图。如图5所示，该化学结构式文件140a将化合物识别信息与示性式建立对应关系。化合物识别信息是唯一地识别高分子化合物的信息。示性式是表示高分子化合物的示性式(化学结构式)的排列的信息。例如，高分子化合物与淀粉、糖原、纤维素、支链淀粉等对应，但并不限定于此，也可以是其它的高分子化

合物。

[0063] 分支代码表格140b是定义分配给分支符号的代码的表格。图6是表示分支代码表格的数据结构的一个例子的图。如图6所示，该分支代码表格140b将分支符号与压缩代码建立对应关系。分支符号是表示高分子化合物的分支部分的符号。压缩代码是分配给分支符号的压缩代码。压缩代码的“h”表示16进制。

[0064] 基团词典140c是表示化合物的示性式的排列与压缩代码之间的关系的词典信息。图7是表示基团词典的数据结构的一个例子的图。如图7所示，该基团词典140c将压缩代码、名称、组、示性式建立对应关系。压缩代码表示分配给相应的基团(或者官能团)的示性式的压缩代码。名称是相应的名称的一个例子。组表示相应的基团所属的组。名称例是相应的基团的名称的一个例子。示性式表示成为相应的基团的示性式的排列。

[0065] 另外，虽然在图7中省略图示，但基团词典140c保持在图1～图3等中进行了说明的定义了单位字符串与压缩代码之间的关系的信息。

[0066] 基团HMM(Hidden Markov Model:隐马尔可夫模型)140d是在高分子化合物的排列中，将某一基团与共现的可能性较高的其它的基团建立对应关系的信息。在以下的说明中，适当地将某一基团表述为“第一基团”，将与该第一基团共现的其它的基团表述为“第二基团”。

[0067] 图8是表示基团HMM的数据结构的一个例子的图。如图8所示，该基团HMM140d将压缩代码、名称、组以及基团的共现率建立对应关系。压缩代码表示分配给第一基团的示性式的压缩代码。名称是相应的第一基团的名称的一个例子。组表示相应的第一基团所属的组。

[0068] 在基团的共现率中，对与第一基团共现的第二基团，示出第二基团的压缩代码和共现率。例如，若对第一行进行说明，则在高分子化合物中，与第一基团(压缩代码:8000h)共现的第二基团为压缩代码“8028h”的基团、压缩代码“8132h”的基团、以及压缩代码“80F5h”的基团等。另外，第一基团“压缩代码:8000h”与第二基团“压缩代码:8028h”的共现率为“78％”。第一基团“压缩代码:8000h”与第二基团“压缩代码:8132h”的共现率为“63％”。第一基团“压缩代码:8000h”与第二基团“压缩代码:80F5h”的共现率为“33％”。

[0069] 化学结构式压缩文件140e是包含编码后的高分子化合物的示性式(化学结构式)的信息的文件。图9是表示化学结构式压缩文件的数据结构的一个例子的图。如图9所示，该化学结构式压缩文件140e将化合物识别信息与压缩代码排列建立对应关系。化合物识别信息是唯一地识别高分子化合物的信息。压缩代码排列是以单位字符串的单位对示性式进行编码后的信息。

[0070] 第一转置索引140f是表示化学结构式压缩文件140e所包含的单位字符串或者分支符号的压缩代码与该压缩代码的出现位置之间的关系的索引信息。图10是表示第一转置索引的数据结构的一个例子的图。如图10所示，第一转置索引145在横轴取偏移，在纵轴取压缩代码。偏移表示从化学结构式压缩文件140e的前端的压缩代码到相应的压缩代码为止的出现位置。将前端的压缩代码的偏移设为“0”。

[0071] 第一转置索引140f也可以对图9所示的每个化合物识别信息的压缩代码排列定义压缩代码与偏移之间的关系。例如，也可以分别定义化合物识别信息“C101”的压缩代码排列的转置索引、化合物识别信息“C102”的压缩代码排列的转置索引。

[0072] 基团一次结构词典140g是表示高分子化合物的示性式的压缩代码排列(连续的多

个压缩代码)与一个基团一次结构的压缩代码之间的关系的词典信息。基团一次结构表示组合了多个基团的结构。

[0073] 图11是表示基团一次结构词典的数据结构的一个例子的图。如图11所示,该基团一次结构词典140g将压缩代码、组、名称以及压缩代码排列建立对应关系。压缩代码是与基团一次结构对应的压缩代码。组表示相应的基团一次结构所属的组。名称是基团一次结构的名称的一个例子。压缩代码排列是与基团一次结构所包含的多个基团对应的压缩代码的排列,基团的压缩代码与在基团词典140c中定义的基团的压缩代码对应。

[0074] 基团一次结构HMM140h是在高分子化合物的排列中,将某一基团一次结构与共现的可能性较高的其它的基团一次结构建立对应关系的信息。在以下的说明中,适当地将某一基团一次结构表述为“第一基团一次结构”,将与该第一基团一次结构共现的其它的基团一次结构表述为“第二基团一次结构”。

[0075] 图12是表示基团一次结构HMM的数据结构的一个例子的图。如图12所示,该基团一次结构HMM140h将压缩代码、组、名称以及基团一次结构的共现率建立对应关系。压缩代码表示分配给第一基团一次结构的压缩代码。组表示相应的第一基团一次结构所属的组。名称是相应的第基团一次结构的名称的一个例子。

[0076] 基团一次结构的共现率对与第一基团一次结构共现的第二基团一次结构,示出第二基团一次结构的压缩代码和共现率。例如,若对第一行进行说明,则在高分子化合物中,与第一基团一次结构(压缩代码:F00000h)共现的第二基团一次结构的各压缩代码为“F00011h”、“F00101h”、“F00225h”。

[0077] 另外,第一基团一次结构“压缩代码:F00000h”与第二基团“压缩代码:F00011h”的共现率为“78%”。第一基团一次结构“压缩代码:F00000h”与第二基团“压缩代码:F00101h”的共现率为“63%”。第一基团一次结构“压缩代码:F00000h”与第二基团“压缩代码:F00225h”的共现率为“51%”。

[0078] 基团一次结构压缩文件140i是包含以基团一次结构的单位进行编码后的高分子化合物的示性式的信息的文件。图13是表示基团一次结构压缩文件的数据结构的一个例子的图。如图13所示,该基团一次结构压缩文件140i将化学物识别信息与基团一次结构的压缩代码排列建立对应关系。化合物识别信息是唯一地识别高分子化合物的信息。压缩代码排列是以基团一次结构的单位对示性式进行编码后的信息。

[0079] 第二转置索引140j是表示基团一次结构压缩文件140i所包含的基团一次结构的压缩代码与该压缩代码的出现位置之间的关系的索引信息。图14是表示第二转置索引的数据结构的一个例子的图。如图14所示,第二转置索引140j在横轴取偏移,在纵轴取压缩代码。偏移表示从基团一次结构压缩文件140i的前端的压缩代码(基团一次结构的单位的压缩代码)到相应的压缩代码为止的出现位置。将前端的压缩代码的偏移设为“0”。

[0080] 第二转置索引140j也可以对图13所示的每个化合物识别信息的压缩代码排列定义压缩代码与偏移之间的关系。例如,也可以分别定义化合物识别信息“C101”的压缩代码排列的转置索引、化合物识别信息“C102”的压缩代码排列的转置索引。

[0081] 性质管理表格140k是定义性质相似的基团的表格。图15是表示性质管理表格的数据结构的一个例子的图。如图15所示,该性质管理表格140k将性质编号、压缩代码以及名称建立对应关系。性质编号是识别基团一次结构的性质的编号,对具有相似的性质的基团的

各压缩代码分配同一性质编号。压缩代码是基团的压缩代码。名称是基团的名称。

[0082] 在图15所示的例子中，对压缩代码“8004h”、压缩代码“8005h”、压缩代码“8006h”、压缩代码“8007h”分配同一性质代码“I102”。因此，意味着与压缩代码“8004h”、压缩代码“8005h”、压缩代码“8006h”、压缩代码“8007h”对应的各基团具有相似的性质。

[0083] 虽然在图15中省略说明，但该性质管理表格140k也可以保持将性质编号与具有相似的性质的基团的压缩代码建立了对应关系的信息。

[0084] 基团向量表格1401是保持对高分子化合物所包含的各基团分配的压缩代码的向量的表格。图16A是表示基团向量表格的数据结构的一个例子的图。如图16A所示，基团向量表格1401将基团的压缩代码与向量建立对应关系。此外，基团向量表格1401也可以保持与在分支代码表格140b中定义的分支符号对应的向量。

