策略·技能

化合物结构存储与检索方法研究

专利检索咨询中心　颜平辉　时嘉鸿

摘要：本文分析了化学结构加工的重要意义，加工现状和化合物结构存储、检索存在的问题。提出了化合物结构存储和基于结构式的检索方法，解决了化合物结构存储唯一性问题，用表结构对 SMILES 表达式重新描述，在不使用任何插件的情况下实现了化合物结构式检索，并通过测试系统验证了该方法是正确的。

关键词：化学结构　表达式转换　数据存储　结构式检索

一、研究背景

化合物结构标引是中国非专利文献数据深加工的重要内容，目前已加工含有确定结构文献 23 万件，绘制化合物结构 5 万件，这些经过深加工的数据对于医药化学领域专利审查具有重要意义。但是，在化合物结构标引和利用方面还存在一些问题：首先，在化合物结构标引过程中，没有实现化合物结构唯一性存储，这样，具有不同名称相同结构的化合物可能会重复绘制，不仅会造成重复加工，还会导致数据不一致，影响数据质量。虽然利用同义词库可以大幅降低这种现象，即在加工过程中，只要同义词组中任一词绘制了化合物结构，则其它词不需要再绘制结构，只需引用即可。但这种方式并不能解决不同文章中名称不同而结构相同未建立同义词关系的化合物。其次，由于标引化合物结构的软件和版本不同，生成同类型化合物结构文件格式存在差异，标引数据能否正常用于结构式检索需要验证。因此，研究化合物结构存储与结构检索方法对于减少重复加工，提高数据质量，降低加工风险具有重要意义。

化学结构存储与检索方面已有较多研究，主要解决方法都是通过在关系数据库上增加插件来实现，如 CambridgeSoft Oracle Cartridge、Symyx Direct、Chord 和 JChem Cartridge。通过插件实现化学结构存储与查询，对数据库类型有严格的限制，不仅性能上有影响，更主要的原因是插件功能完全依赖于开发商，对一些特定需求想要作修改和调整是很困难的。如果自主实现类似解决方案，时间和资源成本几乎是不可接受的。

因此，本文采用不使用插件的方法，将 Mol 文件转化为具有唯一性 SMILES 表达式存储于数据库，再将 SMILES 表达式解析为易于检索的数据表结构[1]，可直接用 SQL 语句实现化合物结构式检索。该方法最大的特点在于简单，使得自主开发化合物存储与检索系统成为可能，没有了插件限制，可灵活修改系统，满足各种特定需求。此外，不使用插件，构建系统不再受数据库平台限制；稳定性和性能有一定改善；SMILES 表达式数据量极小，易于网络传输。

二、结构表达方式转换

化学结构表达方式转换过程是从 Mol 文件中提取化学结构信息，按照 SMILES 编码方式生成具有唯一性 SMILES 表达式，以实现化学结构唯一存储。

（一）Mol 文件结构

Mol 文件[2]是 Molecular Design Limited

（MDL）公司开发的化学表文件之一。文件中包含一个连接表，连接表是存储化学结构原子、键、相互关系和属性的重要结构。连接表主要由计数行、原子区块、键区块组成，如图 1 所示，根据特定结构和应用需求，还可包括原子列表区块、结构文本描述区块和属性区块。计数行主要说明结构的原子、键、列表总数，是否为手性结构，连接表版本号。原子区块主要记录原子的坐标（x,y,z）、元素符号、电荷数、氢原子数、化合价、反应物类型、反应物数量、构型信息，一行记录一个原子的信息。键区块描述了键连接的原子、键类型、立体性质、拓扑关系，一行记录一个键的信息。

