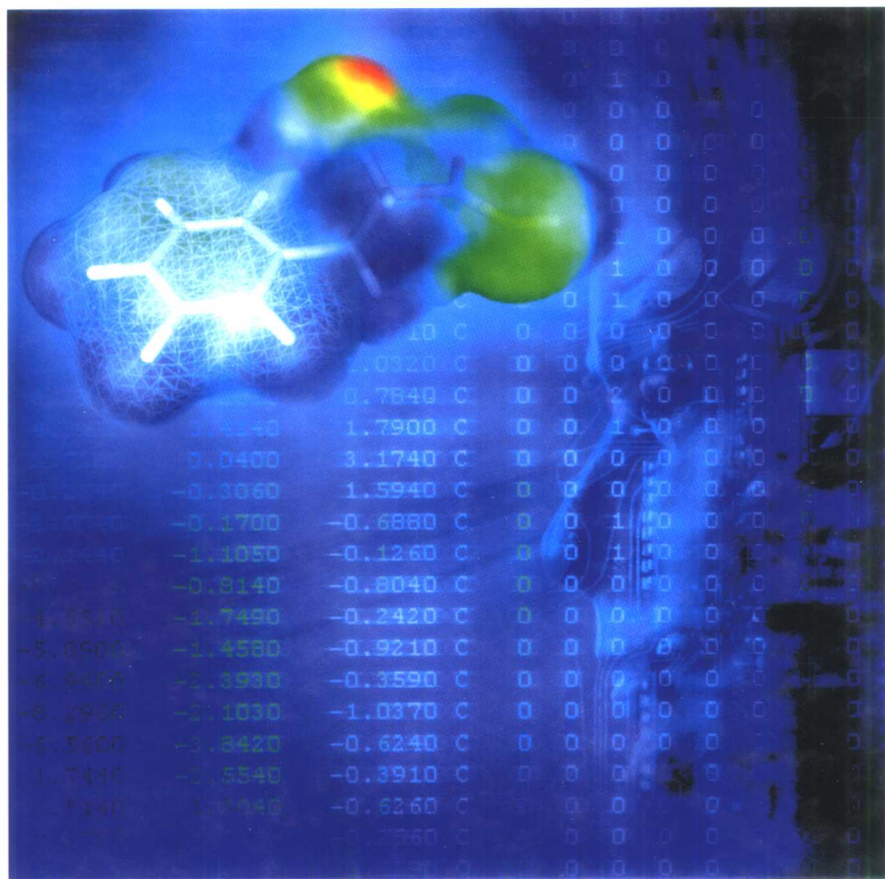


化学信息学教程

[德] 约翰·加斯泰格尔 托马斯·恩格尔 编著
梁逸曾 徐 峻 姚建华 等译



化学工业出版社
化学与应用化学出版中心

国外优秀化学著作译丛

化学信息学教程

[德] 约翰·加斯泰格尔 托马斯·恩格尔 编著
梁逸曾 徐 峻 姚建华 等译



化学工业出版社
化学与应用化学出版中心

· 北 京 ·

(京) 新登字 039 号

图书在版编目 (CIP) 数据

化学信息学教程/[德]加斯泰格尔 (Gasteiger, J.)
[德]恩格尔 (Engel, T.) 编著; 梁逸曾等译. —北京:
化学工业出版社, 2005. 1
书名原文: Chemoinformatics
ISBN 7-5025-6139-0

I. 化… II. ①加…②恩…③梁… III. 计算机
应用-化学-情报检索 IV. G252.7

中国版本图书馆CIP数据核字 (2004) 第 096681 号

Chemoinformatics/by Johann Gasteiger, Thomas Engel.
ISBN 3-527-30681-1
Copyright©2003 by Wiley-VCH. All rights reserved.
Authorized translation from the English language edition published by
Wiley-VCH.