[0085] 基团一次结构向量表格140m是保持对高分子化合物所包含的各基团一次结构分配的压缩代码的向量的表格。图16B是表示基团一次结构向量表格的数据结构的一个例子的图。如图16B所示，基团一次结构向量表格140m将基团一次结构的压缩代码与向量建立对应关系。

[0086] 迁移表格140n是保持高分子化合物的向量的信息的表格。图16C是表示迁移表格的数据结构的一个例子的图。如图16C所示，迁移表格140n具有化合物识别信息和多个向量。化合物识别信息是唯一地识别高分子化合物的信息。多个向量表示高分子化合物所包含的各基团一次结构的向量。例如，向量(n)是从高分子化合物的前端起位于第n个的基团一次结构的向量。

[0087] 返回到图4的说明。控制部150具有获取部151、第一编码部152、第二编码部153、向量计算部154以及相似度评价部155。控制部150例如由CPU(Central Processing Unit:中央处理器)或者MPU(Micro Processing Unit:微处理器)实现。另外，例如也可以通过ASIC(Application specific Integrated Circuit:专用集成电路)、FPGA(Field Programmable gate Array:现场可编程门阵列)等集成电路执行控制部150。

[0088] 获取部151是经由网络从外部装置等获取各种信息的处理部。例如，获取部151获取化学结构式文件140a，并将获取到的化学结构式文件140a储存于存储部140。获取部151也可以获取基团词典140c、基团HMM140d、基团一次结构词典140g、基团一次结构HMM140h并储存于存储部140。

[0089] 第一编码部152是从化学结构式文件140a获取高分子化合物的示性式并对获取到的示性式进行编码的处理部。以下，对第一编码部152的处理的一个例子进行说明。

[0090] 第一编码部152通过对高分子化合物的示性式执行词素解析，将示性式所包含的排列分割为单位字符串或者分支符号。第一编码部152执行词素解析并分割为单位字符串或者分支符号的处理与在图1、图2、图3中进行了说明的处理对应。

[0091] 第一编码部152在成为分割候补的单位字符串存在多个的情况下，也可以基于基团HMM140d，选择共现率较高一方的单位字符串进行分割。

[0092] 第一编码部152在将高分子化合物的示性式分割为单位字符串或者分支符号之后，使用分支代码表格140b以及基团词典140c，对示性式进行编码，从而生成压缩代码排列。第一编码部152生成的压缩代码排列与“第一代码排列”对应。第一编码部152将化合物识别信息与压缩代码排列建立对应关系，并登记于化学结构式压缩文件140e。

[0093] 第一编码部152在对高分子化合物的示性式的单位字符串或者分支符号分配压缩代码的情况下，将压缩代码的种类与偏移之间的关系登记于第一转置索引140f。例如，在高分子化合物的示性式的压缩代码排列(第一代码排列)中，压缩代码“8000h”的偏移为“1”的情况下，第一编码部152在第一转置索引140f的偏移“1”的列与压缩代码“8000h”的行交叉的部分设定“1”。

[0094] 第一编码部152通过从化学结构式文件140a获取与未选择的化合物识别信息对应的示性式的信息，并反复执行上述处理，来生成化学结构式压缩文件140e。

[0095] 第二编码部153是从化学结构式压缩文件140e获取高分子化合物的压缩代码排列(第一代码排列)，并以基团一次结构的单位对获取到的压缩代码排列进行编码的处理部。在基团一次结构词典140g中定义高分子化合物的压缩代码排列与基团一次结构的压缩代码之间的关系。

[0096] 图17是用于说明第二编码部的处理的图。在图17中，压缩代码排列15是登记于化学结构式压缩文件140e的与化合物识别信息“C101”对应的压缩代码排列15a。压缩代码排列15a所包含的各压缩代码是在基团词典140c中定义的单位代码的压缩代码或者分支符号的压缩代码。

[0097] 例如，假设在基团一次结构词典140g中，压缩代码排列15a的各压缩代码X1、X2、…、X19与基团一次结构的压缩代码“F00112h”建立了对应关系。该情况下，第二编码部153对各压缩代码X1、X2、…、X19分配基团一次结构的压缩代码“F00112h”。

[0098] 假设在基团一次结构词典140g中，压缩代码排列15a的各压缩代码Y1、Y2、…、X38与基团一次结构的压缩代码“F00332h”建立了对应关系。该情况下，第二编码部153对各压缩代码Y1、Y2、…、Y38分配基团一次结构的压缩代码“F00332h”。

[0099] 假设在基团一次结构词典140g中，压缩代码排列15a的各压缩代码Z1、Z2、…、Z20与基团一次结构的压缩代码“F00110h”建立了对应关系。该情况下，第二编码部153对各压缩代码Z1、Z2、…、Z20分配基团一次结构的压缩代码“F00110h”。

[0100] 第二编码部153通过基于基团一次结构词典140g反复执行上述处理，生成以基团一次结构的单位对压缩代码排列15a进行编码后的压缩代码排列15b。压缩代码排列15b与“第二代码排列”对应。第二编码部153将化合物识别信息与压缩代码排列建立对应关系，并登记于基团一次结构压缩文件140i。

[0101] 第二编码部153在对压缩代码排列15a与基团一次结构词典140g进行比较来进行编码的过程中，在基团一次结构的单位的多个压缩代码成为编码的候补的情况下，也可以基于基团一次结构HMM140h，选择共现率较高一方的压缩代码进行分配。

[0102] 第二编码部153在以基团一次结构的单位进行编码的情况下，将压缩代码的种类与偏移之间的关系登记于第二转置索引140j。例如，在高分子化合物的示性式的压缩代码排列(第二代码排列)中，压缩代码“F00112h”的偏移为“0”。在该情况下，第二编码部153在第一转置索引140f的偏移“0”的列与压缩代码“F00112h”的行交叉的部分设定“1”。

[0103] 第二编码部153通过从化学结构式压缩文件140e获取与未选择的化合物识别信息对应的示性式的压缩代码排列，并反复执行上述处理，来生成基团一次结构压缩文件140i。

[0104] 另外，第二编码部153将基团一次结构的压缩代码与该基团一次结构的压缩代码所对应的多个基团的压缩代码、分支符号的压缩代码之间的关系输出到向量计算部154。

[0105] 返回到图4的说明。向量计算部154是从基团一次结构压缩文件140i获取高分子化合物的压缩代码排列(第二代码排列),并计算高分子化合物的向量的处理部。向量计算部154从基团一次结构压缩文件140i获取基团一次结构的压缩代码,并通过执行以下的处理来计算获取到的压缩代码的向量。向量计算部154从上述的第二编码部153获取基团一次结构的压缩代码与多个基团的压缩代码以及分支符号的压缩代码之间的关系。

[0106] 向量计算部154通过将每个基团的压缩代码埋入至庞加莱空间等向量空间,分别计算与压缩代码对应的向量。向量计算部154对于分支符号的压缩代码,既可以也通过埋入至庞加莱空间来计算向量,也可以预先分配向量。向量计算部154将基团的压缩代码与向量建立对应关系,并登记于基团向量表格140l。

[0107] 向量计算部154通过将分配给基团一次结构所包含的各基团的压缩代码、分支符号的压缩代码的向量相加,计算基团一次结构的向量。向量计算部154将基团一次结构的压缩代码与向量建立对应关系,并登记于基团一次结构向量表格140m。

[0108] 在高分子化合物包含有多个基团一次结构。向量计算部154将化合物识别信息与每个基团一次结构的向量建立对应关系,并登记于迁移表格140n。向量计算部154按照基团一次结构出现的顺序,将基团一次结构的向量登记于迁移表格140n。

[0109] 这里,向量计算部154将基团的压缩代码埋入至庞加莱空间来计算向量的处理是被称为庞加莱嵌入(Poincare Embeddings)的技术。庞加莱嵌入例如使用非专利文献“Valentin Khrulkovl et al.“Hyperbolic Image Embeddings”Cornell University,2019 April 3”等所记载的技术即可。

[0110] 在庞加莱嵌入中,根据埋入至庞加莱空间的位置分配向量,另外,有越为相似的信息,埋入至越接近的位置这样的特征。向量计算部154基于性质管理表格140k,确定具有相似的性质的各基团的压缩代码。

[0111] 图18是表示庞加莱空间的一个例子的图。如图15所说明的那样,与压缩代码“8004h”、压缩代码“8005h”、压缩代码“8006h”、压缩代码“8007h”对应的基团具有相似的性质。因此,向量计算部154将压缩代码“8004h”、压缩代码“8005h”、压缩代码“8006h”、压缩代码“8007h”埋入至庞加莱空间P上的相互接近的位置,并附加与位置对应的向量。

