第七卷 质 谱 学 报 Vol.7
第四期 JOURNAL OF CHINESE MASS SPECTROMETRY SOCIETY No.4

会话式质谱库检索系统与应用

朱伟雄 董蓉卿

（军事医学科学院仪器中心室）

〔摘要〕 本文介绍一种适合于在微型或小型计算机上工作的会话式质谱库检索系统，供质谱工作者参考使用。

随着微电子技术的进步和电子计算机技术的发展，电子计算机的应用已深入到色谱、质谱、红外光谱、核磁及顺磁共振谱仪等领域。尤其是色谱——质谱——电子计算机联用，使电子计算机在有机质谱中的应用显得日益重要。质谱数据的采集，处理和解析均由电子计算机自动实现，它为识别未知化合物属性显示了新的活力。

质谱谱图的压缩汇编

原始的质谱谱图一般包含大量的质量和强度数据，而这些信息并不是在以后的解析算法中都需要利用的。为了减少谱库存储空间和提高检索速度，原始质谱数据需要按事先规定的某种编码方式规格化和压缩成为简谱。为此人们设计了各种压缩编码方式。其方法的取舍主要取决于所用的解析算法模型。如对每一质谱图在质量范围内选取5，6，8或10个最强峰。或将全谱以7，14或20原子质量单位为间隔划分为多个小质量区间，然后在每一区间内选取1—2个最强峰。也有的是采用二元编码法，对每一质量数按是否存在大于规定阈强度的峰进行编码。有峰为1，无峰为0。

总之方法甚多并在质谱库检索中都获得了一定程度的成功。但同时也存在某些不足。如质量划分从哪个质量开始以及每一质量区间的大小和从中选峰的数量都较大地影响检索效果。至于在全谱范围内取6—10个最强峰的方法将由于丢失较多的对解析有重要意义的峰而逐渐不被采纳。也曾有文献报导，低于0.3%的质量峰对谱的唯一性贡献很小，但若仅取高于1%的质量峰又将丢失大量的有用信息。因此，为避免上述弊病和尽量多地保存包含结构信息的峰数据，舍弃无明显价值的一般峰数据，我们成功地研究了一种新的质谱数据压缩法——按质谱谱图的自然峰簇选取特征峰。

新的质谱数据压缩法的先决条件是必须首先判定谱中各峰簇的界限。在质谱谱图中所谓峰簇是指一组连续的质量峰。我们定义凡峰距≤2AMU的二个连续质量峰属于同一峰簇，而峰距≥3AMU的则为不同峰簇。由于峰簇内各质量峰峰强的不规则性，因此为正确地确定峰簇界限，必须经过二次五点光滑处理使峰簇呈类高斯分布状。然后由程序自动跟踪寻找峰谷以确定全谱中每一峰簇的界限，并根据确定的峰簇界限选取一个最强峰作为该峰簇的特征峰汇编入谱库。此法的优点是使相对丰度比较小的分子离子峰也能被汇编收入谱库，并且压缩谱基本保留原始谱图的形状和尽可能多的有用峰数据，从而改善了检索效果。而对于 Bie-

1985年11月16日收

mann检索简化法（在每14AMU区间内选取二个最强峰），则往往不能保证分子离子峰被选入或有可能当二个峰簇同属一个质量区间及由于一个峰簇的峰强明显地高于另一个而仅选取一个峰簇的峰数据。谱图的压缩示意见图3。谱图数据的压缩率根据200种食品中挥发性化合物的统计，平均小于1/3（在1/2与1/6间分布）。

质谱谱图库的结构

会话式质谱库由质谱库数据文件（简称谱数据文件）和质谱库索引文件（简称索引文件）两部分组成。每一文件均包含若干定长记录。数据文件的记录长度为512字节，索引文件的记录长度为20字节。每一文件的记录1均作为储存关联信息用，以便在每次启动程序时保持对记录地址的跟踪。

