非专利文献数据加工中化合物结构标引常见问题解析

专利检索咨询中心 李雪芹 武芳 马琴琴

摘要: 包括化合物中英文名称、化学结构和职能符等内容的化合物结构信息是医药领域文献数据加工的重要内容。准确、全面地反映化合物信息对于构建数据库有着重要的意义。本文针对化合物结构标引过程中出现的常见问题进行了归纳和总结，以期对规范化合物标引，完善化合物结构数据库提出一些建议。

关键词: 化合物结构 标引 非专利文献 检索

在化合物结构标引过程中，很多非专利文献只提及了化合物名称，并没有对其结构进行描述，给化合物结构标引工作带来困难。利用《中国药物专利数据库西药词典》、《中国药典》、CAweb等各种资源对化合物进行检索成为必不可少的工作。由于数据来源的不同，检索到的化合物结构会有差异。如果不进行甄别，就会造成标引的不一致。本文就一些实例对

化合物结构标引中遇到的问题进行分析说明，并提出解决办法。

一、由化合物名称确定结构时存在的问题

标引化合物为结构相近的一类化合物组成的混合物的情况。此类化合物包括一些糖苷类抗生素，如庆大霉素、阿克拉霉素、卡那霉素、制霉菌素、新霉素、乙酰螺旋霉素，以及谷维素等。例如，制霉菌素是多组分多烯大环内酯类抗真菌抗生素，由文献《制霉素中制霉菌素A1、A3和多真菌素B含量比例的研究》[1]可知，其主要组分确定为：制霉菌素A1、A3和多真菌素B。如果在标引过程中将制霉菌素A1作为制霉菌素的结构提交，并不合适（能够明确所指结构为制霉菌素A1的情况除外）。针对上述情况，在标引时需要区分，慎重引用。

标引化合物为一类化合物的统称的情况。在标引时，要充分检索各种资源，明确其实际含义，保证标引结构的准确性。如茶多酚，为茶叶中多酚类物质的总称（词条含义参见Wikipedia），但在中国药物专利数据库西药词典（以下简称西药词典）中茶多酚对应的结构为表棓儿茶素-3-五倍子酸酯（(-)-epigallocatechin 3-O-gallate，CAS:989-51-5）。鞣质，其含义为一类结构复杂的酚类化合物（词条含义参见Wikipedia），但在西药词典中对应的结构为柯里拉京（Corilagin，CAS:23094-69-1）。对于这类情况，标引员可以首先从名称上质疑，然后通过检索各种资源确证引用概念的真实含义，避免错误的发生。

标引化合物是复方药物制剂的商品名的情况。此种情况不应做为一个化合物标引结构。如《高效耐酶的广谱抗生素泰能》[2]一文中泰能为亚胺培南与西司他丁钠按1:1组成的复方制剂（亚胺培南与西司他丁钠的结构如图1），虽然检索结果显示，亚胺培南-西拉司丁是作为一个化合物给出了CAS号（92309-29-0）；但其英文名称为Cilastatin-imipenem mixt，相当于亚胺培南、西拉司丁两个组分的混合物，与舍曲林天冬氨酸盐这类化合物不同。因此，不要将泰能作为一个化合物标引，可将西司他丁钠和亚胺培南分开标引。

因出版年限早晚导致化合物结构与名称不相对应的情况。一些早期发表的文献与后来发表的文献存在化合物名称不同而结构相同的情况，如山柰酚和山柰素（图2）。山柰素本身结构不同于山柰酚，但在一些早期的文献中，例如《兴安柴胡化学成分研究（Ⅰ）》[3]，山柰素指代的是山柰酚。针对上述情况，应当提交文献实际指代的结构。判定文章实际指代的结构