[0112] 向量计算部154从前端开始依次计算高分子化合物的压缩代码排列所包含的各基团一次结构的压缩代码的向量,并依次登记于迁移表格140n。向量计算部154通过从基团一次结构压缩文件140i获取与未选择的化合物识别信息对应的示性式的压缩代码排列,并反复执行上述处理,来生成迁移表格140n。

[0113] 相似度评价部155是通过对登记于迁移表格140n的各化合物识别信息所对应的向量进行比较来评价高分子化合物的相似度的处理部。例如,相似度评价部155计算各高分子化合物的向量的距离,将距离小于阈值的高分子化合物的组确定为彼此相似的高分子化合物。

[0114] 例如,相似度评价部155也可以将集成了登记于迁移表格140n的高分子化合物所对应的各基团一次结构的向量后的向量作为高分子化合物的向量,并计算各高分子化合物的向量的距离。相似度评价部155既可以使评价结果输出到显示部130并显示,也可以通知给外部装置等。

[0115] 接着,对本实施例所涉及的信息处理装置100的处理顺序的一个例子进行说明。图

19是表示本实施例所涉及的信息处理装置的处理顺序的流程图。如图19所示，信息处理装置100的第一编码部152从化学结构式文件140a获取高分子化合物的示性式的排列(步骤S101)。

[0116] 第一编码部152执行词素解析，将示性式的排列分割为单位字符串或者分支符号(步骤S102)。第一编码部152基于分支代码表格140b、基团词典140c、基团HMM140d，对单位字符串、分支符号分配压缩代码，生成第一转置索引140f、化学结构式压缩文件140e。另外，信息处理装置100的向量计算部154通过埋入至庞加莱空间来生成基团的向量(步骤S103)。

[0117] 信息处理装置100的第二编码部153从化学结构式压缩文件140e获取高分子化合物的压缩代码排列(步骤S104)。第二编码部153基于基团一次结构词典140g、基团一次结构HMM140h，以基团一次结构单位生成压缩编码排列，并生成第二转置索引140j、基团一次结构压缩文件140i(步骤S105)。

[0118] 信息处理装置100的向量计算部154通过将构成储存于基团一次结构压缩文件140i的基团一次结构单位的基团的向量相加，计算基团一次结构的向量，并生成迁移表格140n(步骤S106)。信息处理装置100的相似度评价部155基于迁移表格140n，执行相似度评价(步骤S107)。

[0119] 接下来，对本实施例所涉及的信息处理装置100的效果进行说明。信息处理装置100通过将高分子化合物的示性式分割为单位字符串、分支符号并分配压缩代码，来生成压缩代码排列，并且，转换为每个基团一次结构的各压缩代码。信息处理装置通过基于每个基团的压缩代码，将成为相似的性质的基团汇集并埋入至庞加莱空间，来生成用于适当地表现化合物的性质的基团的向量。通过使用通过这样的信息处理装置100的处理生成的向量执行相似度的评价，能够使相似度评价的精度提高。

[0120] 信息处理装置100使用构成基团一次结构的压缩代码，将各基团的向量相加，计算基团的一次结构的向量，并生成迁移表格。因此，能够对性质相似的各基团一次结构分配高精度的向量。

[0121] 信息处理装置100通过对储存于迁移表格140n的各高分子化合物所对应的基团一次结构的向量进行比较，能够精度良好地评价具有相似的特性的高分子化合物。

[0122] 然而，上述的本实施例所涉及的信息处理装置100的处理是一个例子，也可以执行其它的处理。如图1～图3所说明的那样，信息处理装置100在对高分子化合物的示性式进行编码的情况下，按照基团所包含的单位字符串、分支符号的粒度分配压缩代码，但并不限定于此。

[0123] 信息处理装置100也可以将基团视为最小单位，对位于分支部分的基团分配分支的基团固有的压缩代码，来进行编码。图20是用于说明本实施例所涉及的信息处理装置的其它的处理的图。在图20中，作为高分子化合物的一个例子，使用糖原。

[0124] 连接多个基团构成图20所示的高分子化合物30。着眼于高分子化合物30的区域30a进行说明。在区域30a包含有多个基团31和分支部分的基团32。信息处理装置100的第一编码部152基于基团词典140c，对多个基团31分配压缩代码。另外，第一编码部152对分支部分的基团32分配分支部分的固有的压缩代码。

[0125] 接下来，对实现与上述实施例所示的信息处理装置100相同的功能的计算机的硬件构成的一个例子进行说明。图21是表示实现与实施例的信息处理装置相同的功能的计算

机的硬件构成的一个例子的图。

[0126] 如图21所示，计算机200具有执行各种运算处理的CPU201、接受来自用户的数据的输入的输入装置202、以及显示器203。另外，计算机200具有经由有线或者无线网络与外部装置等之间进行数据的授受的通信装置204和接口装置205。另外，计算机200具有暂时存储各种信息的RAM206和硬盘装置207。而且，各装置201～207与总线208连接。

[0127] 硬盘装置207具有获取程序207a、第一编码程序207b、第二编码程序207c、向量计算程序207d以及相似度评价程序207e。另外，CPU201读出各程序207a～207e并在RAM206展开。

[0128] 获取程序207a作为获取工序206a发挥作用。第一编码程序207b作为第一编码工序206b发挥作用。第二编码程序207c作为第二编码工序206c发挥作用。向量计算程序207d作为向量计算工序206d发挥作用。相似度评价程序207e作为相似度评价工序206e发挥作用。

[0129] 获取工序206a的处理与获取部151的处理对应。第一编码工序206b的处理与第一编码部152的处理对应。第二编码工序206c的处理与第二编码部153的处理对应。向量计算工序206d的处理与向量计算部154的处理对应。相似度评价工序206e的处理与相似度评价部155的处理对应。

[0130] 此外，各程序207a～207e也可以并不一定从最初开始存储于硬盘装置307。例如，也可以使各程序预先存储于插入到计算机200的软盘(FD)、CD-ROM、DVD、光磁盘、IC卡等“便携式物理介质”。而且，计算机200读出各程序207a～207e并执行。

[0131] 附图标记说明：100…信息处理装置，110…通信部，120…输入部，130…显示部，140…存储部，140a…化学结构式文件，140b…分支代码表格，140c…基团词典，140d…基团HMM，140e…化学结构式压缩文件，140f…第一转置索引，140g…基团一次结构词典，140h…基团一次结构HMM，140i…基团一次结构压缩文件，140j…第二转置索引，140k…性质管理表格，140l…基团向量表格，140m…基团一次结构向量表格，140n…迁移表格，150…控制部，151…获取部，152…第一编码部，153…第二编码部，154…向量计算部，155…相似度评价部。