本书中文简体字版由 Wiley-VCH 出版公司授权化学工业出版社独家出版发行。
未经许可, 不得以任何方式复制或抄袭本书的任何部分。

北京市版权局著作权合同登记号: 01-2004-2890

国外优秀化学著作译丛

化学信息学教程

[德] 约翰·加斯泰格尔 托马斯·恩格尔 编著

梁逸曾 徐峻 姚建华 等译

责任编辑: 梁虹

文字编辑: 李晓红

责任校对: 蒋宇

封面设计: 郑小红

*

化学工业出版社 出版发行

化学与应用化学出版中心

(北京市朝阳区惠新里3号 邮政编码 100029)

发行电话: (010) 64982530

<http://www.cip.com.cn>

*

新华书店北京发行所经销

大厂聚鑫印刷有限责任公司印刷

三河市延风装订厂装订

开本 720mm×1000mm 1/16 印张 35½ 字数 625 千字

2005 年 1 月第 1 版 2005 年 1 月北京第 1 次印刷

ISBN 7-5025-6139-0/O·67

定 价: 78.00 元

版权所有 违者必究

该书如有缺页、倒页、脱页者, 本社发行部负责退换

译者的话

此书的快速翻译出版产生于一个偶然。由于我们正在筹划着开一次“化学计量学和生物信息学的国际会议 (CCBA-2004, International Conference on Chemometrics and Bioinformatics in Asia)”。姚建华向我推荐了 Johann Gasteiger 教授。阅读了 Johann Gasteiger 教授的简历后,我为他对计算机应用于化学研究所做的出色工作所折服。当然,他应是我们大会所邀请的最佳对象。我当即给他发出了大会的邀请,并为他能欣然接受大会的邀请来中国而感到由衷高兴。没想到的是,几天后又惊喜地收到了 Johann Gasteiger 教授寄来他刚出版的新作——化学信息学教程 (Chemoinformatics——A Textbook)。粗粗浏览此书后,我脑海马上萌生出要将其尽快介绍给中国同行的念头。于是,我急忙与我们有良好合作关系的化学工业出版社化学与应用化学出版中心梁虹通话,询问她是否有翻译出版这本书的意愿,回答是肯定的。我将此书推荐给了我的恩师俞汝勤院士,并请求他和我一起来写出版推荐。俞汝勤院士慎重地阅读此书后,欣然应允。正好,当时已移居美国的徐峻博士 (现在美国 Discovery Partners International, 药物公司任化学信息部主任) 应邀来我校讲学,我与他谈及此事时方知就是他向 Gasteiger 教授推荐了我们研究小组来翻译此书,故有我收到此书的故事。徐峻博士是 Gasteiger 教授的好朋友,并向我推荐了几名合适的译作者。

经过版权洽谈,化学工业出版社顺利地获得原出版公司——Wiley-VCH 的授权。为把本书尽快介绍到中国来, Gasteiger 教授亦与 Wiley VCH 出版公司进行了联系,才使得这本书得以快速翻译出版。在此,我们谨向化学工业出版社和原作者致以衷心的感谢。

我们组成的一个翻译小组实际是一个广泛的合作小组。本书大部分翻译由中南大学化学计量学研究小组成员完成,梁逸曾从事前言、第 1 章和第 11 章的翻译,沈菊文、刘伟合译第 2 章,向征翻译第 8 章,胡黔南博士翻译第 9 章,附录和索引部分则由梁逸曾,刘伟等共同翻译;徐峻博士翻译第 6 章和第 10 章中有关药物设计的应用部分;廖晨钟博士翻译第 7 章及第 10 章的分子模拟的应用部分;第 3 章、第 4 章、第 5 章以及第 10 章中相应的应用部分由中国科学院上海有机化学研究所计算机化学实验室译出。在此值得提出的是,原

先这些翻译部分很多应由袁身刚先生负责，可由于袁身刚先生的溘然逝世（在此谨表我们对袁身刚先生的深切悼念！），使得这些部分的翻译皆由姚建华教授负责承担。另外，全书所有章节的审校工作皆由梁逸曾教授承担。可以说，此书的翻译出版是一个集体努力的结果。

化学信息学（Cheminformatics，又称 Chemoinformatics 或 Chemo-informatics）与它的姊妹学科生物信息学（Bioinformatics）一样，是近年来受到广泛关注的一门横跨化学、信息科学、数学和统计学的边沿学科。它始于 20 世纪 60 年代以后与计算机有关的各个化学学科，例如计算化学（Computational Chemistry）、计算机化学（Computer Chemistry），化学计量学（Chemometrics），结构与活性关系（QSAR），化学信息（Chemical information）等学科。化学信息学研究如何运用计算机技术来处理化学数据。1961 年，美国化学会出版了化学文献杂志（the Journal of Chemical Documentation），该杂志于 1974 年更名为化学信息与计算机科学杂志（the Journal of Chemical Information & Computer Sciences）以与飞速发展的学科现状相适应。20 世纪 80 年代后期，大规模的化学、生物学实验技术和高效有机化学合成技术的成熟，对化学信息处理提出更高和更迫切的要求。人们发现了处理化学信息的许多新概念、新领域和新方法，于是，化学信息学的提法被广泛接受。今天，化学信息学已经成为药物研究和生命科学各领域不可或缺的热门工具。