可从文章提供的信息，如化合物的结构鉴定信息，以及文章英文摘要中其对应的英文名称等信息进行确定。

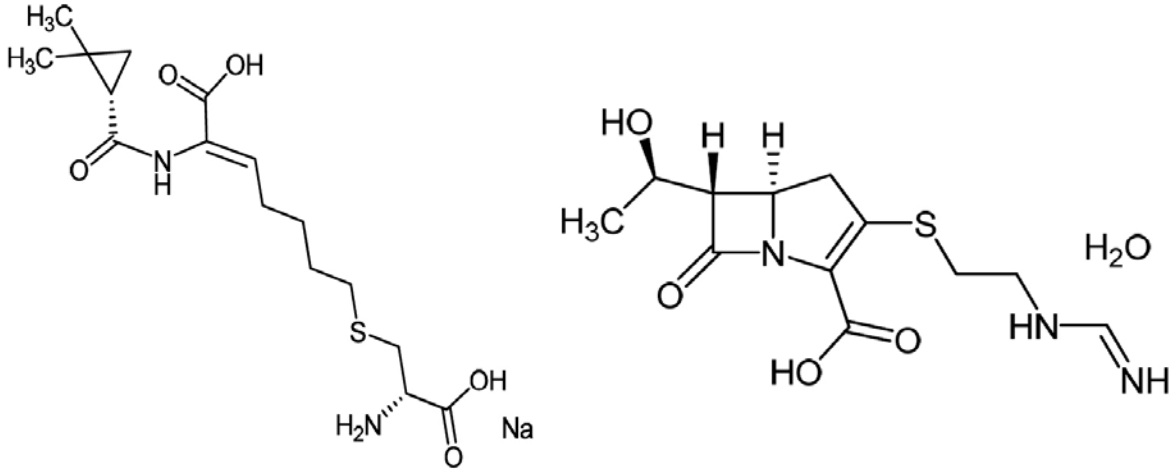


图 1 西司他丁钠（左）和亚胺培南（右）

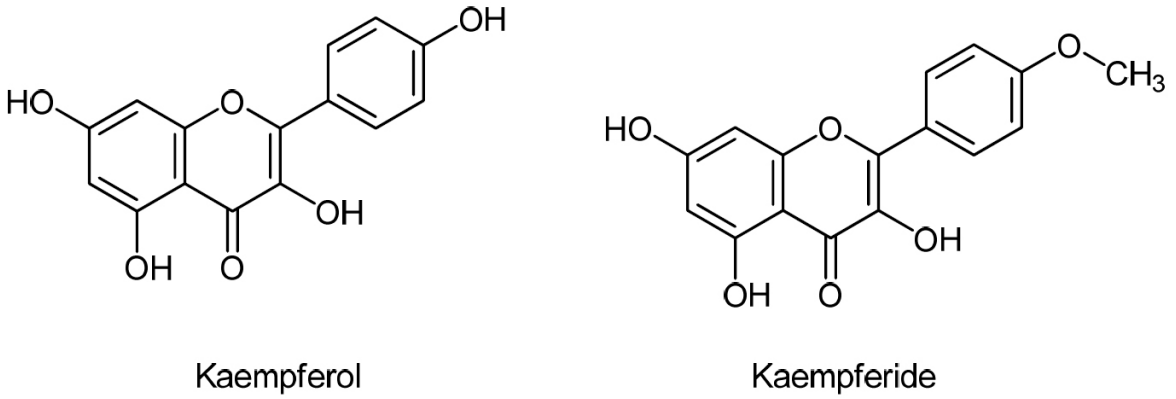


图 2 山柰酚（左）和山柰素（右）

二、化合物结构及绘制中存在的问题。

化合物结构中结晶水的确定。一些化合物，检索各种数据库可知，同一名称，对应若干个结构，一些含有结晶水，而一些不含结晶水，给标引工作带来困难。此时，应当核对美国《化学文摘》（简称 CA）或者其他化学领域较为权威的数据库，确定其实际对应的结构。如阿仑特罗，查证 CA，为不含结晶水的结构（图 3 左图）。但是如果文章明确指出所涉及的结构含有结晶水且与 CA 有不一致的情况，要以原文为准。

结构易混淆的化合物。如《青藤生物碱成分的研究》[4] 一文中的青藤碱（sinomenine）和青风藤碱（sinoacutine），其结构如图 4 所示。某些数据库将上述两个概念等同，实

际应为不同的化合物，在标引结构时应当注意。具体做法可以从文献中查找其对应的英文名称（这些英文名称可能出现在文章正文，也可能出现在英文摘要中），之后通过CA确证。如果原文中没有相应的英文名称，可以利用各种外网资源进行检索，在检索时要注意，尽量检索较为权威的网站，如Wikipedia，google scholar，CAweb等等，以保证数据的准确性。

