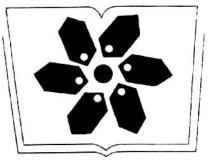
计算机化学化工丛书

Computer, Chemistry and Chemical Engineering Series

药物设计中的分子模型化方法

周家驹 王 亭 著

科学出版社



中国科学院科学出版基金资助出版

计算机化学化工丛书

药物设计中的分子模型化方法

周家驹　王　亭　著

科学出版社

2001

内 容 简 介

本书为《计算机化学化工丛书》之一。

本书以小分子的化学结构信息处理为核心，介绍药物设计中的分子模型化方法。全书分为九章：前三章为第一部分，是一些通用的方法、原理和技术，分别介绍分子结构的计算机处理、分子的相互作用、分子结构信息数据库；后六章为第二部分，分别介绍人工智能中的遗传算法，定量结构活性关系（QSAR）研究、结构匹配和数据库搜索技术、虚拟受体模型方法、药物与受体的对接技术和模拟肽学等六个专题。

本书是一本面向应用、介绍计算机辅助药物分子设计方法的入门书，适合从事计算机化学的研究人员及从事新医药、新农药研制与开发的研究人员阅读和参考，也可供化学各专业大学生及研究生阅读。本书有助于读者了解计算机处理化学信息的基本方法和用于分子设计的各种实用技术。

图书在版编目（CIP）数据

药物设计中的分子模型化方法/周家驹，王亭著. －北京：科学出版社，2001

（计算机化学化工丛书/许志宏主编）

ISBN-7-03-0098690-X

Ⅰ.药… Ⅱ.①周… ②王… Ⅲ. 药物-设计-分子结构-化学模型 Ⅳ. TQ460

中国版本图书馆 CIP 数据核字（2000）第 81844 号

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

源海印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2001 年 4 月第 一 版 开本：850×1168 1/32

2001 年 4 月第一次印刷 印张：7 7/8

印数：1—3 000 字数：198 000

定价：21.00 元

（如有印装质量问题，我社负责调换〈北燕〉）

《计算机化学化工丛书》
编　委　会

《计算机化学化工丛书》序

化学化工是信息量特别大的一门学科。到2000年4月底,在CAS录的化学物质约2400万种,它们的各种性质加上多元体学的性质,可以说是一个无边无际的数据海洋。于是,化学数据库的建设就成为20世纪后20年国际上的一件大事。中国科学院从1979年开始建设化学数据库,迄今已经整整20年。其学院从1979年开始建设化学数据库,迄今已经整整20年。其间,多次得到国家和中国科学院的奖励。

长期以来,人们不仅希望能定性地掌握而且希望能定量地了解化学化工学科的规律,而计算机的能力及相关技术高速发展,正在帮助人们一步一步地实现着这个愿望。从理论化学的计算、谱图解析、化学计量学、化工过程模拟、热力学的复杂计算,都在计算机的帮助下得到了很大的发展。

上述基础工作为化学化工领域的工作者增加了很大的自由度:可以用计算机帮助对化合物的谱图解析,帮助选择合成路线,帮助进行药物分子设计,可以进行新过程、新技术的开发,可以进行大型工业装置的设计,可以对工厂的生产过程进行优化,……等等。

在改革开放的20年里,我们的计算机化学从无到有,形居了一支生机勃勃的研究队伍,也形成了一个很大的用户和读者群体。到21世纪,这个群体更需要有能力利用计算机帮助自己的工作,所以本套丛书中也包含一些计算机化学化工的教材,以利于化学化工本科生和研究生的培养和工程人员自学。

所以,我们希望通过本套丛书介绍一些解决问题的方法,帮助读者在遇到问题时,知道如何去解决问题。为此,要求作者在自己的著作中,要给出软件、数据的出处、网络地址或光盘。时代发展

得很快，仅做到这一点显然还是不够的。我们特别注意到近年来，Internet 网络的高速发展已经给我们的时代带来了巨大变化。到 1999 年 7 月，Internet 已经是一个连接 5600 多万个节点的一个网络系统，它上面的文本信息已经超过 600 亿字节。这些信息一个最大的长处是时间的滞后最少，易于通过计算机获取。

如果能将科学数据库在网上的功能，由数据的存取扩大到运算、绘图、模拟等多方面，必将极大地推动科学数据库工作的发展和广泛使用。在 21 世纪，将逐步可以做到，人们在用户端将数据从库中取出，在服务器端程序系统上计算，结果以图形或多媒体方式输出到用户端。据了解，我们有些作者在自己的工作中已经能够在网上实现数据查询、计算、绘图、三维图形显示等。

进入 21 世纪，Internet 网络系统的应用将更加普及，Internet 网络的客户/服务器的应用将进入千家万户，进入教室和办公室的各个角落。所以，如果能将科学数据库和计算程序库在网络服务器上实现，那么它的普及应用，将会随着计算机网络的推广而推广。

如果有的作者，目前仅能给出单机版本的软件，也欢迎他们能再作一点工作，很快能达到上网服务的目的。相信进入 21 世纪不久，在放户上人们就有可能逐步享受到多种媒体的全方位的科学信息的服务。