德国爱尔朗根-纽伦堡大学的 Johann Gasteiger 教授是一位在国际上享有盛誉的化学信息学专家，他在本领域从事科研和教学长达 30 多年。Gasteiger 教授的研究兴趣宽广，从计算分子结构电荷分布的经验模型、药物设计、化学反应模拟、有机合成设计、小分子二维和三维结构产生、到各种基于化学结构的性质预测、用于搜索巨大数据空间的遗传算法和各种机器学习算法，他和他的合作者们都多有建树。考虑到化学和生命科学对化学信息学人才的需要，Gasteiger 教授和他的同事 Thomas Engel 博士于 2003 年组织化学信息学领域的专家编写了《化学信息学教程》一书。该书系统地介绍了化学信息学的各个研究方向，深入浅出地评述了化学信息学各领域的现状和今后的发展动向。是一本既适合作为大学高年级学生和研究生教科书，又可以作为化学界和制药界科技人员自修用的参考书。因此，该书英文版问世之后，很快被国际上许多大学列为化学信息学的主要教材。我们将其翻译介绍出来，希望能为我国化学信息学的发展做出微薄贡献。

梁逸曾 徐峻

2004 年 8 月

序

从数学的观点说来，两个半只鸡应等于一只鸡。单从质量的观点来看，这是正确的。然而，这样的鸡绝不是一只活鸡。更进一步，这两个半只鸡还应是镜像对称的才行。很明显，数字还只讲了一半事实。信息以全然不相干的外观进入我们的生活。化学，一门研究物质及其转换的科学，展现了一种很宽广的信息的多样性，至今它已成为包含了有关各种物质的化学结构、性质和反应的巨大知识体。然而，尽管在前两个世纪中，化学研究所取得的成就已使之从早期的化学技艺转化成为一门复杂的自然科学，但化学还仍然是一门缺乏全面的理论基础的学科，如准确预计结构与活性或结构与功能关系的理论。例如，为什么于1979年从一种夏威夷 coral 珊瑚分离出来的 Palytoxin ($C_{129}H_{223}N_2O_{54}$) 会成为一种最毒的天然产物。它的64个手性碳和6个烯键可给出 10^{21} 种可能的异构体，如将 Palytoxin 酸全合成出来，其困难程度将可与首次攀登 Everest 山相比（1989年 Kishi 等在哈佛大学的演讲）。

目前，如果我们想消化和掌握化学学科多年累积的，以及现在有如潮水般涌来的化学信息和知识，我们将必须付出相当大的精力。在1960年以前，化合物数目的增长，包括天然发现的和实验室合成的，还是处于一种线性积累的状态，在大约150年时间里增加到100万；然而，在此之后，则是以指数增长的速度在增长，到2000年，已达到1800万。这也正是自1965年由计算机技术的发明为化学学科带来的革命性的变化之一。基于物理、数学和信息科学的技术和方法进入化学研究也达到前所未闻的一种程度，此中既包括化学实验室中出现的各种新仪器，也包括大量的几十年来就认为几乎是不可能的那些基于数学模型或量子力学的计算。例如，利用计算机来对水分子的膜转运进行模拟，通过 aquaporins 作为界质，可在一皮秒内产生以10个纳秒最小标度的 10^5 个原子的空间位置的时间依赖曲线（Grubmmuller 等，MPI Gottingen）。通过计算机网络可在这样巨大的数据阵中进行数据搜索和处理，然后还可可在其中通过前后数据来进行评价（“数据发掘”）。另一方面，在化学技术中还产生了组合合成，可很快产生大量数据。总的来说，尤其对于化学和制药行业，研究者用于分析和处理数据的时间已超过产生这些数据的时间。然而，在十几年