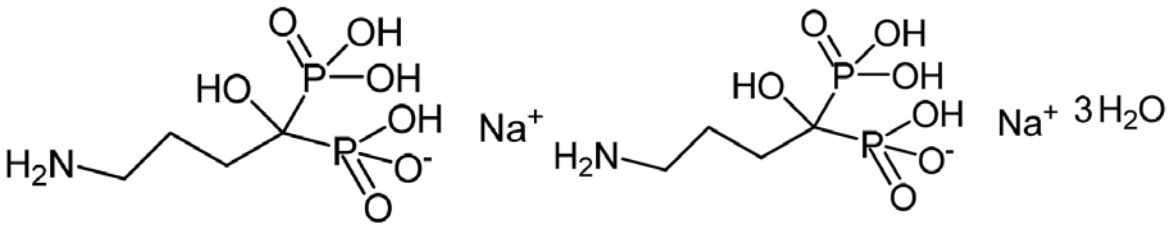
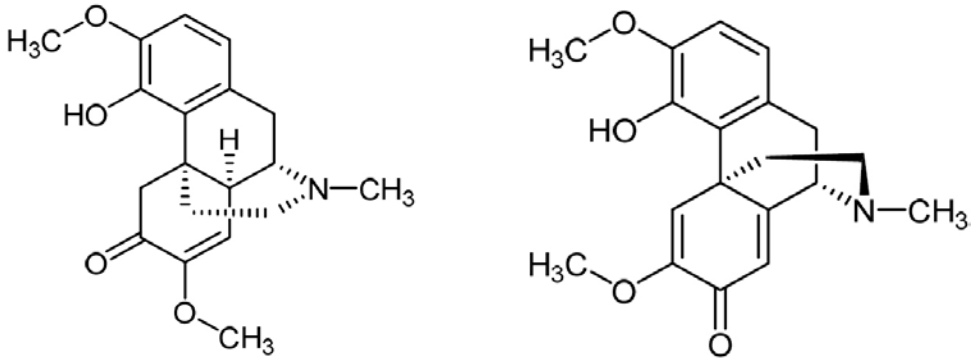


图3 阿仑特罗（左）和含结晶水的阿仑特罗（右）



sinomenine　　sinoacutine

图4 青藤碱（左）及青风藤碱（右）

化合物本身有确切结构，但其盐结构不确定。此时应当通过CA查证，避免“想当然”，如盐酸川芎嗪、醋酸奥曲肽（结构见图5）。盐酸川芎嗪经CA检索，英文系统名称为：2, 3, 5, 6-Tetramethyl pyrazine, Hydrochloride(1:x)，且分子式为$C_8H_{12}N_2 \cdot xHCl$。由于不能确定川芎嗪与盐酸的比例，因此盐酸川芎嗪为结构不确定的化合物。在标引过程中，如遇到盐酸川芎嗪，建议标引并提交川芎嗪的结构。

利用chemdraw软件绘制结构可能产生错误。chemdraw软件可以利用化合物的英文名称绘制化合物结构，但绘制完毕应该进行结构确认。如二巯丙磺钠，对应的英文名称Sodium Dimercaptopropanesulfonate，此时利用chemdraw生成的结构为图6中右图所示，经检索，此结构是错误的，正确结构为左图所示。此种现象的发生是由于常用的英文名对于巯基基团的具

体位置没有说明，在核对CA后，确定其英文名称为1-Propanesulfonicacid, 2, 3-dimercapto-,sodium salt，确定巯基位置为2位和3位，据此名称则可以绘制出正确的结构，如图6左图所示。

