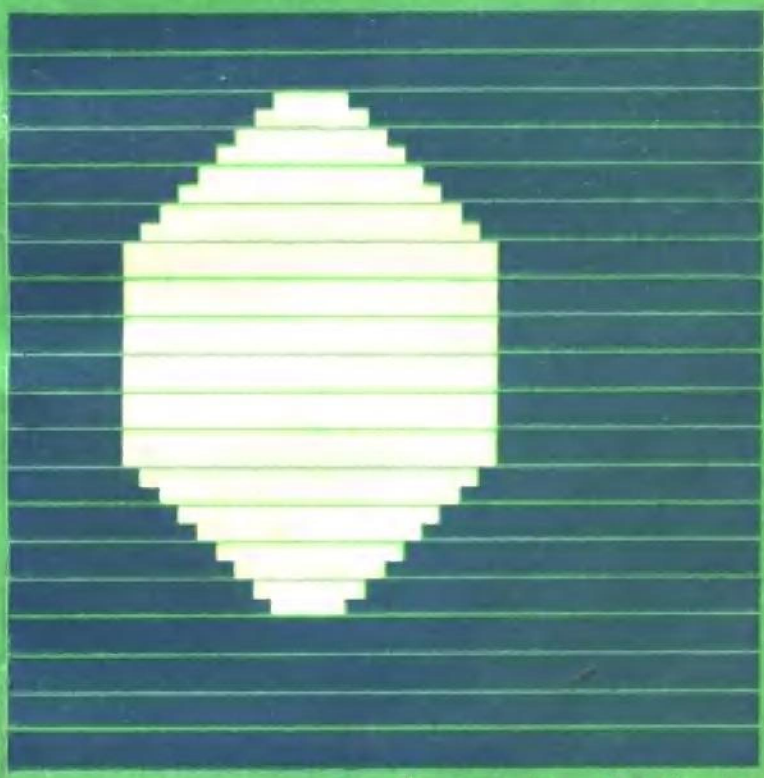
计算机化学

方法及应用

许禄　郭传杰　编著

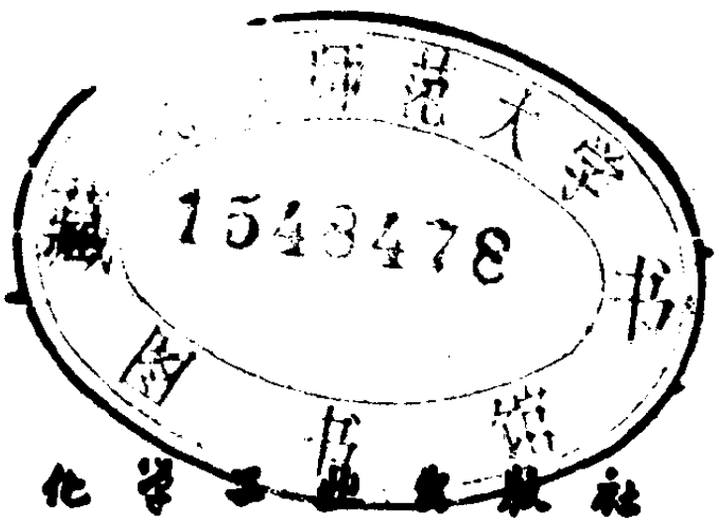


化学工业出版社

计算机化学

方法及应用

许 禄 郭传杰 编著



化学工业出版社

内 容 提 要

本书系统叙述了浩繁化学信息的计算机处理技术，概括了计算机化学，尤其是计算机辅助结构解析等方面的现况。内容主要包括：化合物结构的编码、输入、输出和检索；有机化合物的主要谱图手段；人工智能和专家系统在有机结构解析中的应用；化合物的反应信息及计算机辅助合成路线方案设计；结构-活性相关性研究；文献型数据库和联机检索系统及整个学科的发展趋势和去向。本书在大量文献的基础上，结合我国的研究工作编著而成。书后附有主要程序系统名称缩写一览表。