前则还是处于一个相反的境地。

在 20 世纪 70 年代, 化学家们进行研究时, 就越来越多地碰到不断变化的“化学-信息-计算机”(Chemistry-Information-Computer, CIC) 三头统治的现象。即采用计算机和信息技术来产生数据, 将不同数据源组合起来, 进而将数据转成信息, 再将信息转化成知识, 以达到解决化学问题的最终目的。例如, 我们在进行有机合成的规划和设计, 药物设计和结构解析等的研究就正是这样。这就导致了一个新的化学专门技术的产生, 它已与在 200 多年前建立的, 包含首先是杂志、然后二级文献, 再查阅化学文摘等检索系统的传统的化学档案式的化学信息获取方法, 已有了明显不同的特征。

在 20 世纪 80 年代, 计算机网络的出现又为跨越任一距离的信息快速流开辟了新纪元。开始时, 它们的重要性还没得到化学界的充分认识, 这一情况可用 Karl Valentin 最近的一段话来表述: “计算机网络就是人们不想在需要时才想拥有, 而是简单地必须想拥有的一种工具, 因为人们总是可能在需要使用它的状态之中。”(A computer network is something that one does want to be in the need to have, nevertheless simply must want to have, because one always might be in need to use it.) 1986 年 Johann Gasteiger 在慕尼黑共同组织了德国化学会 (GDCh) 的 Task Force CIC。在同一年内, 他开始了 CIC 的有关化学中计算机软件开发研讨班, 受到了一致的认同和接受。直到今天, 这些年度会议已成为报告在不同 CIC 领域的最新成果, 也包括化学信息系统的展示和传播的讨论场所。Task Force CIC 与德国化学会 (GDCh) 的化学信息分部结合, 成为现在的德国化学会 (GDCh) 的“化学-信息-计算机”(Chemistry-Information-Computer, CIC) 分部。

20 世纪 90 年代因特网的出现, 更是刺激了计算机网络在化学中的应用。结果就使得在进入 21 世纪的时刻, 有长期从事化学与计算机科学交叉研究的先行者的成果正式被广泛认识和接受, 而形成了一门完整的新的交叉学科: “化学信息学”。它将包含化学信息的设计、创造、组织、处理、检索、分析、传播和使用 (G. Paris)。作为它的孪生姐妹, 更早一点发展的生物信息学, 主要研究基因和蛋白质, 而化学信息学则主要是处理小分子问题。然而, 在研究小分子如何与蛋白质进行相互作用时, 它们之间的区别就变得模糊起来。从生命科学的观点来看, 它们之间的边界线就完全没法划分。

一个年轻的科学分支是与它的学生们一起成长的。作为报答, 学生应该付股金。工业就业对化学信息学家的要求越来越高, 所以这个领域也将期望成为很好的产业。这样, 怎样和在哪里能学到化学信息学就将成为一个问题。一个好地方当然是去 Erlangen。Johann Gasteiger 和他的研究小组甚至是在不知该

研究领域怎么命名的情况下，进行化学信息学方面的研究 25 年。1991 年，Gasteiger 教授获得了德国化学会的 Gmelin Beilstein 奖章，由于他在化学信息学方面（CIC）的成就，于 1997 年又获得了美国化学会的化学信息分部的 Herman Skolnik 奖。由于这本面向本科生和研究生的，全面而又深刻的化学信息学教科书的出现，Gasteiger 教授和他的研究小组为建立起更多的“Erlangen”提供了可靠的基石。我热忱地向我的学术同事们推荐这本书。现在应该是各大学开始在一个较广范围内进行化学信息学教育的好时机。它可以是化学学习中的一门学位课程，如本科毕业后的硕士研究生的学位课，也可以像生物信息学那样，成为系列的化学信息学课程。无论是哪种情况，这本书以及它的附加材料将为教学或自学提供适当的基础。

Dieter Ziesow

德国化学会（GDCh）化学-信息-计算机（CIC）分会主席

前 言

计算机已进入了我们日常生活的每一个方面。很清楚，科学家和工程师们在将计算机用于解决实际问题的这个过程中起到了一个领导作用。在化学研究中，化学家们很早就认识到，大量数据的处理只能依靠电子设施，将数据存储于数据库中来完成。只有通过这种方式，大量现存的化合物（3500 万）的数据才能被处理。这样，早在 20 世纪 60 年代，有关化学数据库的工作就已启动。更进一步，有关化合物结构与其物理、化学或者是生物性质之间的关系是非常复杂的，这就要求有高度复杂的计算，或是要求在对相应数据进行分析后进而对这些性质进行预测。很多国家的化学协会都意识到计算机在各自领域的重要性，并成立了相应的分会（分部），这些分会的主要任务就是怎样将计算机用于化学研究。