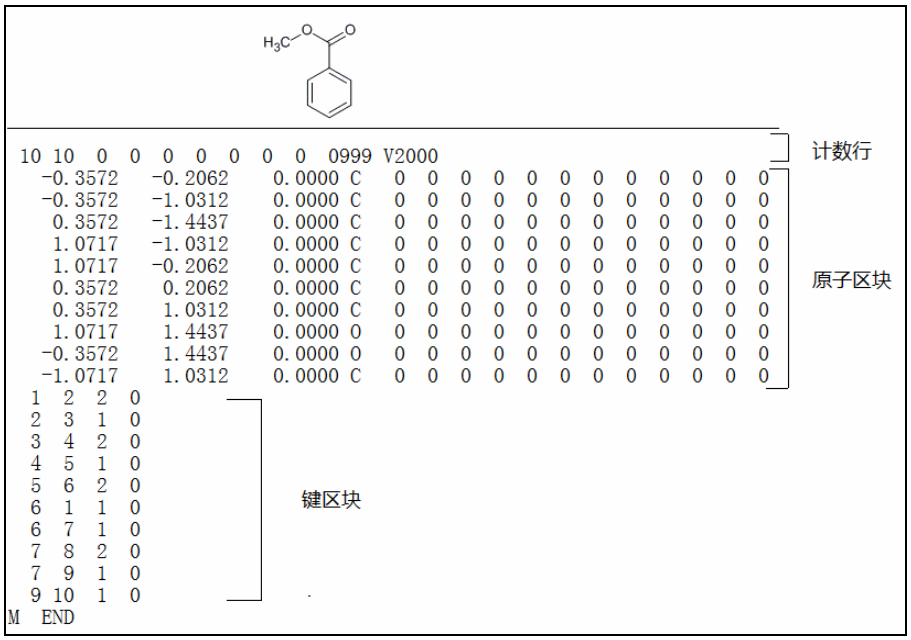


图 1　Mol 文件结构

（二）SMILES 编码规则

简化分子线性输入规范 (Simplified molecular input line entry specification，SMILES)[3]，是一种用ASCII码描述分子结构的规范。SMILES 由 Arthur Weininger 和 David Weininger 于20 世纪 80 年代晚期开发，主要由日光化学信息系统有限公司修改和扩展，其编码规则如下：

1. 原子，用在方括号内的化学元素符号表示。 例如[Au]表示“金”。 有机物中的C、N、O、P、S、Br、Cl、I等原子可以省略方括号，其他元素必须包括在方括号之内。氢原子常被省略，对于省略了方括号的原子，用氢原子补足价数。例如，水的 SMILES 编码为 O，乙醇为 CCO。

2. 键，用“ ”、“=”、“#”、“:”分别表示单键、双键、三键和芳香键，一般单键和芳香键省略。如二氧化碳表示为 O=C=O，氰化氢表示为 C#N。

3. 环，如果结构中有环，则要打开，断开处的两个原子用同一个数字标记，表示原子间有键相连。环己烷(C6H12)表示为 C1CCCCC1，芳环中的原子用小写字母表示。

4. 支链，碳链上的分支用圆括号表示。如丙酸表示为 CCC(=O)O，三氟甲烷表示为 FC(F)F 或 C(F)(F)F。

5. 双键两侧的结构分别用符号/和\表示，例如，F/C=C/F 表示反二氟乙烯。

6. 离子化和物，将包含正负离子的结构单独编码，将带电荷原子和所带电荷单独写在方括号中，两部分用“.”连接。

7. 手性，用“@”和“@@”表示，“@”表示从 SMILES 编码的第一个原子看起，其它与手性原子连接的原子或基团呈逆时针排列；“@@”表示从 SMILES 编码的第一

个原子看起，其它与手性原子连接的原子或基团呈顺时针排列。

（三）唯一性编码方法

从 Mol 文件读取相应的信息，按照 SMILES 编码规则，生成 SMIlES 表达式。通常，一个确定结构可以写出多个符合上述编码规则的表达式，只有实现了确定结构唯一线性编码，才能在数据库中作主键列或唯一值列，以避免化学结构重复绘制和存储。CANGEN 方法[4]解决了这一问题，CANGEN 方法由 CANON 和 GENES 算法组成，CANON 算法将分子结构看作由节点（原子）和边（键）构成的图，在分子拓扑结构的基础上给每个原子赋 ID 唯一标识原子，原子 ID 确定主要依据原子不变性，原子不变性包括连接度、非氢键数、原子序数、原子电荷符号、原子电荷绝对值、氢原子连接数等，并结合相应函数计算已保证原子 ID 唯一；在 CANON 计算结果的基础上，GENES 算法中将分子结构视为树，选择 ID 值最小的原子作为起点，根据原子 ID 确定连接方式，深度优先遍历树生成唯一编码。