本书对象为化学和计算机两大领域中的应用开发、研究科技人员及管理人员，也可供有关高等院校的师生学习、参考。

计 算 机 化 学
方 法 及 应 用
许禄 郭传杰 编著

责任编辑：夏叶清
封面设计：郑小红

化学工业出版社出版发行
（北京和平里七区十六号楼）
化学工业出版社印刷厂印刷
豆各庄装订厂装订
新华书店北京发行所经销

开本787×1092 1/32印张 10 3/4 字数 245 千字
1990年10月 第1版 1990年10月 北京第1次印刷
印 数 1—2030
ISBN 7-5025-0736-1/TQ·419
定 价6.40元

前　言

计算机的应用日渐广泛，其应用水平也日渐提高，特别是微计算机的大量涌现，使计算机的应用得以极大普及。目前，科学家们正在潜心于第五代计算机的研究，一旦这一“坚冰”被打破，它必将对于科学（当然包括化学）的发展产生更为深刻的影响。

计算机的应用可粗略地划归为两类，即数值方法和非数值方法。早期应用主要为数值方法。尔后，非数值方法发展异常迅速，如科学数据库、化合物结构解析中的人工智能和分子设计等。

在我国化学界，从70年代末期开始了计算机辅助结构解析的研究工作，特别是和大型分析测试仪器相关联的谱图库，如质谱、红外光谱和核磁共振波谱等，在80年代中期相继建成。人工智能中的专家系统日益普遍地发展起来。将模式识别方法用于化合物的分类及结构-活性相关性研究，引起了化学家们的广泛兴趣。因而，目前能够提供一本有关方面的参考书是很有意义的。为此，我们根据近年来的研究工作，并且查阅了大量的文献，写成了本书。

全书共九章，比较系统地介绍了化合物的结构编码（第二章），结构的输入、输出和结构的检索（第三章），有机化合物的主要谱图手段（第四章）。第五、六两章分别介绍了属于人工智能范畴的模式识别和专家系统在有机化合物结构解析中的应用。第七章为化合物的反应信息及计算机辅助合成路线方案设计。

第八章为目前在国际上开展得甚为广泛的结构-活性（或性质）相关性研究。另外，概述了文献型数据库、联机检索系统（第一章）及值得今后重视的几个发展方向（第九章）。书中各个部分有一定联系，但又基本独立成章。

本书在有限的篇幅内，力图使读者对计算机化学（特别是计算机辅助结构解析等）的概貌有一了解，因而，在某些章节（如专家系统）的叙述中，并没有着力去对某个单一系统作详尽剖析。书中所涉及的数值方法，仅着重介绍方法的基本概念和结论，而未作详细推导和证明。

清华大学化学系胡鑫尧副教授对全书进行了审阅。同时，在该书编写、出版过程中，得到了中国科学院长春应用化学研究所和北京化学所许多同志的热情支持和帮助，特别是化工出版社黄志学编审、夏叶清编辑付出了大量辛勤劳动，在此一并致谢。

由于作者学识浅陋，缺点和错误在所难免，敬请读者不吝赐教、批评和指正。

作者

一九八八年十二月

目　　录

第一章　化学信息与计算机

1.1　化学信息的产生及计算机处理的必要性

在中世纪，炼丹术士的秘方是口授的，即师傅将配方和技艺传给他的徒弟。后来，开始以插图或图形方法代替口传，到15世纪出现了第一本教科书，于是，诞生了化学。化学信息的传输就是以教科书这种初级形式进行的，并且由古一直延续至今。所不同的是，在古代的化学书籍中没有理论，也无化学名称，而仅以简略的符号来表示化合物。在化学科学的发展进程中，随着Avogadro和Boyle气体定律的发现以及Dalton、Berzelius和Cannizzaro的近代结构理论的确立而首次产生了“化学结构”一词，从而，可以区分于非结构的化学信息。

从19世纪开始，科学技术的飞速发展引起了信息的大爆炸，Price[1]曾对广阔的学科领域中信息增长的速率进行了研究，观察了不同的数据增长模式。当一项新兴领域出现时，信息增长率就加大，如液晶的出现即为一例。再如干扰素，在最初的10个月内，其文献量成倍地增长（图1.1）[2]。

