计算机与应用化学
Computers and Applied Chemistry
研究论文
Vol.35, No.8
August 28, 2018
CODEN JYYHE6/ISSN 1001-4160
DOI: 10.16866/j.com.app.chem201808002

植物化学成分数据库分子式数据处理

徐挺军，陈维明

中国科学院上海有机化学研究所信息中心，上海市，200032

摘要：为解决专业型化学数据库化合物检索和筛选难题，从研究化合物分子式的定义和基本构成入手，总结和归纳植物化学成分数据库中化合物分子式数据的类型及特点，得出的分子式表达式通式，设计算法程序分析和处理分子式数据，并生成分子式特征数据。在植物化学成分数据库中，结合分子式特征数据对化合物进行检索和筛选，实现了化合物的分子式综合检索技术。介绍分子式综合检索的应用实例，以此来丰富植物化学成分数据库的检索和筛选方式。分子式数据处理方法解决了用户在分子式检索时输入与数据库表达不一致的问题，提高了化合物检索和筛选效率。分子式数据处理方法还可以应用于功能材料数据库、药物化学数据库、天然产物数据库等专业型化学数据库的化合物检索和筛选。

关键词：分子式数据处理；分子式特征；化合物检索；药物筛选；化学数据库挖掘

中图分类号：TQ015.9; TP391.9; O6-39　　**文献标识码：**A

文章编号：1001-4160(2018)08-619-624　　**DOI:**10.16866/j.com.app.chem201808002

1　引　言

化合物检索和筛选是化学数据库中最主要的功能，常见检索方式有：化合物名称检索、化合物结构检索、化合物标示信息检索（如CAS Registry Number， SRN等）、化合物分子式检索等。传统的化合物分子式检索是利用数据库查询功能，直接匹配或者是模糊匹配输入分子式字符串与数据库表中分子式数据，检索出与输入分子式相等或相似的结果[1]。由于化学数据库使用者对于化学分子式的输入规范不尽一致，而数据库中的分子式数据根据其来源和生成途径不同，其表达方式也各有不同，这就造成传统的化合物分子式检索方式往往无法比较精确地定位某一个或者某一类化合物。例如，想要在Reaxys®中用分子式检索小檗碱类化合物，使用小檗碱的分子式“C20H18NO4”出发，检索分子式中含有“C20H18NO4”的所有化合物，检索结果有1800多个物质，如果不依靠化

投稿日期：2018-06-07；录用日期：2018-07-27；网络出版日期：2018-08-25
基金项目：中国科学院信息化专项科学大数据工程(XXH135)；上海市化学化工数据共享服务平台(18DZ2294000)
作者简介：徐挺军(1984--)，男，浙江海宁，工程硕士．研究领域：化学数据库．
联 系 人：徐挺军(1984--)，E-mail: xutingjun@sioc.ac.cn

合物结构、名称等进行二次筛选，就无法简单地检索到比较精确的结果集。

对于专业型化学数据库而言，例如天然产物数据库、含能材料数据库、化合物谱图数据库等，由于其数据来源、数据加工方式等原因，化合物的结构信息不完整，想要开发子结构检索来建立化合物筛选机制较为困难。而分子式数据分布广泛，通过数据收集或者是专业化学软件（如ChemOffice®、Accelrys®等）计算都较容易得到。植物化学成分数据库的数据来自研究文献，其中成分化合物名称根据其来源文献不同而表达不一（俗名和系统命名），但基本为有机大分子化合物，想要通过输入化合物系统名称或者画完整的化学结构去检索和筛选往往比较繁琐且无法得到精确的结果[2]。而分子式的输入相对比较简单，且植物化学成分数据库中所有的化合物成分都有其确切的分子式。将植物化学成分数据库中的分子式数据进行预处理，规范和统一分子式数据的表达方式；使用计算机程序分析植物化学成分数据库中分子式数据的特征，并将分析结果数据存储于数据库中，从分子式数据中挖掘更深层次的化合物特征数据信息；数据库前端检索程序结合基础的分子式数据和分子式分析结果特征数据对化合物进行检索和筛选，从而实现化合物分子式更加精确的检索和筛选是可行的。