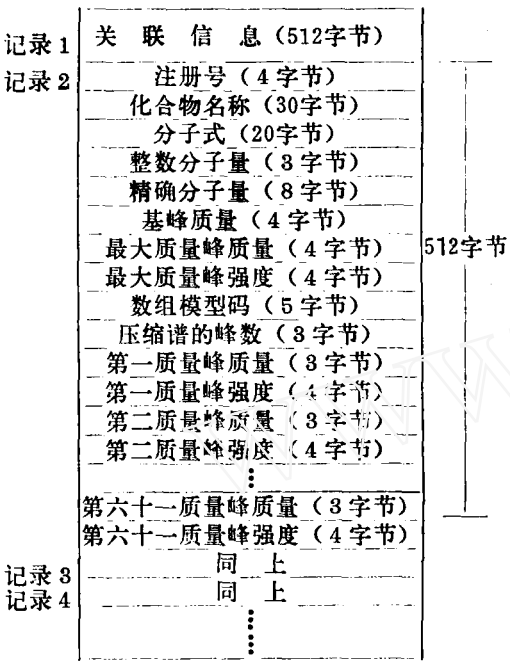

图1 质谱库数据文件格式

系统的库生成程序包含插入，修改，删除，列表等功能。执行插入操作时在谱数据文件和索引文件中产生新的记录。每一谱数据文件记录储存一个压缩的谱图资料。资料由两部分组成：前部为化合物的描述信息；后部为经压缩的化合物谱数据（可储存61个峰数据）。谱数据文件大小可根据欲储存的谱数和可利用的磁盘空间设定，不被结构限制。谱数据文件的格式见图1。

索引文件是专为加快检索速度而设计的，我们采用分子量作为关键字。在作质谱库插入操作时新输入的质谱数据将按其分子量由程序产生索引记录插入索引文件中。程序使索引记录按化合物分子量从小到大排列。索引文件包括主索引区和索引扩展区两部分。主索引区内含699个记录。除记录1作为关联信息供系统使用外，其余记录可容纳698种不同分子量的质谱索引。记录700后为索引扩展区，供不同分子量化合物保存数量众多的索引。索引文件的格式见图2。由图可见每一记录包含5项，第五项为索引扩展指针（也可称勾链地址）。如无索引扩展，其内容为0。如有索引扩展，则为扩展的索引记录号。根据主索引区的设定和结构计算，可容纳2094个谱索引。然而索引扩展区则可根据实际需要任意延伸不受限制。

为保证质谱数据的完整性和防止频繁的记录拷贝和转移，在做谱删除操作时本系统不采用实际删去然后重新排序的方法，而是在索引文件和谱数据文件的被删除记录位置加注特殊的标记，使库中留有“空洞”。待下次执行库谱插入操作时，系统能自动根据加注的特殊标记优先填补“空洞”，从而使质谱数据库始终保持最紧凑的空间结构和获得较高的处理速度。关于质谱库生成程序的流程见图4。