这套丛书是我国多位化学化工学科的专家、教授、学者多年辛勤劳动的成果，也是科学出版社、国家自然科学基金委员会优秀研究成果专著出版基金和中国科学院科学出版社基金大力支持的产物，希望它的出版能促进我国 21 世纪计算机化学化工学科的发展，并有助于相关学科发展。

《计算机化学化工丛书》编委会

2000 年 5 月

前　言

分子的结构信息在计算机内部的表达、存储、转换技术，以及研究结构-性质关系的分子模型化方法是计算机化学的基本问题之一。本书以小分子结构信息处理为核心，围绕作者实验室的研究工作，介绍药物设计中的化学结构信息处理和分子模型化方法，是一本面向应用、介绍计算机辅助分子设计方法的入门书。

全书分为两部分。第一部分（第一至三章）是计算机化学中一些通用的方法、原理和技术：第一章介绍分子结构信息的计算机处理，包括结构信息在计算机中的表达、存储和显示，以及分子结构的几何优化、构象分析、构象搜索等重要概念；第二章介绍分子的相互作用，包括对生物过程至关重要的分子间的各种弱相互作用（非键相互作用），即范德华作用、静电作用、氢键作用和疏水作用；第三章介绍分子结构信息数据库，它们是从事分子模型化研究必不可少的工具。通过对中药化学成分及生物活性数据库的描述，引入了交叉学科数据库的概念。

第二部分（第四至九章）是分子模型化方法中的几个专题：第四章介绍遗传算法，包括了一种基于线杂交和面变异思想的新算法；第五章介绍定量结构活性关系（QSAR）研究，包括传统的2D-QSAR 和近年的 3D-QSAR。其中强调了结构信息自动数值化的思想，因为只有实现了结构信息的数值化，才有可能顺理成章地使用各种数学工具，而只有实现了自动数值化，QSAR 方法才能实现大面积的普及应用；第六章介绍结构匹配和数据库搜索技术，其中二维的 GMA 算法和三维的 3DFS 软件都是有特色的成果，而基于三维药效团的大型数据库搜索法可用来发现新的先导化合物；第七章介绍的虚拟受体模型方法是一种很实用的技术，特别是对于受体结构未知的情形；第八章介绍药物与受体的对接技术。对接是基

于受体结构的药物设计中的核心技术，对药物设计感兴趣的读者不可以不知道对接的概念；第九章介绍近年来药物设计中的一个热门话题——模拟肽学。有理由认为，在 21 世纪模拟肽学将在药物设计中占据重要地位，特别是各种非肽的模拟物。

本书论及一个典型的交叉学科研究领域，涉及基础知识比较多，如量子化学基础、分子力学、有机立体化学、生物化学、药物化学等。这些内容不可能在本书中都做介绍，有需要的读者可阅读有关书籍。

本书第一、二、五、六、八章由王亭执笔，第三、七、九各章由周家驹执笔，第四章由何险峰执笔，全书由周家驹定稿。

书中涉及到的 LCC 实验室的一些研究工作大都是集体完成的，对此作出贡献的有：谢前、陈红明、孙红梅、石乐明、张跃敏、何敏、曹凌霄、李松梅、蔡萍、张晓晨、肖贞、谢桂荣、严新建、唐武成等。多年来，各级领导、科学界前辈和同仁对 LCC 的分子模型化研究给予了支持、指导和帮助，我们表示由衷的感谢。他们是：原化工部凌秋萍司长，郑秀兰、赵忠华、余昌申高工，科技部邵立勤副司长，中国科学院郁小民副局长，南开大学李正名先生，北京大学唐有祺先生、徐筱杰教授，北京大学药学院王夔先生、李仁利教授，中国农业大学陈万义先生，中国科学院生态研究中心白迺彬研究员，中国科学院化工冶金研究所郭慕荪先生和计算机化学开放实验室许志宏先生、杨章远、温浩研究员。

最后要感谢中国科学院科学出版基金的支持，以及科学出版社刘俊来编辑为此书的出版所付出的辛勤劳动。

限于作者的水平，书中难免有错误及不妥之处，恳切期望读者批评指正。

作　者

2000 年 3 月于中国科学院化工冶金研究所

通讯处：周家驹，北京 353 信箱，100080，E-mail: jjzhou@lcc.icm.ac.cn

目　录

第一章　分子结构的计算机处理

1.1　分子结构信息的表达

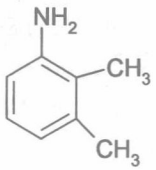
分子在计算机中是如何表达的？这是关心分子模型化方法的人需要了解的首要问题。当我们描述一个分子时最习惯的做法是画出它的分子结构图，以顶点代表原子，边代表原子之间的连接键。然而这样的图形方式却不适合于计算机处理，了解计算机基础知识的人都知道，计算机最擅长接受和处理的是符号、数字以及由符号和数字组成的文件。因此，为了使计算机能方便地处理分子结构信息，需要采取计算机化的方式。