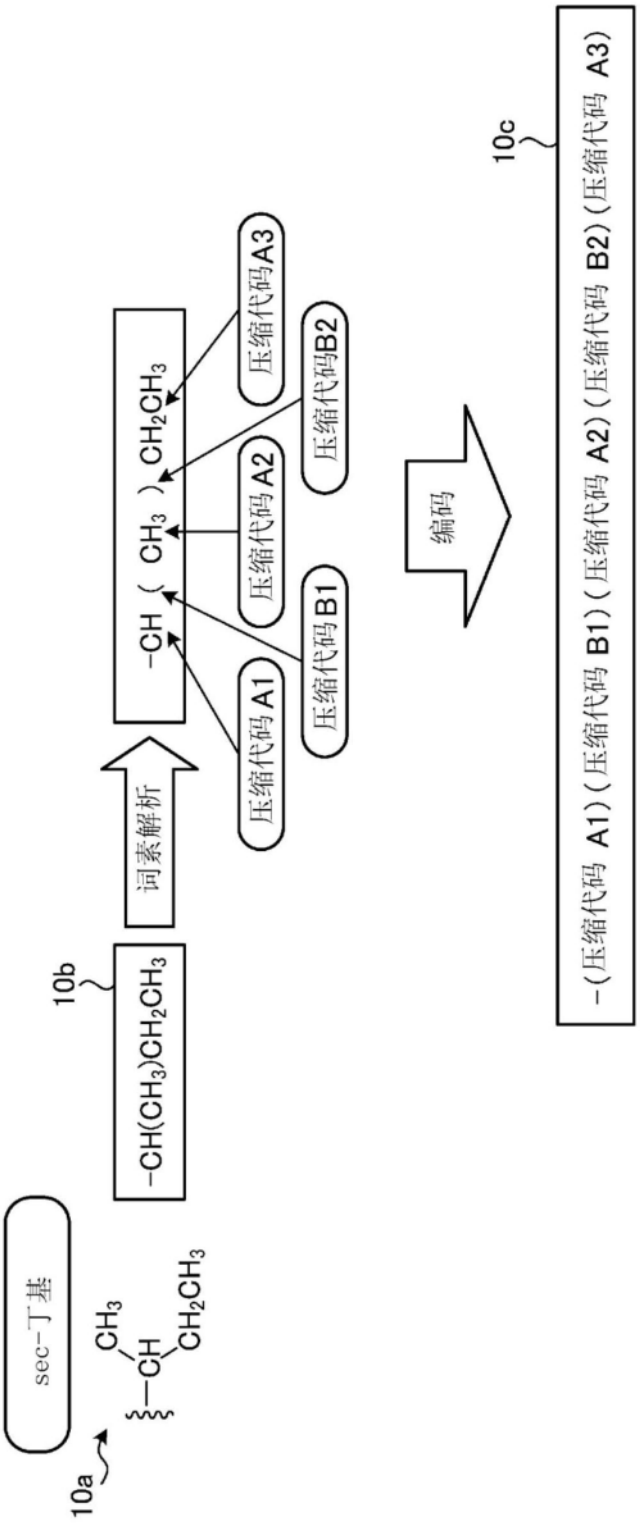

sec-丁基
10a
-CH(CH3)CH2CH3
10b
词素解析
-CH(CH3) CH2CH3
压缩代码A1
压缩代码A2
压缩代码A3
压缩代码B1
压缩代码B2
编码
10c
-(压缩代码 A1)(压缩代码 B1)(压缩代码 A2)(压缩代码 B2)(压缩代码 A3)

图1

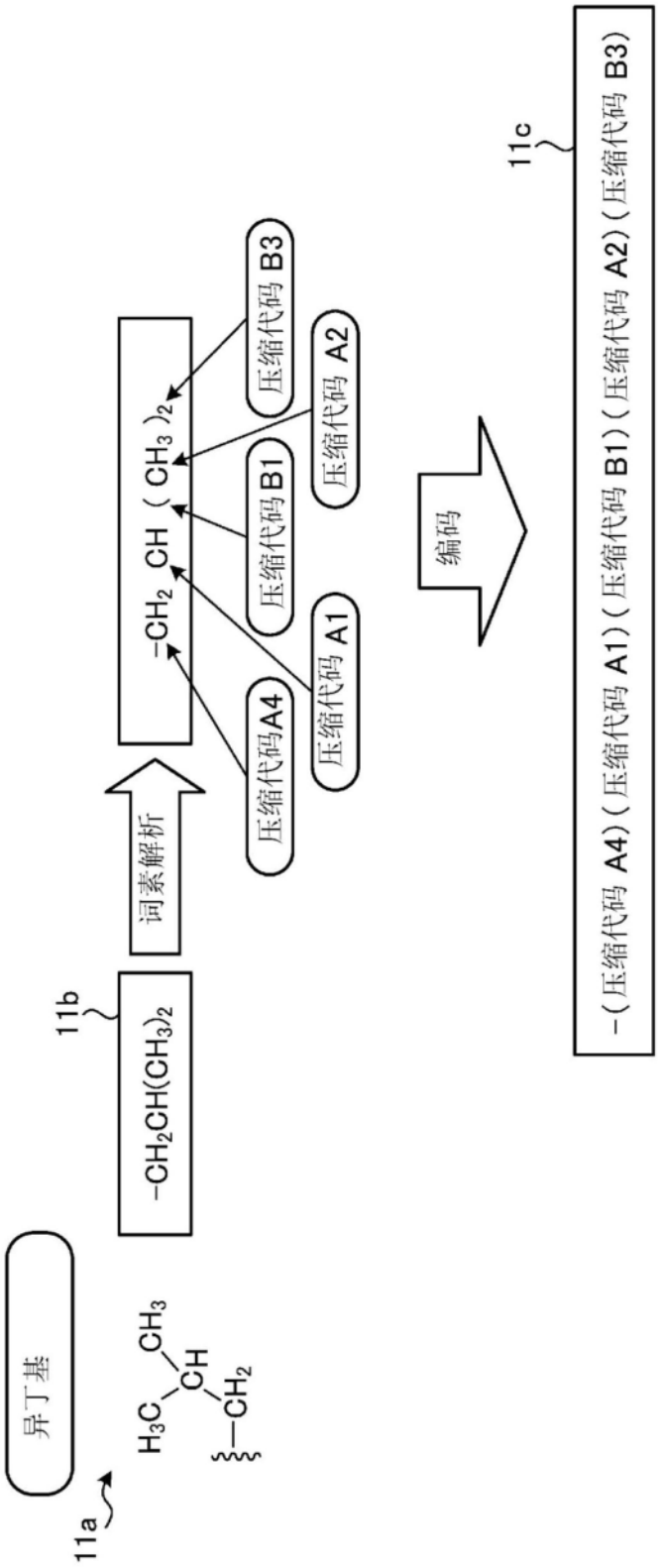

异丁基
11a
H_3C CH_3 CH CH_2
$-CH_2CH(CH_3)_2$
11b
词素解析
$-CH_2$ CH (CH_3)$_2$
压缩代码A4
压缩代码 A1
压缩代码 B1
压缩代码 A2
压缩代码 B3
编码
11c
-(压缩代码 A4)(压缩代码 A1)(压缩代码 B1)(压缩代码 A2)(压缩代码 B3)

图2

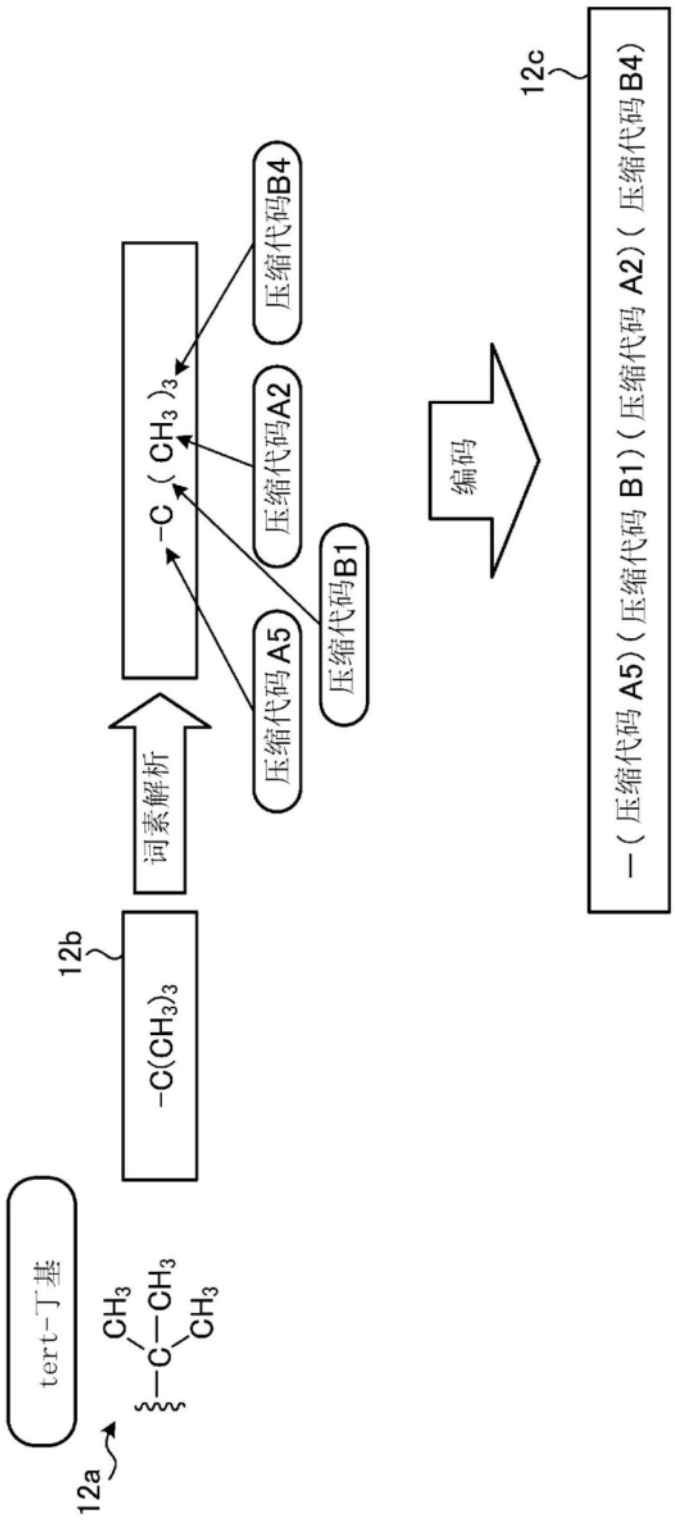

tert-丁基
12a
12b
12c
词素解析
编码
压缩代码A5
压缩代码B1
压缩代码A2
压缩代码B4
—（压缩代码 A5）（压缩代码 B1）（压缩代码 A2）（ 压缩代码B4）