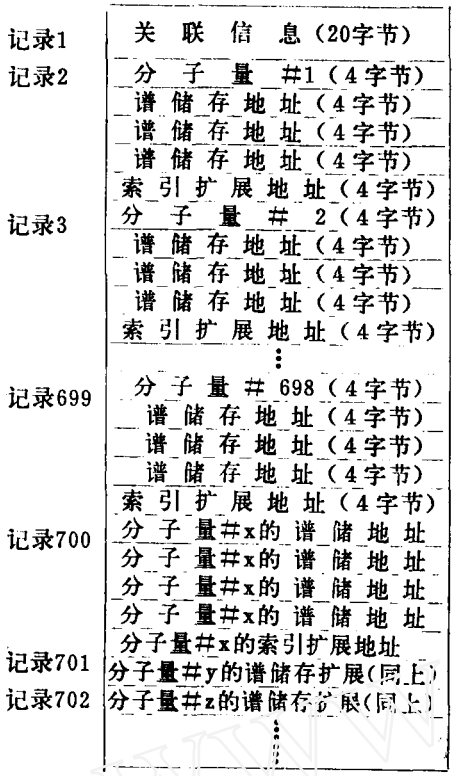

图2　索引文件格式

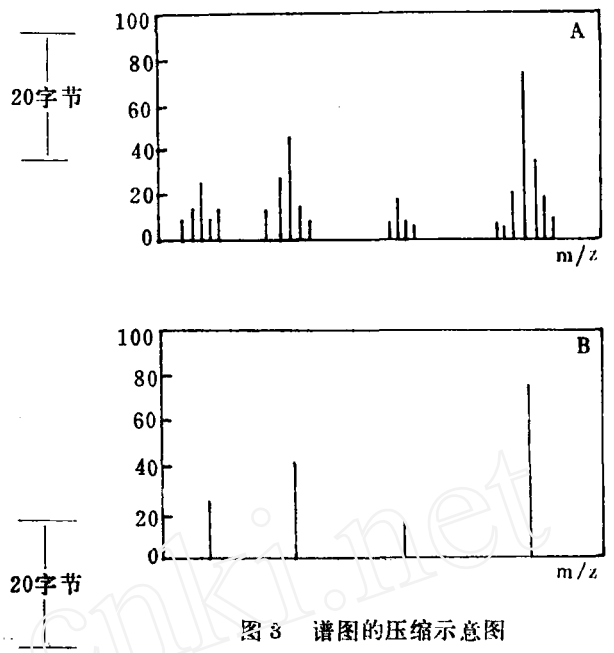

图3　谱图的压缩示意图
A：原始谱　B：简化谱

质谱谱图库的检索

设计一个谱比较检索系统时，一般均要考虑以下四个因素：（1）尽可能地缩短检索时间，能为使用者所接受；（2）不能因仪器类型或操作条件不同而较大地影响检索结果；（3）当未知化合物谱图包含在谱图库中时，应在检索结果列表中名列前茅；（4）当未知化合物谱图不包含在谱图库中时，也能在检索结果列表中提供几个同系物或相似的化合物。本系统根据上述要求设计了检索算法。系统复合应用了直接检索和二级检索。直接检索用于按化合物名称、分子式、分子量、分子量复合基峰四种方式进行的谱检索。二级检索用于单纯按谱、分子量复合谱、可变预检索条件三种谱检索方式。

不管是直接检索或二级检索，在本系统中均通过谱索引文件寻找谱记录所在位置。索引文件的结构由上述已知为一系列由分子量大小排序的记录组成。记录内包含分子量和相关的谱图储存地址指示器，指示器指示谱图在质谱数据文件中的位置。在按化合物名称和分子式作直接检索时是从索引的始端直到索引末端逐个比较有关项而实现检索的。由于化合物名称或分子式是文字数字字符串，因此以上两种检索方式的检索过程实际上都是执行字符串的逻辑比较。然而，对于分子量和分子量复合基峰检索，为了加快检索速度程序采用了对半搜索法。

对于按谱检索，为了缩短检索时间程序采用二级检索。整个检索过程分为预检索和精细比较两步。预检索采用树状结构判别。它的目的是通过选定的预选条件剔除与欲检索的未知化合物差别较大的库谱，筛选出与未知化合物有一定共同特征的一组库谱供精细比较。我们设计的预选条件是：（1）判定两个谱图的质量复盖范围是否在三倍以内（在按分子量复合谱检索时，本条件不适用），（2）进行基峰强度比较，判定二谱基峰与其对应峰是否每一

峰的强度至少是另一峰的25%（在按可变预检索条件检索时，本条件可按需修改）；（3）进行同系离子峰矩阵模型比较，判定二谱间同系离子峰矩阵码是否满足66%以上的相似要求（在按可变预检索条件检索时，本条件可按需修改）。同系离子峰矩阵码是根据形成的MOD 14离子系列谱，然后从中挑选6个最强峰取为1，其余设定为0而组成的。它是一个5位8进制编码，在一定程度上能反映同系物或相似化合物的谱图特征。

对满足上述三个预检索条件的库谱才做精细比较，它计算各预选谱与未知谱的相似因子。本系统中使用Biemann相似因子表示相似性。

$$\text{相似因子} = \frac{\text{匹配峰的平均加权比}}{\text{不匹配峰的分数} + 1}$$

公式保证了二谱完全符合时相似因子为1，不符合时相似因子接近0。

最后，程序依照计算所得的相似因子挑选其中5个具有最大相似因子值的库谱打印输出。关于质谱库检索程序的流程见图5。

质谱库检索系统的软件特点与应用

考虑到软件的可移植性，本系统软件使用PDP-11 FORTRAN IV-PLUS语言中的标准语句书写。

为加快系统处理速度和适应微型或小型计算机存储容量小的特点，本系统软件不采用程序复盖方式，而是将整个软件分成质谱库生成和检索两个独立的应用程序部分。每一程序约占36K字节内存空间。为尽可能地减小每一独立应用程序所占的内存空间，对于系统中经常使用的功能模块均以子例程子程序方式编写和调用。