在过去40多年中，化学文献每年都约以5倍的速度增加；同时，新实验技术所产生的信息量也极其浩繁。而且，近代科学的显著特点是学科之间相互渗透。如生物学家，不仅需要熟知生物化学，而且需要了解和掌握化学、医学及物理学等；物理学家，不仅需要熟知物理学，而且需要了解和掌握数学、化学及工程科学等。同样，化学家不仅需要熟知化学，而且需要

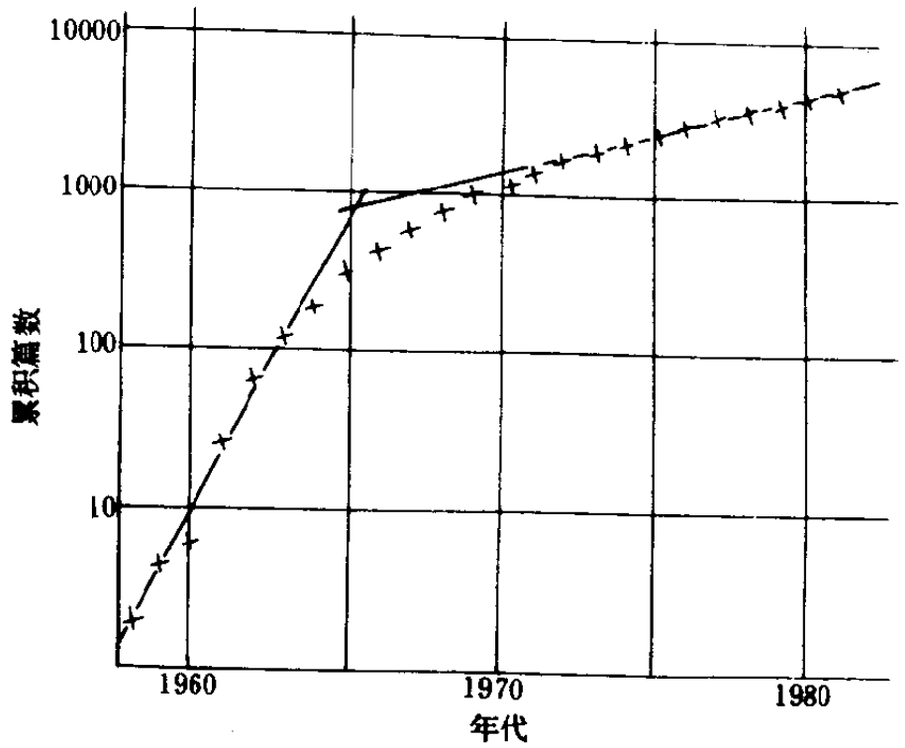


图 1.1　干扰素文献增长情况

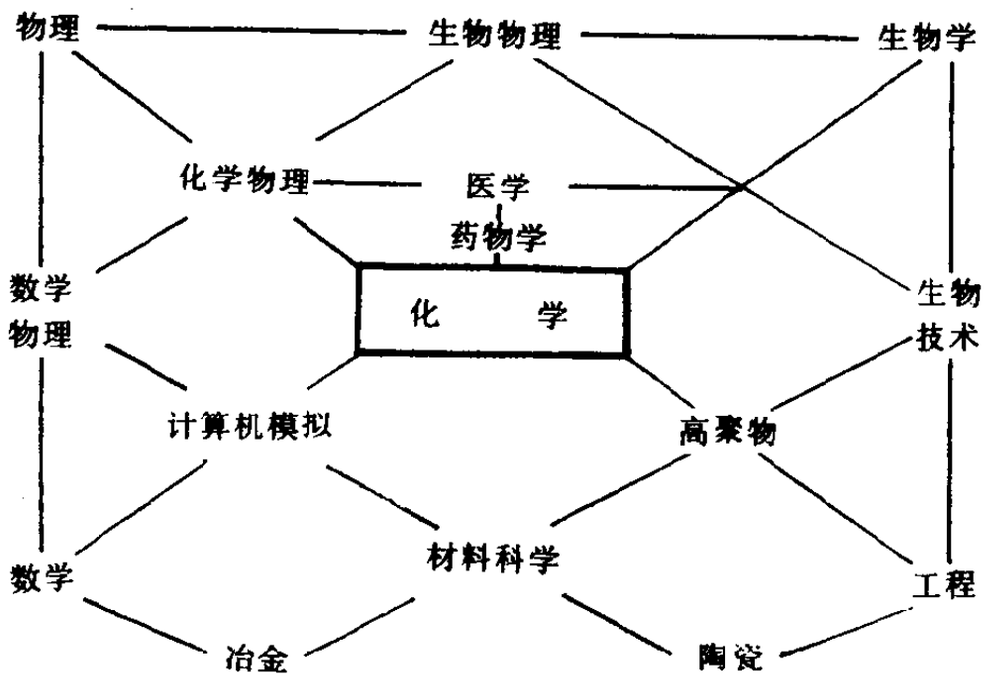


图 1.2　近代化学与其他学科的相关图

了解和掌握生物学、医学及物理学等等。在近代科学中，化学与其他学科领域的相关网络示于图1.2。由图可见，化学家为了进行科学研究，所需要的信息范围甚广，所涉猎的学科范围甚宽。因而，在信息时代中，对化学若仅仅以传统方式进行数据的处理，则会与实际需要极不适应，由此，客观上提出了应用计算机技术的要求。于是，化学必然地与计算机结下不解之缘。运用计算机进行化学信息的管理是势在必行。

1.2 数值型数据库

60年代初，世界上一些发达国家开始应用计算机进行数据的存贮和检索，实现了现代化的信息管理。随着计算机硬件及软件系统的发展，从70年代以来，数据库有了突飞猛进的发展。据统计，世界上商用数据库的增长情况为：1975年约390个，1978年为1100多个，1982年达到1800多个。

数据库可分为目录型、数值型和管理型三类。目录型，如文献检索；管理型，如人事档案材料；数值型主要是科学数据库。化学中数值型数据库又可粗略地分成二类：（1）结构数据的数据库；（2）非结构数据的数据库。前者，是与物质化学结构有关的信息，如质谱、核磁和红外光谱等各种谱图数据，以及化合物结构本身的相关数据等。后者包括各种物质的物化参数、热力学性质、相平衡以及传递过程方面的数据等。有关谱图的数据库将在第四章予以介绍；结构的编码、存贮及检索将在第三章予以讨论，而本节仅拟对数值型化学数据库做一概略性描述。