图3

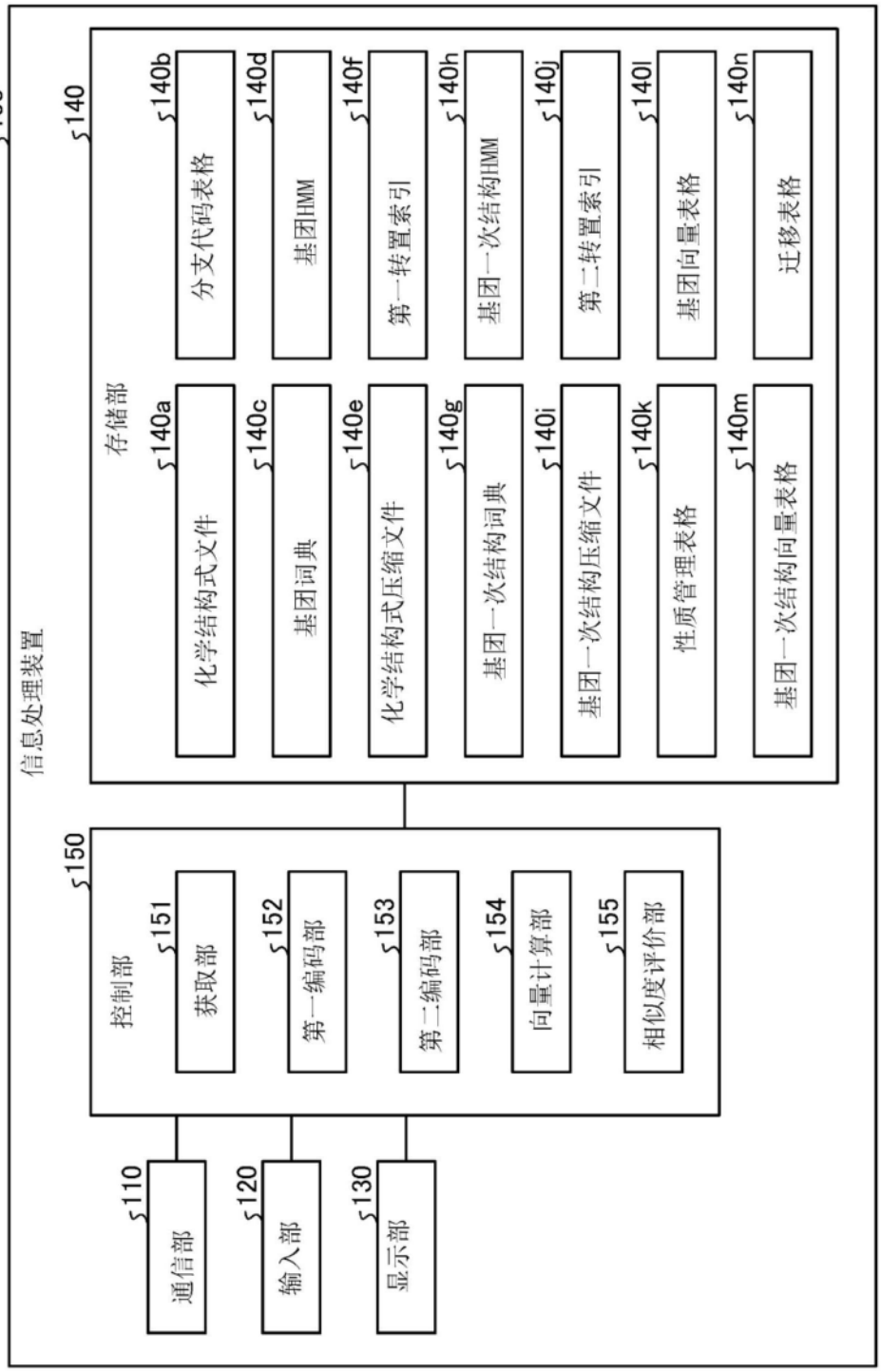

100
信息处理装置
140
存储部
140a
化学结构式文件
140b
分支代码表格
140c
基团词典
140d
基团HMM
140e
化学结构式压缩文件
140f
第一转置索引
140g
基团一次结构词典
140h
基团一次结构HMM
140i
基团一次结构压缩文件
140j
第二转置索引
140k
性质管理表格
140l
基团向量表格
140m
基团一次结构向量表格
140n
迁移表格
150
控制部
151
获取部
152
第一编码部
153
第二编码部
154
向量计算部
155
相似度评价部
110
通信部
120
输入部
130
显示部

图4

140a

化合物识别信息	示性式
C101	表示化合物识别信息“C101”的示性式的代码排列的信息
C102	表示化合物识别信息“C102”的示性式的代码排列的信息
C103	表示化合物识别信息“C103”的示性式的代码排列的信息
···	···

图5

140b

分支符号	压缩代码
(	30h
)	31h
$)_2$	32h
$)_3$	33h

图6

140c

压缩代码	组	名称例	示性式
8000h	甲基	甲烷	CH3
8001h	乙基	乙烷	CH3CH2
8002h	丙基	丙烷	CH3CH2CH2
8003h	丁基	丁烷	CH3CH2CH2CH2
…	…	…	…
8013h	硝基	硝基…	NO2
…	…	…	…
805Ah	羟基	…羟基醇	OH
805Bh	醛基	…醛	OHO
…	…	…	…