程序中建立和使用的谱数据文件和索引文件以直接存取记录方式驻盘保存，供系统软件随时调用而不常驻内存。又索引文件记录采用开环勾链方式，因而储存谱图的数量仅取决于磁盘的容量而不受文件格式和内存的限制。

整个软件操作由人机对话方式实现，易于为不懂计算机语言的质谱工作者掌握。

本系统软件已应用于特殊毒物的鉴定，证明效果较好。对于谱图已包含在库中的欲检索毒剂均获得最高的相似因子值，而与欲检索毒剂有相似结构的其它毒剂排列在后未发现错判的现象。此外我们也曾利用本系统对随机挑选的200个食品中的挥发性化合物进行检索试验，同样获得令人十分满意的效果。因而本质谱检索系统具有实际应用价值。对于其它类化合物的检索效果尚待今后验证。

讨 论

本文介绍的会话式质谱库检索系统对库结构和谱数据压缩汇编法做了一些探索，较适合于推广移植到其它类型的以磁盘为基础的微型或小型计算机上工作。我们认为它可以作为已开发的质谱库检索方法之一推荐给质谱工作者参考，并在今后的实际使用中改进和完善。

鉴于建立专用质谱库的目的和对象不同，我们在编制程序时仅考虑分子量小于或等于700AMU和谱质量峰个数小于300的化合物。因而若需建立包含较大分子量和质量峰数据较多的专用质谱库时，需增大程序中的有关数组的维数和扩大索引文件中的主索引区。