1.2.1 现代化学数据库的特征

数值型化学数据库有如下几个主要特征：

1. 拥有大量的原始数据。我们称之为一级数据，它即是一

般意义上所理解的数据源。例如，目前的谱图库中碳-13已达50000多张谱图，质谱达到约70000张，红外光谱达100000多张；美国NIH/EPA的毒性库所包含的化合物有20多万个。

2．在一级数据的基础上，可以衍生二级、三级数据。如由碳-13谱图系数，即化学位移和多重性，可派生出化学位移-子结构相关表（表1.1），用于化合物结构的人工智能解析。

表1.1 化学位移-子结构相关表

编码	结构	^{1}HNMR	^{13}C NMR
1	CH_3	84.0 - 60.0	26.02 - 31.13
2	CH_3-C-	96.0 - 72.0	24.47 - 33.57
3	CH_3	90.0 - 60.0	25.48 - 34.04
⋮	⋮	⋮ ⋮	⋮ ⋮
10	CH_3	90.0 - 48.0	14.72 - 36.27
11	$-C-$	90.0 - 48.0	10.12 - 36.27
12	CH_3	89.4 - 36.0	6.80 - 32.61
⋮	⋮	⋮ ⋮	⋮ ⋮
106		220.0 - 108.0	20.46 - 56.42
107	$-CH_2-$	162.0 - 102.0	18.88 - 53.09
108		180.0 - 114.0	13.27 - 54.14
109		144.0 - 0.0	14.66 - 57.16
⋮	⋮	⋮ ⋮	⋮ ⋮
153	$-O-$	— —	— —
⋮	⋮	⋮ ⋮	⋮ ⋮

3．向智能化方向扩展。数值型数据库总是要在一级或二级信息检索的基础上，进一步扩展功能，使其具备一定智能，即过渡到知识库和专家系统的范畴。如利用萃取分配比、萃取离子的浓度和溶液的酸度等基本数据进行三维空间中响应曲面