图7

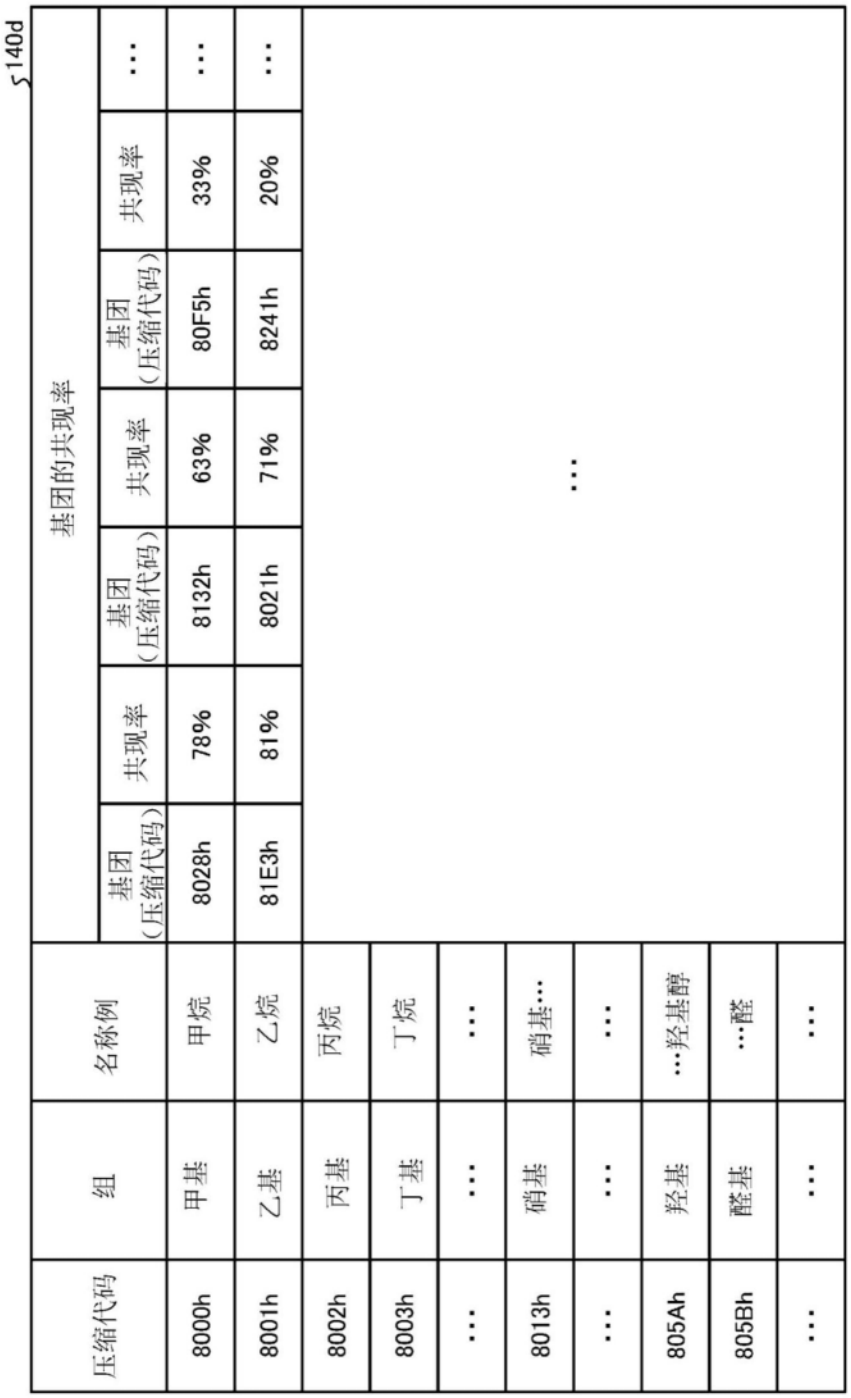

140d
基团的共现率
压缩代码
组
名称例
基团（压缩代码）
共现率
基团（压缩代码）
共现率
基团（压缩代码）
共现率
8000h
甲基
甲烷
8028h
78%
8132h
63%
80F5h
33%
8001h
乙基
乙烷
81E3h
81%
8021h
71%
8241h
20%
8002h
丙基
丙烷
8003h
丁基
丁烷
8013h
硝基
硝基…
805Ah
羟基
…羟基醇
805Bh
醛基
…醛

图8

140e

化合物识别信息	压缩代码排列
C101	对化合物识别信息“C101”的示性式进行了编码的压缩代码排列
C102	对化合物识别信息“C102”的示性式进行了编码的压缩代码排列
C103	对化合物识别信息“C103”的示性式进行了编码的压缩代码排列
…	…

图9

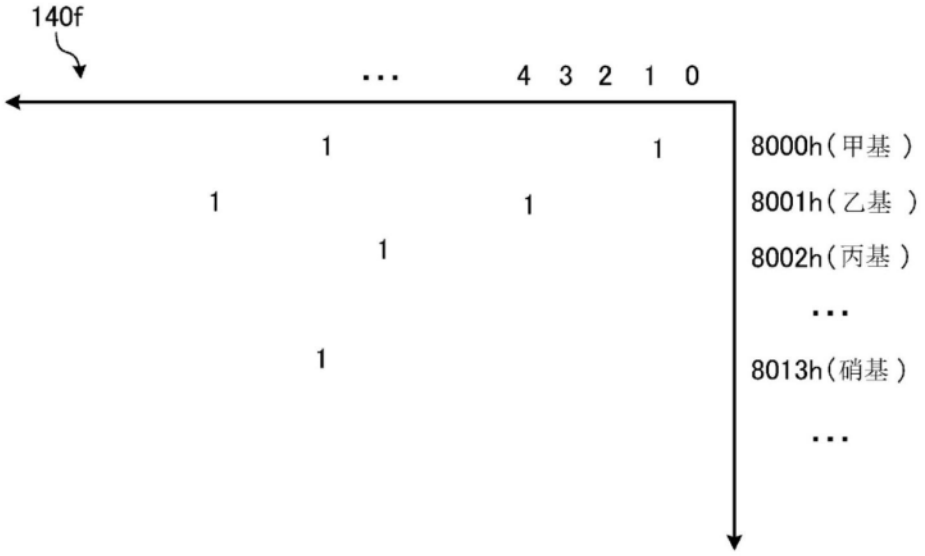


图10

140g

压缩代码	组	名称	压缩代码排列
F0O000h	G1	α一次结构	02h8028h·····03h
F0O001h	G1	β一次结构	02h80F5h····03h
F0O002h	G1	γ一次结构	02h81E3h·····03h
F0O003h	G1	δ一次结构	·····
···	···	···	·····

图11

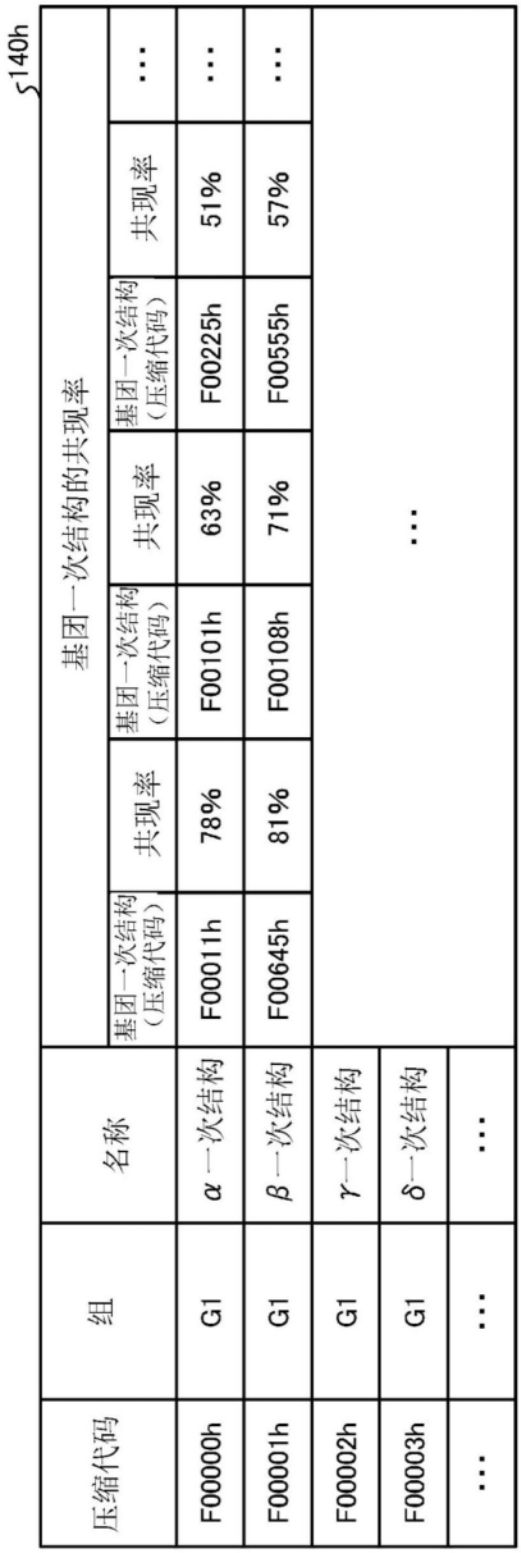

140h
压缩代码
组
名称
基团一次结构的共现率
基团一次结构（压缩代码）
共现率
基团一次结构（压缩代码）
共现率
基团一次结构（压缩代码）
共现率
F00000h
G1
α一次结构
F00011h
78%
F00101h
63%
F00225h
51%
F00001h
G1
β一次结构
F00645h
81%
F00108h
71%
F00555h
57%
F00002h
G1
γ一次结构
F00003h
G1
δ一次结构

图12

140i

化合物识别信息	基团一次结构的压缩代码排列				
C101	F00112h	F00332h	F00110h	F00460h	···
C102	F00232h	F00582h	F00874h	F00209h	···
C103	F00102h	F00082h	F00874h	F00009h	···
···	···				

图13

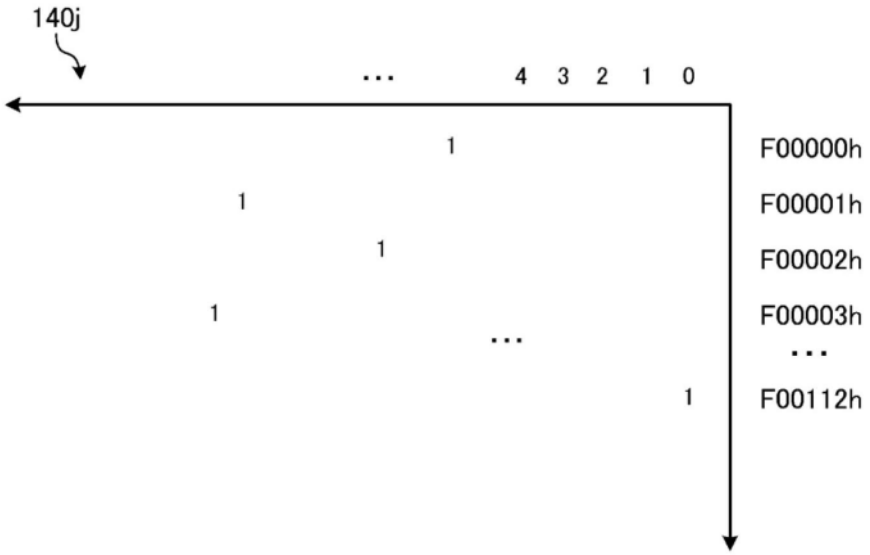


图14

140k

性质编号	压缩代码	名称
I101	8000h	甲烷
	8120h	···
	8234h	···
I102	8004h	乙烷
	8005h	···
	8006h	···
	8007h	···
···	···	···