2 分子式数据处理

首先，分析化合物分子式的定义和基本构成。由分子式的定义来看，化合物分子式是由组成化合物的元素符号与该元素在该化合物中所占比例的数字组成的，数字为正整数，比例为1时通常省略[3]。植物化学成分数据库中的分子式类型有：无机化合物按照元素符号首字母的英文字母顺序排列，比如：水的分子式“H2O”；有机化合物一般碳元素和氢元素在前，其他元素按照元素符号英文字母顺序排列，比如苯甲酸钠的分子式“C7H5NaO2”；其他类型的分子式结构还有晶体或者有机化合物盐使用点号“.”分隔，“.”之前为晶体或者有机化合物盐的阴阳离子，“.”之后的数字为晶体或者是有机化合物盐的阴阳离子比例，比例为1时通常省略，接着是晶体或者有机化合物盐的阴阳离子，比如硼砂的分子式“B4Na2O7.10H2O”；聚合物使用单体化合物的分子式表示，并且使用括号“（）n”，比如聚乙醇酸的分子式“(C3H5NO)n”等……[4]根据以上化合物分子式的定义以及植物化学成分数据库中化合物分子式数据的构成类型，总结得出植物化学成分数据库中分子式的通用表达式，如公式1所示：

$$
\begin{aligned}
&Molecular\ Formula\\
&= [< Atomic\ symbol >< Digit >]_1^n \mid [< Atomic\ symbol >< Digit >]_1^n .< Ritio\\
&> [< Atomic\ symbol >< Digit >]_1^n \mid ([< Atomic\ symbol >< Digit >]_1^n)n
\end{aligned}
$$

Formula 1. General expression of molecular formula.

公式1化合物分子式通用表达式

其中，“$[< Atomic\ symbol >< Digit >]_1^n$”表示分子式基本单元，由元素符号“$< Atomic\ symbol >$”和元素比例数“$< Digit >$”组成，一个完整的分子式可由1到n个分子式基本单元组成；晶体和有机盐分别由阴阳离子的分子式表示，阴阳离子间使用“.”分隔，“.”后面使用

“<Ritio>”表示晶体和盐阴阳离子比例，有整数或者分数的情况，数字为1时通常省略。“$([<Atomic\ symbol><Digit>]_1^n)n$”表示聚合物分子式，由聚合物单体的分子式通式和“( )n”组成；“<Atomic symbol>”元素符号为一个或者两个英文字母，元素符号的首字母为大写，有两个英文字母组成的第二个英文字母为小写；“<Digit>”数字为正整数，数字为1时通常省略。

然后，根据以上化合物分子式的通用表达式，设计分析和处理分子式数据的程序算法，如图1所示：

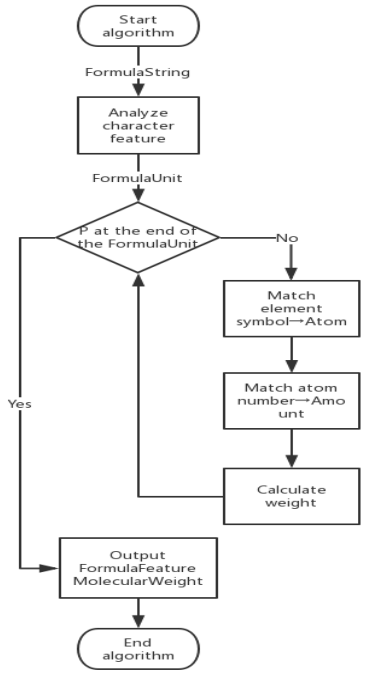


Fig.1 Molecular formula data processing algorithm

图1 分子式数据处理算法

然后，根据以上算法设计分子式数据分析和处理工具程序软件，使用计算机程序批量处理植物化学成分数据库中的分子式数据。程序读入分子式字符串*FormulaString*后，根据分子式通用表达式解析其字符特征，比如分子式中的“（）”、“.”等，以此判断该分子式的类型，并且将分析所得的分子式特征数据存入分子式特征变量*FormulaFeature*中；根据该分子式的类型，截取其中的分子式单元字符串，分别存入分子式单元字符串变量*FormulaUnit*中，普通的化合物分子式单元字符串为一个，晶体和有机盐分子式单元字符串为两个或多个，聚合物为“（）”中单体的分子式单元字符串；依次遍历分子式单元字符串*FormulaUnit*，根据分子式通用表达式分别解析出元素符号和元素数量存入元素变量*Atom*和元素数量变量*Amount*中，并且根据元素的原子量和元素数量计算该分子式基本单元的分子量，并且累加存入化合物分子量变量*MolecularWeight*；最后输出化合物分子量数据*MolecularWeight*和化合物分子式特征数据*FormulaFeature*。