三、数据库结构与结构解析

（一）数据库结构

数据库存储了化合物结构基本信息和子结构检索辅助信息，包括一个基本信息表和五个辅助信息表，如图 2 所示。

基本信息表是指 sqlmol_compound 表，存储化合物核心信息，包含“compoundid”、“SMILES”、“Mol”、“Gif”、“CName”等字段。“compounded”字段为整数自增字段，为表主键，以便其它表引用；“SMILES”存储了化合物 SMILES 表达式，具有唯一性约束，实现化合物结构唯一存储；“Mol”字段存储了该结构的 mol 文件，用于结构编辑和数据交换；“Gif”字段存储了化合物结构的 Gif 图像文件，用于结构快速展示；以及名称、来源、登记号等信息用于常规属性检索。

辅助信息表是根据基本信息表中 SMILES 字段，采用数据表方式重构化合物结构信息，以便通过常规数据库实现基于结构的化合物检索。辅助信息表包括 SMILES 符号表（sqlmol_symbol,见表 1）、键类型表（sql_bond_type, 见表 4）、元素类型表（sql_element_type,见表 5）、化合物元素组成表（sql_compound_elements,见表 6）、化合物键组成表（sql_compound_bonds，见表 7）。SMILES 符号表用于 SMILES 表达式解析，化合物元素组成表描述了化合物原子构成信息，化合物键组成表包含了键组成信息和键连接关系信息，键类型表和元素类型表存储了键和元素常量信息。

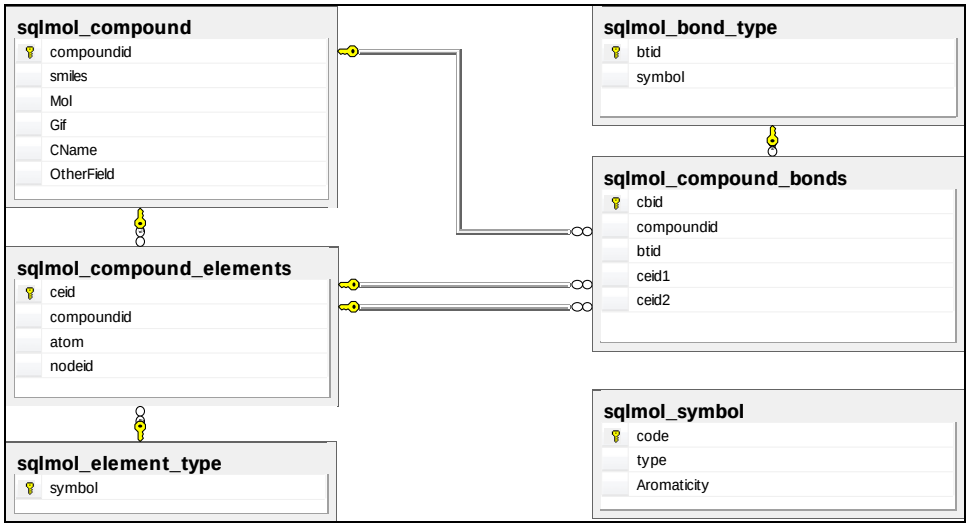


图 2 化合物数据库结构

（二）结构解析

结构解析是将基本表中的 SMILES 字段值按照数据表表达化合物结构的要求，填充到化合物元素组成表和化合物键组成表的过程。 该过程分为三个步骤，以解析“S(C)(C)=O”为例：其过程如下：

第一步，依次扫描 SMILES 表达式字符串，对照 SMILES 符号表（表 1），获得单个具有特定意义的符号和代表符号类型、芳香性和下一次读取的偏移量，扫描结果如表 2 所示。

第二步，根据前一步的结果，按照 SMILES 编码规则，得到以元素为中心的中间结果表 3，该中间结果表通过 pnodeid 指出与该元素相连的前一个元素，从而描述了各元素的连接关系，并说明了元素间的键类型。

第三步，将中间结果表 3 的信息分为两部分，元素组成存入表 6 化合物元素组成表，键组成和连接关系存入表 7 化合物键组成表。

表 1 SMILES 符号表

Code	Type	Aromaticity
C	a	False
O	a	False
S	a	False
=	b	False
#	b	False
(	r	False
)	r	False

表 2 SMILES 表达式字符解析结果

Order	Code	Type	Aromaticity	Offset
1	S	a	0	1
2	(	r	0	1
3	C	a	0	1
4	)	r	0	1
5	(	r	0	1
6	C	a	0	1
7	)	r	0	1
8	=	b	0	1
9	O	a	0	1