图15

140l

压缩代码	向量
8000h	Vec1－1
8001h	Vec1－2
8002h	Vec1－3
···	···
8013h	Vec1－3
···	···
805Ah	Vec1－10
805Bh	Vec1－11
···	···

图16A

140m

压缩代码	向量
F00000h	Vec2－1
F00001h	Vec2－2
F00002h	Vec2－3
F00003h	Vec2－4
···	···

图16B

140n

化合物识别信息	向量(1)	向量(2)	向量(3)	···
C101	0.28213	0.14975	0.18975	···
C102	0.97325	0.23800	0.15575	···
C103	0.309085	0.13800	0.14925	···
···	···	···	···	···

图16C

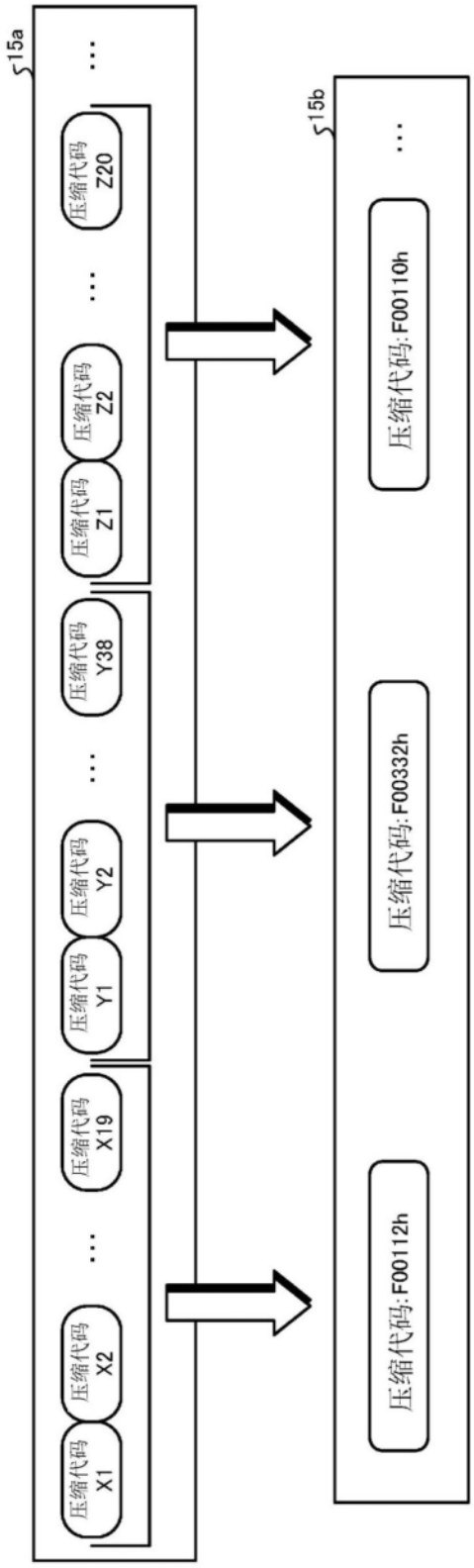

15a
15b
压缩代码 X1
压缩代码 X2
…
压缩代码 X19
压缩代码 Y1
压缩代码 Y2
…
压缩代码 Y38
压缩代码 Z1
压缩代码 Z2
…
压缩代码 Z20
…
压缩代码:F00112h
压缩代码:F00332h
压缩代码:F00110h
…