然而，从一开始，在为进行量子力学计算而用计算机的理论化学家和采用计算机来进行信息处理和数据分析的化学家之间就存在着一条鸿沟。在美国，就存在着两个分部，即计算化学分部和化学信息分部。在德国，存在着理论化学分部从属于“Deutsche Bunsen-Gesellschaft für physikalische Chemie”和化学信息分部从属于“Gesellschaft Deutsche Chemie”。事实上，在 1989 年，化学信息分部更名为化学-信息-计算机分部（Chemie-information-computer, CIC），以承认采用计算机来处理化学信息的重要性。在这一分部工作的科学家非常积极进行着计算机在化学中应用研究的推广工作。在 1987 年，组成了两个研讨班（workshop），一个是化学软件的开发，另一个是分子模拟，从那以后这两个研讨班就每年都进行一次。

另一方面，德国联邦研究技术部（German Federal Minister of Research and Technology, BMFT 后改名为 BMBF）于 20 世纪 80 年代启动一个资助所谓的科学信息中心（Fachinformationszentren, FIZ）的项目，而且，也开始建立数据库。化学家们都十分庆幸政治家和专家们认识到了化学数据库的重要性。这样，一些目前在国际上公认的知名数据库也开始启动，如 Beilstein 有机化合物数据库；Gmelin 无机和金属无机化合物数据库；Chem Inform RX 反应数据库；以及 SpecInfo 波谱数据库。

然而，尽管这些事件的发生，我们还必须认识到化学家们还只是逐步在接受“计算机是一种很需要的、日常工作的工具”的说法。但是，他们逐渐地（或勉强地）在接受这样一种说法：我们在经常使用数据库来检索信息，我们在经常使用量子化学或是分子力学的程序计算（大多数是事后的），以增加我们对化学现象的理解。更进一步，随着组合化学和高通量筛选的出现，我们可以清楚地看到，由这些技术所产生出来的潮水般的信息是只能由计算机方可处理的。

这样，计算机将继续深入化学的各个方面，我们也必须为我们的下一代化学家作出准备，使他们能尽快地熟悉这些计算机方法。实际上，我们也将看到，为解决化学研究中出现的问题，在化学中，不同的计算机应用方法正在不断地增加。所以，从化学整体的观点来对这些方法进行梳理是有必要的。将在化学中应用不同的计算机方法归类并整理，并将它们组成一门自身的分支学科，化学信息学。

在这本教科书中，我们将化学信息学进行了全面总结，并将它们集成起来成为课程，可以作为化学信息学整个课程的一个基础。

这本书所包含的内容可以说是建立在我们研究小组 25 年来研究与发展的成果之上的。所以，首先我得感谢我所有的，现在和过去的同事，感谢他们与我一道共同在这一新的研究领域进行的冒险和探索。事实上，这本教科书完全是由我们研究小组的成员完成。这样，我们可以方便地对全书各章进行调整，以达到使全书在内容取舍方面达到平衡，并尽量展示整个化学信息学领域的成果。然而，每个作者的个人表达的写作方式和特点在此过程中是不可能完全消失的，我们希望这可能还能让读者在阅读和学习过程中体会到某些生动的经验。将此书的写作置于他们日常工作的首位，无论如何将是一项艰巨的任务。所以，在此我必须感谢他们与我一道完成这一旅行。

我们还同样要感谢德国联邦教育和研究部（German Federal Minister of Education and Research, BMBF）对由科学信息中心（FIZ）化学部组织的“化学中的网络教育”的资助，就是在此项目中，我们为化学信息学发展了网络学习（eLearning）的工具。

此外，我们感谢 Wiley-VCH 出版社的 Gudrun Walter 博士在完成此书过程中对我们的鼓励，感谢 Romy Kirsten 博士在处理此书编辑中的愉快合作。

我们谨希望这本教科书将引起广大读者对化学信息学的兴趣，并希望此书能使他们为此研究领域而感到激动，就像我们经常为之激动一样。

Johann Gasteiger

2003 年 5 月，于 Erlangen

内 容 提 要

本书系统介绍了化学信息学的各个研究方向，深入浅出地评述了化学信息学各领域的现状和今后的发展动向。本书重点阐述了化学分子结构及反应的表征，数据类型与数据库，检索方法，数据分析方法，以及化学信息学在结构解析，反应模拟，合成与药物设计方面的应用。同时还介绍了常用计算机软件与因特网资源。