表 3 SMILES 表达式解析中间结果

nodeid	code	pnodeid	bond	bondid
1	S	0	-	b1_0
3	C	1	-	b3_1
6	C	1	-	b6_1
9	O	1	=	b9_1

表 4 键类型表

btid	Symbol
26	C-S
29	0=S

表 5 元素类型表

symbol
C
O
S

表 6 化合物元素组成表

ceid	compoundid	atom	nodeid
10848	763	C	3
10849	763	C	6
10850	763	O	9
10851	763	S	1

表 7 化合物键组成表

cbid	compoundid	btid	ceid1	ceid2
10878	763	26	10848	10851
10879	763	26	10849	10851
10880	763	29	10850	10851

四、基于结构式的查询方法

基于结构式查询的方法主要分为两类，精确查找和子结构检索。

（一）精确查找

利用化合物结构 SMILES 编码在数据库中存储的唯一特性，可将输入结构信息用同样方法编码，用数据库常规属性检索方式精确匹配 SMILES 表达式获得结果。

（二）子结构检索

利用化合物结构存储的辅助信息，按照

前一节方法解码输入的 SMILES 表达式，获得与数据库中表 6 和表 7 相同结构的数据，在此基础上动态生成 Sql 语句并执行查询。

由表 6 可知化合物的原子构成信息，由表 7 可知键类型信息和连接关系。由于键包含了原子构成信息，因此子结构检索的实质为查找包含输入结构所有键且键连接关系相同的化合物。据此，动态 SQL 语句生成分为以下两步：

第一步，键组成信息匹配。由表 7 可看到，“S(C)(C)=O”由三个键组成，键类型 ID（btid）值为“26”，“26”，“29”，分别代表“C-S”，“C-S”，“0=S”键，表达出键类型 ID 同时等于这三个值即可。由于表 7 中键类型 ID 值位于同一列，因此需用化合物 ID 与自身连接形成 From 子句，再进行键类型 ID 条件限制，如果结构中存在键类型相同键实例不同的情况，需进一步限制，即键实例 ID 不相等，说明是不同键实例，语句片段如下：

From sqlmol_compound_bonds b3_1, sqlmol_compound_bonds b6_1, sqlmol_compound_bonds b9_1 where b3_1.compoundid=b6_1.compoundid and b3_1.compoundid=b9_1.compoundid and b3_1.btid=26 and b6_1.btid=26 and b9_1.btid=29 and b3_1.cbid<>b6_1.cbid

第二步，键连接关系匹配。表 7 中，列 ceid1 和 ceid2 表示了化合物键的连接关系，ceid1 表示键的第 2 个原子（后继节点），ceid2 表示键的第 1 个原子（前驱结点），连接关系建立的过程是对每个键寻找前驱结点的过程。找前驱结点的方法是先在所有后继结点中找，若找不到再从前驱节点中找，前驱节点相同时需避免重复，找到即停止。语句片段如下：

b6_1.ceid2=b3_1.ceid2 and b9_1.ceid2=b3_1.ceid2 and b9_1.ceid2=b6_1.ceid2

完整的查询语句如下：

*select * from sqlmol_compound where compoundid in (select distinct b3_1.compoundid from sqlmol_compound_bonds b3_1, sqlmol_compound_bonds b6_1, sqlmol_compound_bonds b9_1 where b3_1.compoundid=b6_1.compoundid and b3_1.compoundid=b9_1.compoundid and b3_1.btid=26 and b6_1.btid=26 and b9_1.btid=29 and b3_1.cbid<>b6_1.cbid and b6_1.ceid2=b3_1.ceid2 and b9_1.ceid2=b3_1.ceid2 and b9_1.ceid2=b6_1.ceid2)*

执行该语句可查到所有包含“S(C)(C)=O”的结构。

五、方法验证

通过构建测试系统，对上述方法进行了验证。测试系统由 Web 客户端、Web 服务器端、数据库构成。数据库将 SMILES 表达式解析封装为自定义函数，将结构式查询封装为存储过程。Web 客户端采用 JME Molecular Editor 作为化学结构输入工具，查询时，客户端将化学结构解析为 SMILES 表达式，传到 Web 服务器端， Web 服务器端接受客服端的请求，调用数据库存储过程，将结果返回客服端。如图 3 所示，在分子编辑器中输入苯环，返回了具有苯环的化学结构列表。检索结果与预期相同，说明检索方法是正确的。