图17

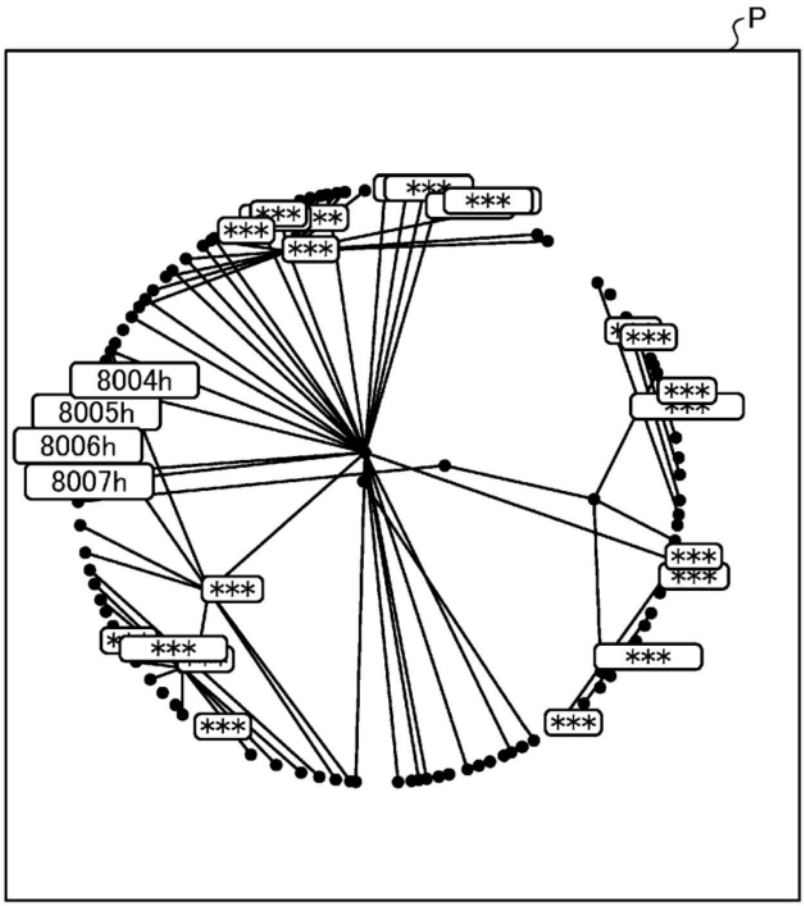


图18

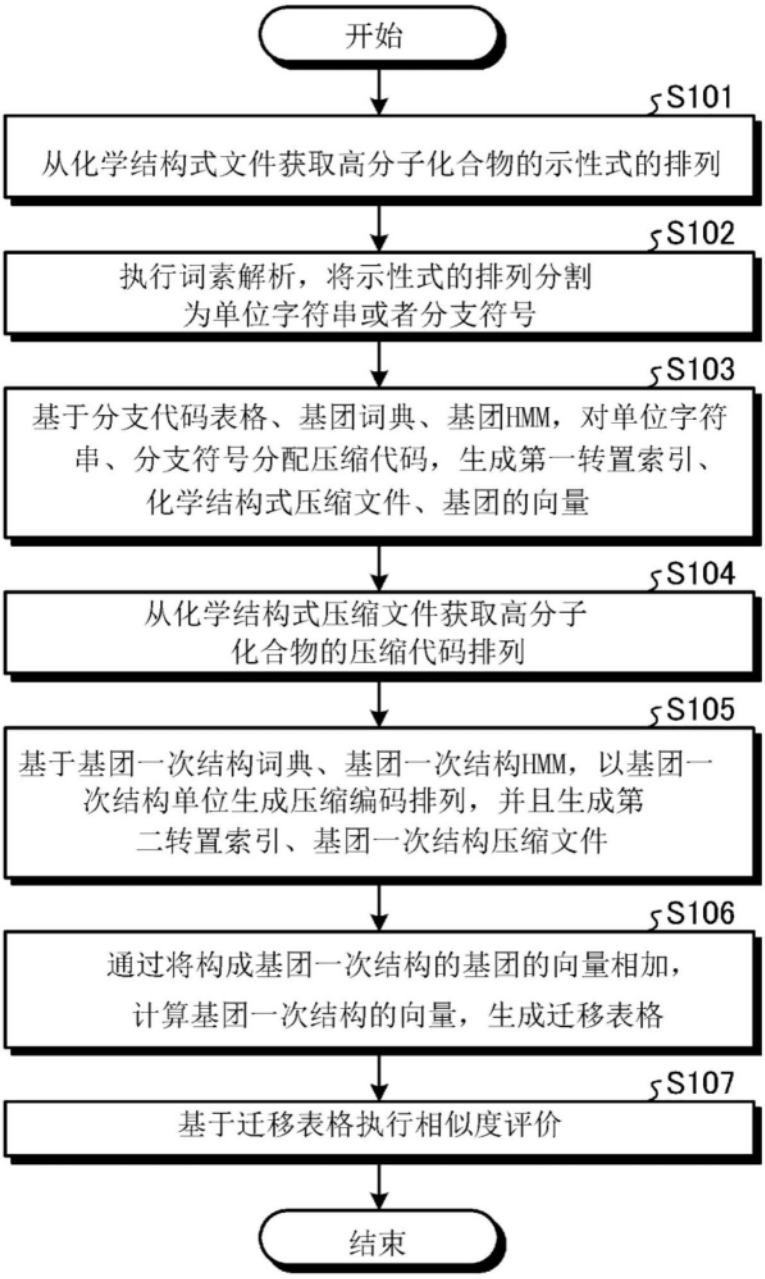

开始
S101
从化学结构式文件获取高分子化合物的示性式的排列
S102
执行词素解析，将示性式的排列分割
为单位字符串或者分支符号
S103
基于分支代码表格、基团词典、基团HMM，对单位字符
串、分支符号分配压缩代码，生成第一转置索引、
化学结构式压缩文件、基团的向量
S104
从化学结构式压缩文件获取高分子
化合物的压缩代码排列
S105
基于基团一次结构词典、基团一次结构HMM，以基团一
次结构单位生成压缩编码排列，并且生成第
二转置索引、基团一次结构压缩文件
S106
通过将构成基团一次结构的基团的向量相加，
计算基团一次结构的向量，生成迁移表格
S107
基于迁移表格执行相似度评价
结束

图19

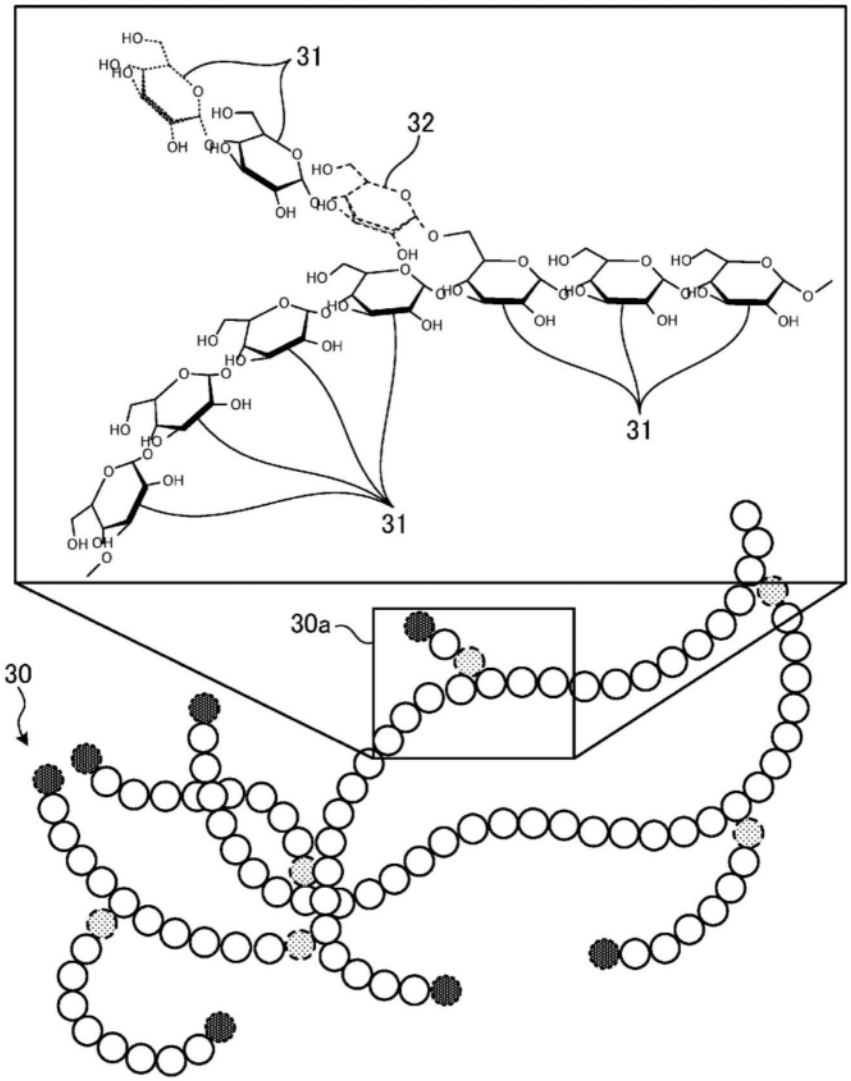


图20

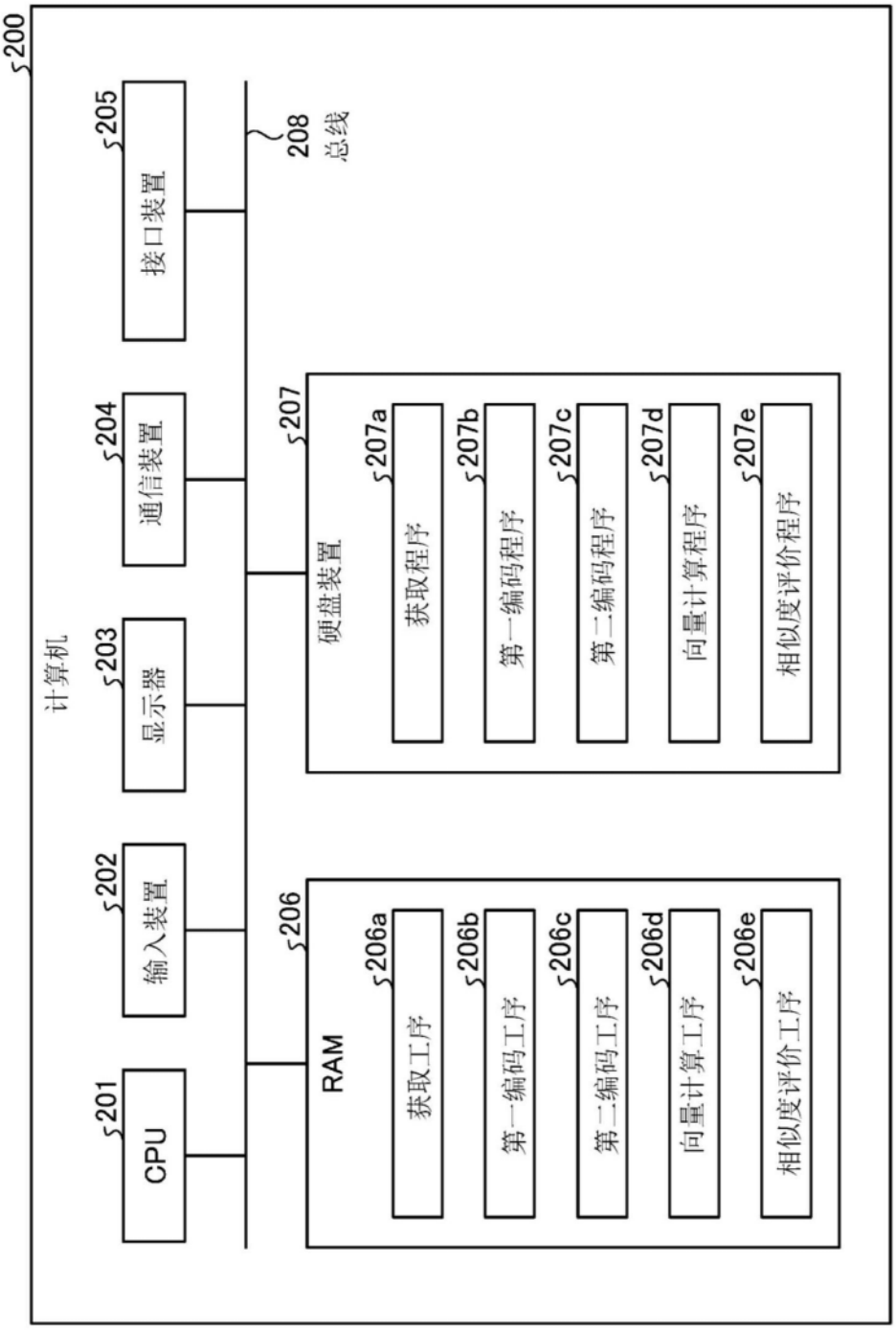

200
计算机
201
CPU
202
输入装置
203
显示器
204
通信装置
205
接口装置
208
总线
206
RAM
206a
获取工序
206b
第一编码工序
206c
第二编码工序
206d
向量计算工序
206e
相似度评价工序
207
硬盘装置
207a
获取程序
207b
第一编码程序
207c
第二编码程序
207d
向量计算程序
207e
相似度评价程序

图21