最后，设计化合物分子式特征数据表，用于存储分子式数据处理程序的运算结果数据，包括分子式、分子量、聚合物标示、聚合物单体、晶体盐标示、晶体盐阴阳离子、晶体盐比例、分子式中常见元素的个数等，设计分子式特征数据表结构，如表1所示：

表1 化合物分子式特征数据表

Table 1 Compound molecular formula feature data table

序号	属性名称	描述说明
1	CompoundID	化合物编号
2	MolecularFormula	分子式
3	Molecularweight	分子量
4	Polymer	聚合物标示
5	PolymerMonomer	聚合物单体
6	CrystalSalt	晶体盐标示
7	Cation	晶体盐阳离子
8	Anion	晶体盐阴离子
9	Ratio	晶体盐比例
10	C	碳原子个数
11	H	氢原子个数
12	O	氧原子个数
13	N	氮原子个数
14	P	磷原子个数
15	S	硫原子个数
16	Cl	氯原子个数
17	Br	溴原子个数
18	I	碘原子个数
19	F	氟原子个数

3 分子式检索和筛选

使用上文设计的算法编写分子式数据分析和处理程序，对植物化学成分数据库中分子式数据逐一进行分析和处理，所得到的分子式特征数据存入分子式特征数据表。利用完整的分子式特征数据结合数据库中原有的成分化合物数据，设计检索程序即可实现成分化合物分子式的综合检索和筛选。检索方式有：传统的分子式精确和模糊检索，例如检索分子式为“C6H10O7”的化合物；分子量精确检索和分子量范围检索，例如检索分子量为100到150的化合物；聚合物检索，例如检索单体为“C2H4”的聚合物；晶体盐检索，例如检索比例为10个结晶水“H2O”的晶体化合物；原子筛选，例如检索碳原子“C”数大于20，且含有氯原子“Cl”数为1的化合物；检索程序还可以结合分子式、分子量、聚合物参数、晶体盐参数、常见原子数等检索参数，对数据库中的化合物进行自由组合综合检索。

植物化学成分数据库中初始的化合物数据为化合物名称数据、化合物结构数据、分子式数据以及化合物标示数据等，原有的天然产物检索方式只有根据化合物名称检索或者输入化合物结构检索，并不支持一定条件下天然产物范围的检索和筛选。通过对植物化学成分数据库中的成分化合物分子式数据的分析和处理，得到化合物分子式的特征数据，即可实现分子式数据综合检索和筛选。植物化学成分数据库的一个最主要功能就是在天然产物中筛选化合物药物[5]。根据近年以来的科学研究，植物中能够提取的生物碱类化学成分具有良好的抗肿瘤作用，并且具有疗效高、耐受性好、都副作用小等特点[6]。在植物化学成分数据库中利用分子式综合检索，直接筛选含氮类有机化合物或者含氮类有机盐，从而可以更加快捷有效地筛选出抗肿瘤天然产物生物碱。例如，筛选含有可抑制黑色素肿瘤生长的小檗碱类生物碱的天然产物[7]：在植物化学成分数据库中检索小檗碱类生物碱及其相关有机盐化合物，检索条件为分子式是“C20H18NO4”的化合物或者阳离子为“C20H18NO4” 有机盐，以及含1个氮原子“N”、4个氧原子“O”、碳原子“C”在18—22个范围内的化合物。根据以上检索条件对植物化学成分数据库的分子式特征数据进行检索，检索部分结果见表2所示。再根据检索结果中化合物对应的结构以及性质数据，进行功能药物化合物筛选，即可得到能够从植物中提取的具有抗肿瘤作用的小檗碱类生物碱。最后结合来源植物数据，得到含有小檗碱类生物碱的天然产物，从而实现天然产物药物筛选过程。