适用于化学系高年级本科生和研究生教材，同时也可从事化学和药学的科研人员提供参考。

目 录

第 1 章 导言	1
1.1 化学的研究领域	1
1.2 化学家的基本问题	2
1.3 化学信息学的范畴	4
1.4 化学信息学的学习	6
1.5 主要任务	8
1.5.1 目标物的表征	8
1.5.2 数据	8
1.5.3 学习	9
1.6 化学信息学的历史	9
1.6.1 结构数据库	9
1.6.2 定量结构与活性关系	10
1.6.3 分子模拟	10
1.6.4 结构解析	10
1.6.5 化学反应与合成设计	10
1.7 本书的内涵	11
1.8 化学信息学的教学	12
第 2 章 化合物的表示法	13
2.1 引言	13
2.2 化学命名	15
2.2.1 化学命名的发展	16
2.2.2 化学元素的表达	16
2.2.2.1 元素的描述	16
2.2.3 (无机)化合物的经验式的描述	16
2.2.3.1 现行表示法	17
2.2.4 有机化合物的经验式表示法	17
2.2.4.1 现行表示法	17

2.2.5 无机和有机化合物的系统命名	17
2.3 线性符号表示法	19
2.3.1 Wiswesser 线性符号表示法	19
2.3.1.1 应用	21
2.3.2 ROSDAL	21
2.3.2.1 应用	22
2.3.3 SMILES 编码	22
2.3.3.1 应用	24
2.3.4 Sybyl 线性符号表示法	24
2.3.4.1 应用	25
2.4 结构的编码	26
2.4.1 图论	26
2.4.1.1 图论基础	27
2.4.2 矩阵表达	29
2.4.2.1 邻接矩阵	29
2.4.2.2 距离矩阵	30
2.4.2.3 原子连接矩阵	30
2.4.2.4 关联矩阵	30
2.4.2.5 键矩阵	30
2.4.2.6 键-电子矩阵	31
2.4.3 连接表	33
2.4.4 化学结构的输入与输出	35
2.4.5 标准的结构交换格式	37
2.4.6 指南: Molfile 和 SDfile	39
2.4.6.1 Molfile 的结构	40
2.4.6.2 SDfile 的结构	43
2.4.6.3 库和工具箱	43
2.5 结构信息的表达	45
2.5.1 环识别	45
2.5.1.1 最小环数目	46
2.5.1.2 所有的环	46
2.5.1.3 最小基本环	47
2.5.2 表达的无歧义性和惟一性	48
2.5.2.1 结构异构体和同构	48

2.5.2.2 标准化	49
2.5.3 摩根算法	50
2.5.3.1 指南：摩根算法	51
2.6 连接表的不足	53
2.6.1 用连接表来表达分子结构的不足	53
2.6.2 用电子体系表达分子结构	54
2.6.2.1 基本概念	54
2.6.2.2 简单的单键和双键	54
2.6.2.3 共轭性和芳香性	54
2.6.2.4 π 体系的正交性	56
2.6.2.5 非键轨道	56
2.6.2.6 带电物质和自由基	56
2.6.2.7 电离状态	57
2.6.2.8 缺电子化合物	57
2.6.2.9 有机金属化合物	57
2.6.3 从 VB 表达中产生 RAMSES	58
2.7 化合物结构的特殊标记	58
2.7.1 Markush 结构	59
2.7.2 碎片编码	59
2.7.2.1 应用	60
2.7.3 指纹	60
2.7.3.1 Hashed 指纹	60
2.7.4 Hash 编码	61
2.7.4.1 应用	62
2.8 立体化学的表征	62
2.8.1 基本概念	62
2.8.2 分子手性和构型异构体的表征	64
2.8.2.1 手性的探测与表征	66
2.8.3 排序列表	67
2.8.4 旋转序列	68
2.8.5 置换说明	68
2.8.6 Molfile 和 SMILES 中的立体化学	69
2.8.6.1 Molfile 中的立体化学	70
2.8.6.2 SMILES 中的立体化学	71