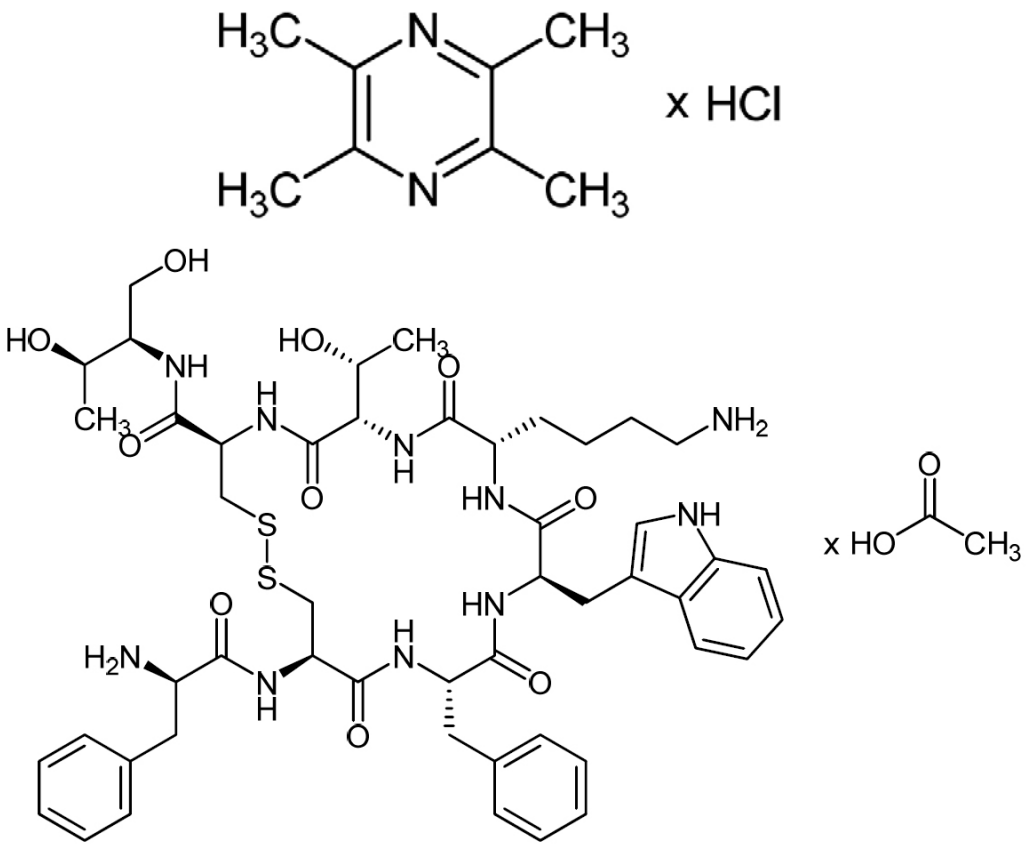


图 5　盐酸川芎嗪（上）和醋酸奥曲肽（下）

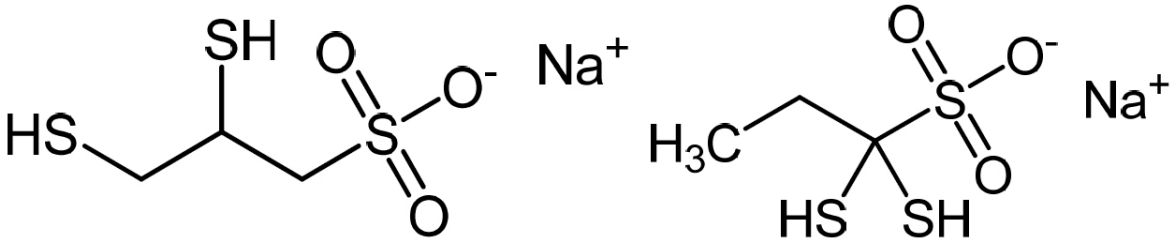


图 6　二巯丙磺钠（左）及其易混结构（右）

化合物结构标引中出现的问题总结其形成原因有以下几点：第一、外网信息资源中的不准确性和不一致性导致标引员将一些不具备特定结构的物质进行标引；第二、文章本身存在指代不清的情况，如同时存在中、英文名称，而两种名称分别指代不同的化合物。这些存在的不确定因素造成化合物结构标引的潜在“陷阱”。标引的化合物应准确反映文章实际指代的

结构，需要对文章中化合物结构信息进行深入挖掘。不能从文章中获知该化合物的其他信息时，可根据实际情况区别对待：文章提供中英文名称信息不一致时，应当以英文名称对应的结构为准；文章只给出中文名称，而其对应的英文名称不唯一时，应该引用其对应的公知、惯用的名称所指代的结构。两种情况最终结构的确定应以 CA 及其他较为权威的数据库为准。

非专利文献中的化合物信息浩如烟海，标引时存在着很多不确定因素，准确地反映化合物结构信息并非易事。但只要充分检索，细心甄别，便可以提高非专利文献数据加工化合物数据库的质量，为审查员的检索提供准确、可靠的信息和便捷的检索途径。

（专利检索咨询中心　杨晓春 审校）

参考文献

1. 汪素岩、王健、孟铮．“制霉素中制霉菌素 A1、A3 和多真菌素 B 含量比例的研究”．《中国药事》1996 年第 10 卷第 1 期第 41-42 页。
2. 王炜．“高效耐酶的广谱抗生素泰能”．《安徽医药》2002 年第 6 卷第 3 期第 24-25 页。
3. 宋治中、贾忠建．“兴安柴胡化学成分研究（Ⅰ）”．《兰州大学学报（自然科学版）》1992 年 28 卷 8 期 99-103 页。
4. 李海滨．“青藤生物碱成分的研究”．《贵阳医学院学报》2006 年 31 卷 4 期第 344-345 页。