本系统采用的是按质谱谱图的自然峰簇选取特征峰，对于无明显峰簇特征的化合物类不适用。因为当质谱图无明显峰簇特征时本压缩汇编法已失去简化谱图和节省存储空间的特点。

本工作承蒙汤炳生，贾敬山，武力民，周杰，杨松成等同志大力协助和卢涌泉副教授的热情指导，特此致谢。

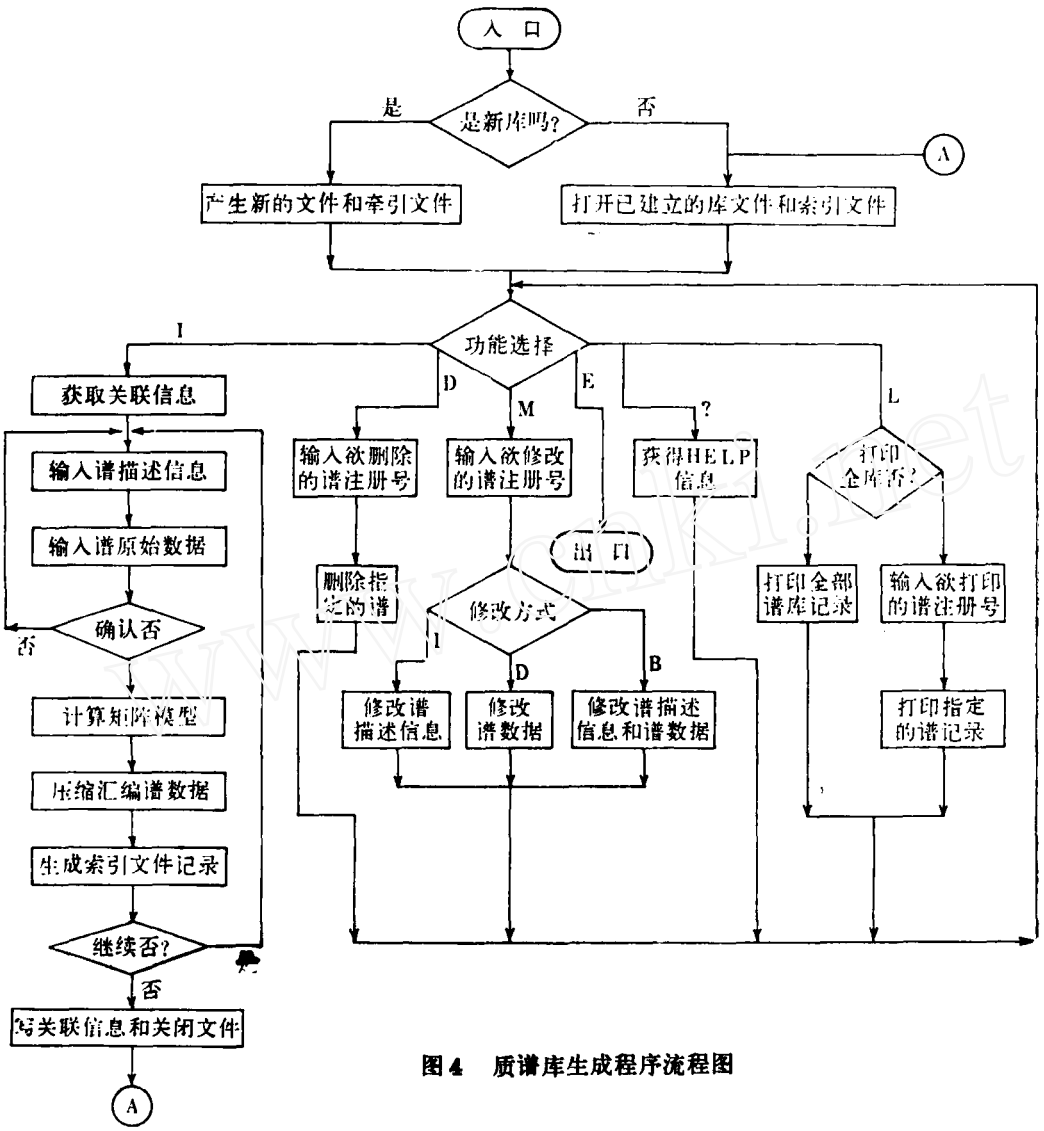

图4　质谱库生成程序流程图

参　考　文　献

〔1〕 Abrahamsson S: Science Tools 14:29, 1967

〔2〕 Crawford LR, et al: Anal Chem 40:1465, 1968

〔3〕 Knock B, et al: Proceedings of The 17th Annual Conference on Mass Spectrometry and Allied Topics, ASTM E14 committee Dallas, 1969

〔4〕 Grotch SL: Anal Chem 42:1214, 1970

〔5〕 Hertz HS, et al: Org Mass Spetrom 4:452, 1970

〔6〕 Hertz HS, et al: Anal Chem 43:681, 1971

〔7〕 Biller JE, et al: Nineteenth Annual Conference on Mass Spectrometry and Allied Topics, Atlanta, Georgia, June, p85, 1971

〔8〕 Mcfadden WH: Techniques of Combined Cas Chromatography-Mass Spectrometry Applications in Organic Analysis, John Wiley, p284, 1973