2.8.7 指南：通过置换基团来解决立体化学问题	72
2.8.7.1 四价碳的立体化学	73
2.8.7.2 双键的立体化学	75
2.9 三维结构的表征	77
2.9.1 分层介绍化学结构表征	77
2.9.2 三维结构的表征	78
2.9.3 为什么需要三维结构以及怎样获得三维结构	80
2.9.4 三维结构的自动生成	82
2.9.5 获取所有构象：何谓构象分析？	88
2.9.6 全体构象的自动生成	89
2.9.7 指南：三维结构编码（PDB，STAR，CIF，mmCIF）	95
2.9.7.1 简介	95
2.9.7.2 PDB 文件格式	96
2.9.7.3 STAR 文件格式和字典	102
2.9.7.4 晶体信息文件（CCDC）	103
2.9.7.5 mmCIF 文件格式	105
2.9.7.6 软件	105
2.10 分子表面	106
2.10.1 范德华表面	108
2.10.2 Connolly 表面	109
2.10.3 溶剂可及表面	109
2.10.4 溶剂排斥表面	110
2.10.5 酶腔表面（Enzyme Cavity Surface，联合表面）	110
2.10.6 基于等值的电子密度表面	110
2.10.7 用实验方法获取表面	111
2.11 分子模型可视化	111
2.11.1 历史回顾	111
2.11.2 结构模型	113
2.11.2.1 金属线结构模型	113
2.11.2.2 帽棍模型	113
2.11.2.3 球棍模型	113
2.11.2.4 空间填充模型	114
2.11.3 生物大分子模型	115
2.11.3.1 圆柱体模型	115

2.11.3.2	带状模型	116
2.11.3.3	管状模型	116
2.11.4	晶体图模型 (Crystallographic Model)	116
2.11.5	分子属性的可视化	116
2.11.5.1	基于等表面的性质 (Properties Based on Isosurfaces)	116
2.12	工具: 化学结构绘制软件——分子编辑软件与显示软件	118
2.12.1	概述	118
2.12.2	分子编辑器	119
2.12.2.1	单机应用	120
2.12.2.2	基于网页 (web) 的应用	124
2.12.3	分子显示器 (Molecule Viewers)	127
2.12.3.1	单机应用	127
2.12.3.2	基于网页 (web) 的应用	129
2.13	基于网页 (Web) 的三维结构产生工具	137
第3章	化学反应表达	148
3.1	引言	148
3.2	反应类型	151
3.3	反应中心	151
3.4	化学反应度	153
3.4.1	物化效应	153
3.4.1.1	电荷分布	154
3.4.1.2	诱导效应	155
3.4.1.3	共振效应	155
3.4.1.4	极化效应	155
3.4.1.5	立体效应	155
3.4.1.6	立体电子效应	156
3.4.2	化学反应度定量的简单方法	156
3.4.2.1	前沿分子轨道理论	156
3.4.2.2	线性自由能关系 (Linear Free Energy Relationships, LFER)	157
3.4.2.3	经验反应度公式	159
3.5	反应分类	159
3.5.1	模型驱动方法	160

3.5.1.1	Hendrickson 方案	160
3.5.1.2	Ugi 方法	161
3.5.1.3	信息化学反应分类	166
3.5.2	数据驱动的方法	166
3.5.2.1	HORACE	166
3.5.2.2	反应前景	167
3.6	反应的立体化学	170
3.7	指南: 反应的立体化学	170
第4章	数据	175
4.1	引言	175
4.1.1	数据、信息和知识	175
4.1.2	数据采集途径	176
4.2	数据采集	177
4.2.1	数据质量的重要性	177
4.2.2	数据复杂度	179
4.2.3	实验数据	179
4.2.4	数据交换	180
4.2.4.1	DAT 文件	180
4.2.4.2	JCAMP-DX 格式	181
4.2.4.3	PMML	182
4.2.5	实际数据和它们潜在的不足	183
4.3	数据预处理	184
4.3.1	中心化、定标度和自标度	184
4.3.2	高级方法	186
4.3.2.1	快速傅里叶变换 (Fast Fourier Transformation, FFT)	186
4.3.2.2	小波变换 (Wavelet transformation, WT)	187
4.3.2.3	奇异值分解 (Singular Value Decomposition, SVD)	187
4.3.3	变量选择	188
4.3.3.1	基于遗传算法 (genetic algorithm, GA) 的方法	188
4.3.3.2	基于正交化 (orthogonalization) 的方法	189
4.3.3.3	基于模拟退火 (Simulated Annealing, SA) 的方法	190
4.3.3.4	基于主成分分析 (principal component analysis, PCA) 的方法	190