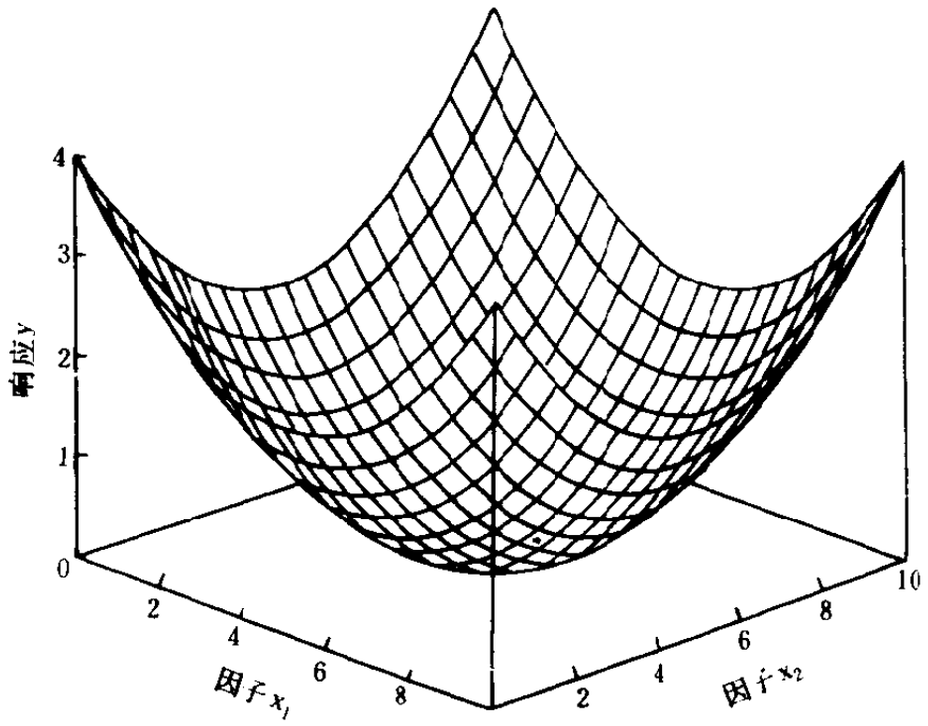


图 1.3　两因子响应曲面

$$Y = 4 - 0.8x_1 - 0.8x_2 + 0.08x_1^2 + 0.08x_2^2$$

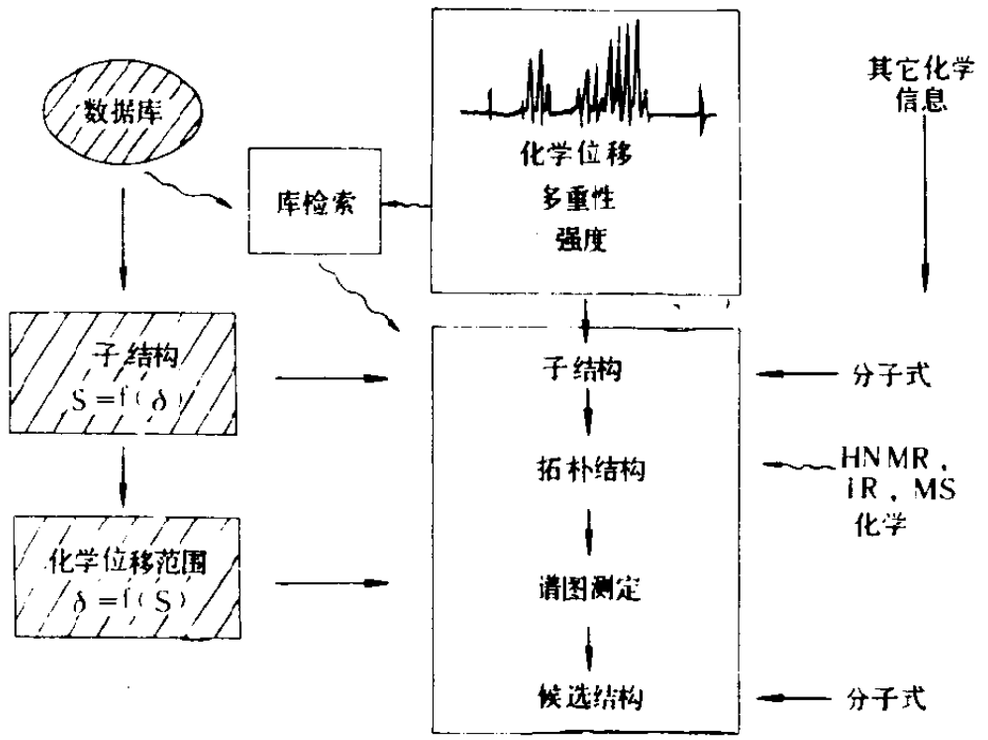


图 1.4　BASF公司的ACCESS系统

图中S 表示子结构；δ 表示化学位移

(response surface) 的研究。图1.3所示为因变量Y对于自变量x_1和x_2的响应曲面及其数学模型。这类响应曲面可为某种流程的调优提供理论依据。

再如联邦德国BASF公司的综合谱图库信息系统中，ACCESS (Automatic Computer Correlation and Evaluation of Spectra based an Substructure)程序包可以自动解析化合物结构，其解析过程示于图1.4。

1.2.2 数据库的作用

数据库不仅是提供原始数据的，也是信息再加工的必须手段，而且能为进一步研究提供条件，开拓通路。因此数据库是基础研究工作之一。

1.2.2.1 化学数据库与人工智能

人工智能的研究（如：有机化合物解析）与数据库的关系尤为密切。

DENDRAL系统是大家所熟知的最早(60年代中叶开始)的专家系统之一（见第六章),此系统是美国斯坦福大学的质谱学家、遗传工程学家、化学家和计算机专家通力协作的一项卓越成果。在该系统中，运用低分辨质谱中的断裂规则及碳-13和氢核磁共振波谱等进行化合物结构的演绎推理。随后就在许多企业、研究机构和大学获得了应用。这一先驱工作的成功振奋了整个人工智能领域。

在DENDRAL系统演绎所得候选化合物的验证中采用了谱数据库[3]。如图1.5所示，程序功能分为两个部分。第一部分,根据子结构环境和H谱化学位移的相关性,建造H谱数据库:（1）程序HCODE，完成子结构编码的自动生成和数据库的构造;（2）程序HDCODE，完成数据库中子结构-化学位移归属。第二部分，未知化合物H谱的分析:（1）程序HNMRP,