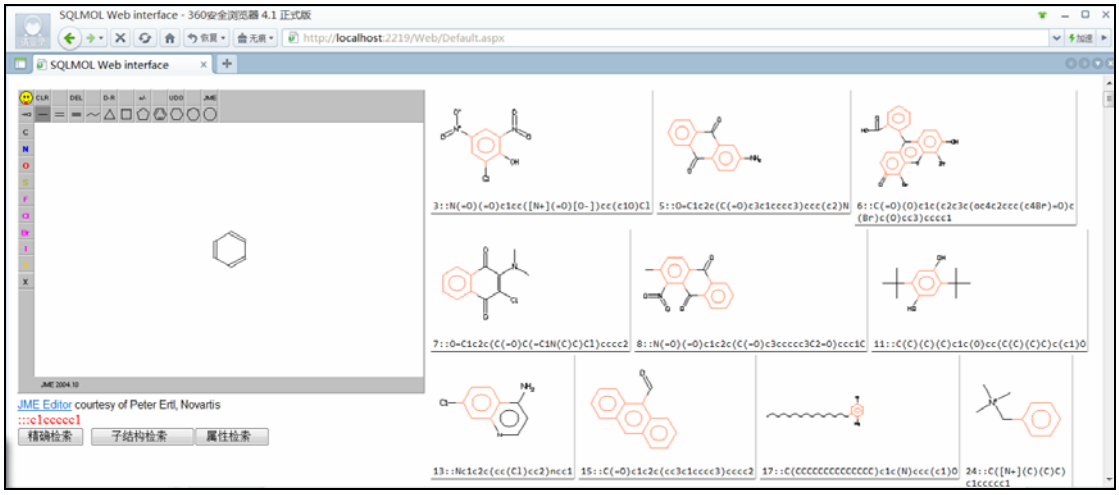


图 3 化学结构检索测试界面

六、结论与讨论

本文阐述了 Mol 文件格式和 SMILES 编码规范，研究了 SMILES 唯一编码的生成方法，化合物存储结构和基于结构式的检索方法，并通过测试系统验证了方法的正确性，为化合物结构加工去除重复数据、避免重复加工提供了方法支持，为化合物结构检索提供了结构检索手段，对数据加工和专利审查工作具有重要现实意义。基于结构式的化合物检索一直都是生物、化学信息学领域的热点和难点，本文虽然做了有益的探索，但在以下几方面还需要进一步完善：在 SMILES 表达式解析方面，对一些特殊的结构还未提供支持，需要进一步完善；在动态生成 SQL 语句方面，还需增强可靠性；客户端分子结构输入工具应自己开发，才能灵活构建应用系统，满足需求；此外测试数据量较小，大数据量下效率还不确定，需进一步验证。

参考文献

[1] Adel Golovin,Kim Henrick. Chemical Substructure Search in SQl. J.Chem.Inf.Model. 2009,49,22-27.

[2] Arthur Dalby, James G.Nourse, W..Douglas Hounshell etc.Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. J.Chem.Inf.Comput.Sci. 1992,32,244-255.

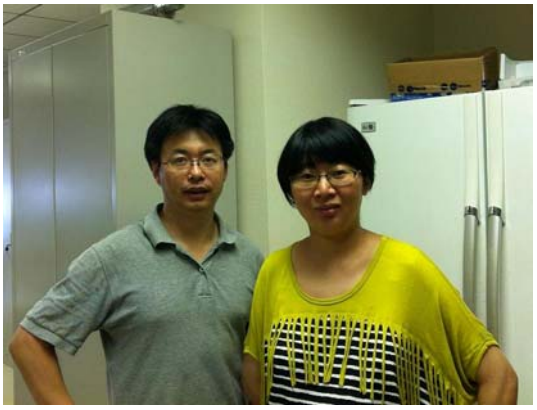
[3] David Weininger. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J.Chem.Inf.Comput. Sci.. 1988, 28, 31-36.

[4] David Weininger, Arthur Weininger, and Joserph L. Weininger. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. J.Chem.Inf.Comput.Sci.. 1989, 29, 97-101

作者简介:

颜平辉（左），计算机与信息系统专业，2008 年入局，在检索咨询中心数据加工处从事软件开发与系统管理工作。

时嘉鸿（右），分析化学专业，2008 年入局，在检索咨询中心数据加工处从事化学结构绘制，医药类期刊深加工工作。



（专利检索咨询中心 杨晓春 审校）