1.1.1　化合物的命名

化合物的命名可以在一定程度上反映分子的结构。早期通常使用简洁的习惯命名法，后来随着化合物种类的急剧增加，这种缺乏规律的命名法逐渐被遵循一套大家所公认的规则的系统命名法所代替。我们所普遍使用的系统命名法是 IUPAC 命名法，该命名规则由国际纯粹与应用化学联合会(IUPAC, International Union of Pure and Applied Chemistry)经过多次修订后于 1979 年正式颁布。例如习惯称为三甲胺(trimethylamine)的化合物的系统命名则是 *N*,*N*-dimethylmethanamine。但是该命名法仍然存在一个缺点，就是对许多化合物无法得到惟一的命名，因而限制了它在化学信息检索系统中的应用。针对这个问题，美国化学文摘社(CAS, Chemical Abstracts Service)提出了一套自己的命名规则，称为 CA 索引名(CA index name)。CA 索引名可以做到对每个化合物只给出一个命名，从而保证了化合物与命名一一对应，便于实现化合物命名的自动化和根据命名对化合物信息的检索。

1.1.2 碎片码、线形码、拓扑码

化合物的命名所表达的结构信息是非常笼统的，而且不便于计算机做进一步的结构分析和处理。在计算机处理化学结构的发展过程中，相继出现了碎片码、线形码、拓扑码等一些表达方法。其中碎片码和线形码都是将一些预先定义好的代表分子结构碎片的符号和数字线形排列起来，这两种表达方法的代表分别是 WLN 码[1]和 SMILES 码[2,3]，如图 1.1 中 2, 3-二甲苯胺的 WLN 码和 SMILES 码分别是 ZR B1 C1 和 Cc1c(cccc1c)N。碎片码和线形码表达的结构信息有紧凑和简明的优点，但同时有一个致命的缺点：难于实现计算机处理化学结构的基本操作——子结构检索。因为用户所提出的提问子结构是不可预测的，当提问子结构与预先定义的结构碎片不相吻合时，就无法得到检索结果。也就是说提问子结构只能限制在预先定义的结构碎片中，而化学结构的千变万化，使得这一限制导致了碎片码和线形码不能被大多数计算机结构处理系统所采用(SMILES 在网络传输中还有应用)。



WLN 码: ZR B1 C1

SMILES 码: Cc1c(cccc1c)N

图 1.1 2, 3-二甲苯胺的 WLN 码和 SMILES 码

拓扑码则利用了图论的知识，将化学结构图看作数学上的图，原子对应于图中的结点，化学键对应于图中的边，原子的不同类型(碳、氧、氮等)赋予结点不同的颜色，化学键的不同性质(单键、双键、三键、芳香键等)赋予边不同的颜色。最著名的拓扑码是美国化学文摘社采用的 Registry III 和法国学者 Dubois 创立的 DARC 系统。拓扑码虽然没有线形码所面临的子结构检索问题，但面临着一个同样重要的问题：惟一性问题。因为一个含 n 个原

子的结构可以有 $n!$ 种不同的编码方式。为了解决编码的惟一性需要采用一些算法，其中最有名的是 Morgan 算法[4]。

1.1.3 连接表

无论何种命名法和编码方法都是力图将分子的复杂的结构信息以简略的并且是惟一的方式表达，从这一点看，有些类似于分子的拓扑指数，是对分子结构信息的高度浓缩。而我们在利用计算机处理分子时，希望面对的结构信息越原始越详尽越好，因为计算机不仅需要处理查询等问题，而且要处理结构检索问题，以及在分子设计中经常用到的分子力学、分子动力学、量子力学等计算。因此，以命名或编码的方式表达分子的结构信息是远远不够的。连接表是目前表达分子结构信息的最完善和最计算机化的方式，它除了可以准确地表达分子的二维信息，还能利用原子的空间坐标精确地表达分子的三维结构信息。

连接表本质上是分子中所有原子、键及其空间关系（有时包括立体化学）的一个列表，在计算机上就是一个文本文件，通常包括原子信息部分（原子类型、电荷、坐标等），键信息部分（键的类型，连接原子等）和其他特殊的结构信息部分。连接表不需要考虑惟一性的问题，原子的序号不影响分子的结构信息，也不影响现有的结构处理算法对分子的识别。可以说连接表是一种非常自由的表达方式，随着计算机存储量和计算速度的飞速发展，连接表已经成为绝大多数化学结构处理系统的首选方式。

连接表有各种各样的格式，几乎每个分子模型化软件的推出都伴随着一个新的连接表格式的出现。比较著名的有化学结构协会(Chemical Structure Association)设计的SMD格式[5]，MDL信息系统公司(MDL Information System)推出的MOL格式[6]和由Tripos公司推出的MOL2格式[7]，其中MOL格式和MOL2格式由于相应公司推出的产品在世界各地比较普及，已成为普遍使用的文件格式。表1.1和表1.2分别是苯分子的MOL格式文件和MOL2格式文件。