4 结 论

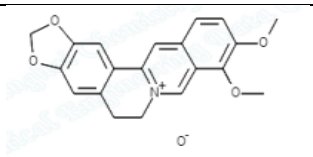
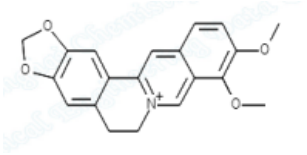
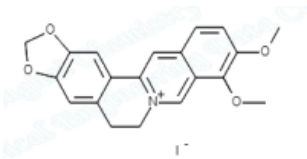
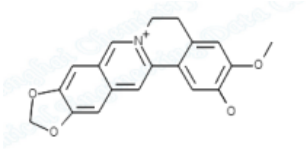
植物化学成分数据库分子式数据的处理过程还可以应用于其他类型的化学数据库，可以根据数据库自身特点开发出其他类型检索和筛选方式，例如含能材料数据库可以使用含有特定数量的“N”、“O”、“F”、“B”元素，并且在一定分子量范围内的检索条件；聚合物物性数据库可以使用聚合物单体分子式结合碳元素数量为检索条件；功能材料数据库则可以使用功能基团或者功能原子分子式检索等，以较容易实现的数据库挖掘技术来实现化学数据库更大的使用价值。对于例如聚合物含有多个重复单元或非重复单元的、盐有两

种以上组分的、分子式中含有同位素的等特殊表达方式的分子式类型，分子式数据处理程序还应根据每种类型的字符特征，挖掘出其内在的化合物分子式特征数据。

上文所介绍的植物化学成分数据库化合物分子式数据处理过程，是通过计算机程序进行数据预处理，得到分子式中更深层次的数据信息。当大型化学数据库中数据量非常庞大时，计算机运行分子式分析程序生成分子式特征数据需要耗费相当大的时间和资源，分析结果产生的分子式特征数据也需要占用一定量的存储空间。但是经过前处理得到的化合物分子式数据更有价值，分子式的检索和筛选也更加有效。另外，分子式综合检索技术可以尝试结合其他检索方式，如子结构检索等，对检索结果集进行二次筛选，从而大大提高化合物检索和筛选的效率。

表2 植物化学成分数据库分子式检索结果示例

Table 2 An example of molecular formula retrieve in plant chemical constituent database

化合物	分子式	来源植物
	C20H18NO4.HO	Corydalis chaerophylla
	C20H18NO4	Evodia rutaecarpa
	C20H18NO4.I	Thalictrum przewalskii
	C19H16NO4	Isopyrum thalictroides

References

1 周蕊. 美国《化学文摘》光盘版和网络版中化学物质检索方法的选择与介绍[J]. 图书馆学研究, 2008(4):73-75.

2 徐挺军. 植物化学成分数据加工系统的研究与设计[D]. 上海交通大学, 2016.

3 IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Blackwell Scientific Publications, Oxford, 1997.

4 Barker P G. Syntactic definition and parsing of molecular formulae: Part 1 Initial syntax definition and parser implementation[J]. Computer Journal, 1975, 18(4):355-359.

5 徐挺军,赵英莉,陈维明.植物化学成分数据库建设[C]. 第十一届科学数据库与信息技术学术研讨会论文集, 中国科学院上海有机化学研究所, 2012: 30-37.

6 张靖, 杨柳, 高文远. 天然抗肿瘤药物研究进展[J]. 中草药, 2010, 41(6):1014-1020.

7 曹明哲, 季宇彬, 辛国松,等. 天然植物中生物碱类抗肿瘤药物研究进展[J]. 亚太传统医药, 2015, 11(7):59-61.

Molecular formula data processing in plant chemical constituent database

XU Tingjun and CHEN Weiming

Information Center, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

Abstract: In order to solve the difficult problem of compound retrieval and screening in specialized chemical database, the definition and the basic composition of molecular formula was studied, and the general expression of molecular formula in the plant chemical constituent database has been gotten, then an algorithm program was designed to analyze and process molecular formula data, and generate molecular formula feature data. In the plant chemical constituent database, the compound data combined with molecular formula feature data can been used for compound retrieval and screen. In order to introduce an application of compounds screening after molecular formula data processing, example of the plant chemical constituent database was taken. The molecular formula data processing method solved the problem of molecular formula inconsistency between user side input and database side representation, and improved the efficiency of compound retrieval and screening. The molecular formula data processing method was applied to specialized chemical database in order to enrich the retrieval methods and technologies, and can be used for functional material compounds screening, organic molecular drugs designing, natural product retrieval, etc.

Keywords: molecular formula data processing; molecular formula feature; compound retrieval; drug screening; chemical database mining

(Received: 2018-06-07; Revised: 2018-07-27; Published: 2018-08-25)