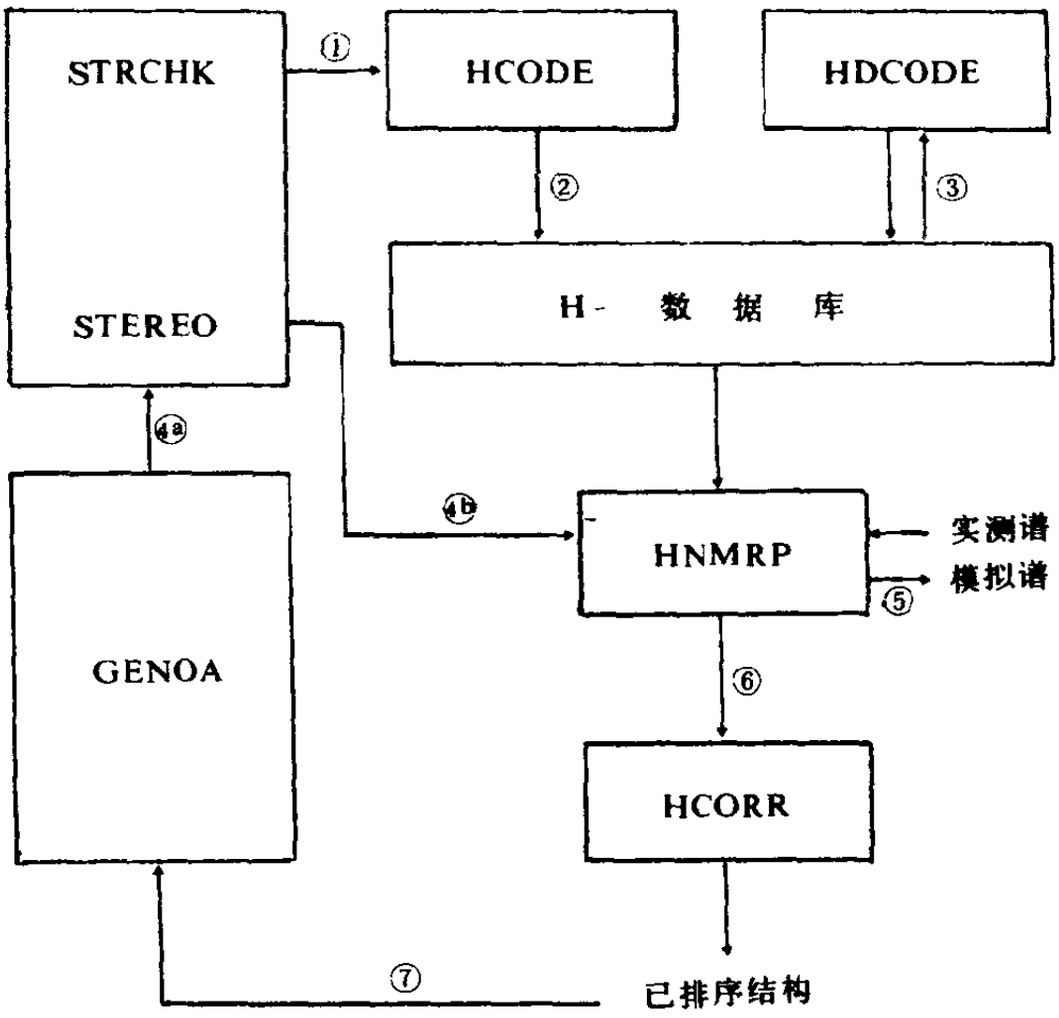


图 1.5 H谱分析程序

① 未知结构说明；② 与数据库联接；③ H谱数据库的维护及检验；4a 同分异构体生成；4b 空间异构体生成；⑤ 波谱预测；⑥ 预测谱与实验谱的比较，并将候选化合物排序；⑦ 将此结果用于约束化合物的最终确证

预测候选化合物的H谱；（2）程序HCORR，比较预测谱与实验谱，并根据二者的符合程度进行候选化合物的排序。

再如，新近已由南斯拉夫化学计量学教授Zupan研制成了CARBON这一结构解析专家系统[4]。此系统的显著特点是它具有浓厚的数据库色彩，Zupan等在70年代中已经建立起他们

的C-13、IR等谱图库信息系统，同时创造了三距离聚类分析方法。CARBON系统主要运用的是C-13谱图数据，目前，此数据库中共有2536张C-13 NMR谱图，30000个已作归属的化学位移。以此数据库为基础，提取不同取代基的化学位移信息，用于谱图的模拟和归属。

专家系统CARBON的知识规则由四部分组成：(i) 相似谱图的系统聚类树（三距离法）；(ii) 不同取代基、不同空间位置及不同构象的化学位移表；(iii)功能团类别的统计和数学表达；(iv) 子结构检索、归属等的图论算法。由此可见，CARBON系统是在坚实的数据库基础和技术的根基之上研制而成的。

1.2.2.2 化学数据库与分子设计

研究活性物质，如药物、杀虫剂及农药等的合成方法，多年来一直是沿用纯经验的方法，这要先合成一系列化合物，然后进行生物实验进行筛选。对于有效物质的改进，通常采用结构的修饰方法，如以穷举法进行各种功能团的取代。这种用经验、以母体衍生物为线索的药物合成方式效率极低。以农药为例，在国外市场上，一种化学药品的需要合成量随着年代变化示于表1.2。若以表1.2增长速估测，到1997年，要创制一种新的农药，须合成的化合物将变成天文数字。也就是说，时至今日，

表 1.2 创制一种农药的合成量

年　代	1956	1964	1970	1972	1979
合成量（篇）	1800	3600	8000	10000	12000

单靠纯经验探索药物的方法已经陷入困境。因此，人类在寻求新的方法来摆脱这种状况。其中的一种途径是运用计算机来辅助