〔9〕 胡鑫尧等：《计算机在分析化学中的应用》，第182页，1983

〔10〕 Spetrosystem MAT200 Operating Manual D1020101

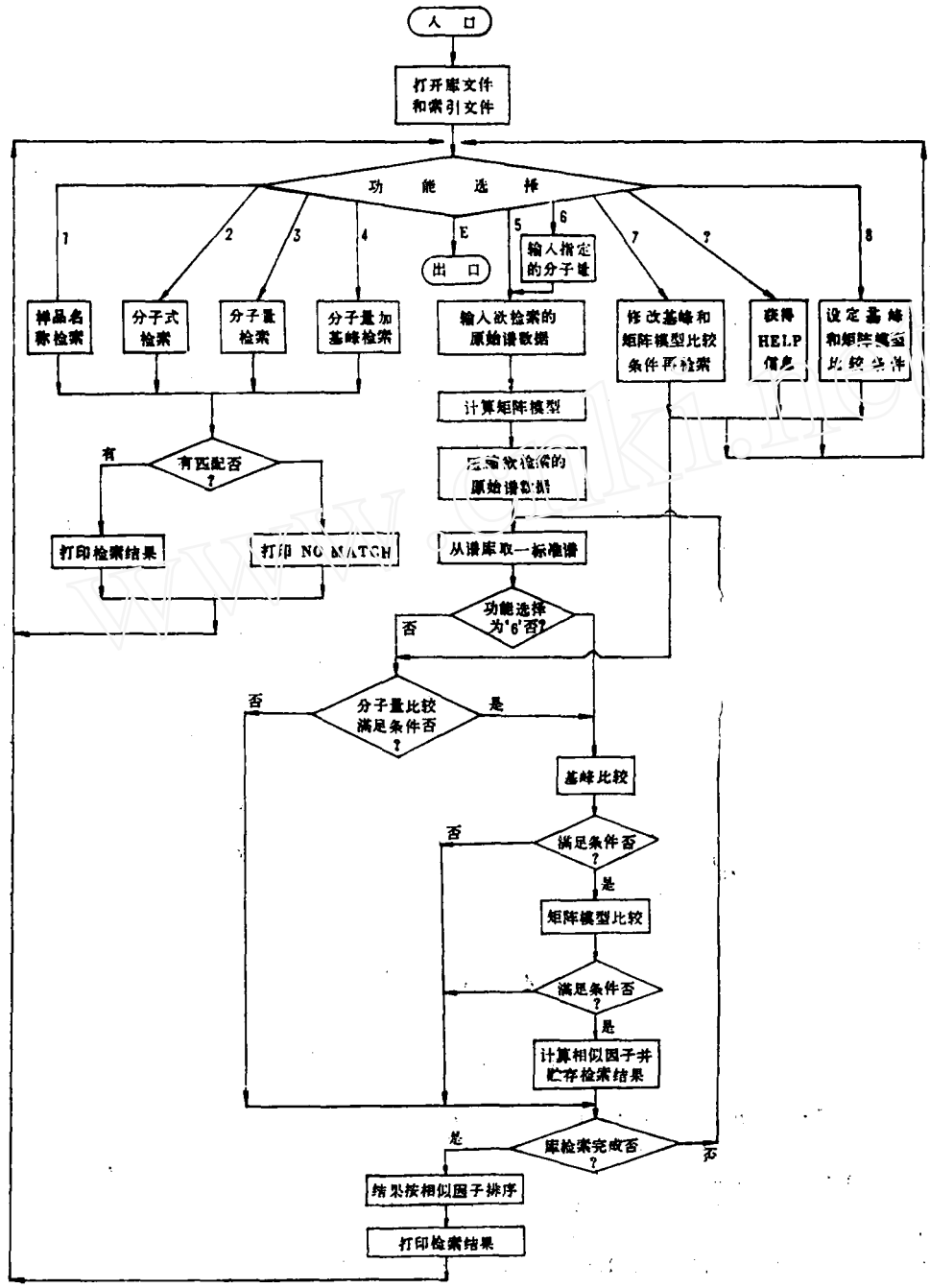

图5 质谱库检索程序流程图

Conversational Mass Spectral Search System and Its Application

Zhu Weixiong　Dong Rongqing

(Instrument Centre, Academy of Military Medical Sciences)

Abstract

This paper describes a conversational mass spectral search system which is suitable for microcomputer or minicomputer and runs on PDP-11/34A computer. It permits user to set up a special mass spectral library for various organic compounds. The software consists of a library generation program and a library search program. These programs were written by PDP-11 FORTRAN IV-PLUS language, and is easy to transfer to other computers. The library generation program is used to produce a data file for simplified mass spectra and a index file for searching. In order to keep integrity of a library, the program was designed to have some additional functions, such as inserting, modifying, deleting, examining etc. The library search program used immediate and dual searching techniques and had seven pattens for searching. The search system was applied to identify some special toxic and volatile compounds in food, the results were satisfactory.

简　讯

QP-1000质谱计用户协作组成立

中国科学院科学仪器厂和日本岛津公司于1986年10月24日在北京举办了QP-1000A GC/MS技术讲座，50余位来自全国的质谱界人士参加。科仪厂介绍了四极质谱的基本原理、组装岛津QP-1000A GC/MS的特征，并用实际仪器示范了各种操作程序。岛津公司北京分析中心的村田武博士报告了高质量数高沸点样品在QP-1000A上作GC/MS分析。

国内12家已有QP-1000的用户在会上成立了用户协作组，推选北京市劳动保护研究所赵子璋同志为负责人，并立即和岛津公司进行了有关仪器维修方式的座谈，一致同意每年召开一次用户协作会议，以促进该仪器的使用效率。

（野　渡）