微计算机质谱库检索系统

梁曦云、张美怡、陈稚芳、刘津琨
（中国科学院化学研究所）

〔**摘要**〕 本文介绍在微计算机上建立的一个质谱库检索系统。这一系统允许建立专一类别化合物的微型谱图库，并可以方便地增加和修改谱图数据及其辅助信息，能够产生和检查检索关键字，并最后从谱图库中检索与未知谱图同类型的化合物谱图。系统方法硬件结构简单，使用CBM/PET2001微计算机联接PET—2040，5.5英寸磁盘系统，包括辅助信息和谱图数据的谱图库存在一个软磁盘上，每一个谱图建立一个文件。系统的管理、维护和应用程序以及各种索引文件存在另一个软磁盘上。系统采用人机对话方式，程序采取功能模块结构，检索方法采取两级检索技术。

引　　言

微电子学的飞速发展使计算机技术大大改观，以微处理器为中心的台式微计算机的功能和容量已达到六十年代大型计算机的水平，而其体积与价格只及后者的几百到几千分之一，因此微计算机作为实验室的研究辅助工具已被确认是必然的趋势。最近两年来微计算机在实验室的应用也普遍受到国内科技界的重视。微计算机在实验室的使用主要可分为两类，一是仪器的控制和数据采集，二是数据和资料的存贮和回收。由于质谱数据的特殊性使它特别适合用计算机处理，因此质谱数据系统发展较早[1-4]，其结果也普遍地推广到商品仪器上。质谱图数据量大，信息复杂，谱图检索一般都由较大的计算机来实现，但以数量有限的单一类型（或有相同性质）化合物谱图来组织小型的谱图库和建立检索系统完全可以在微型计算机上进行。这种小型的质谱数据检索系统虽然概括性不如大型系统，但在经济上和灵活性上的优点都是大型机系统不能达到的。一般中、小型实验室都有与自己课题或任务相关的某种类型的化合物需要识别与鉴定，微型机系统恰好能满足这一需求，随着微型计算机的容量和功能不断地扩大，这种微型机质谱系统将能容纳越来越多的谱图以及辅助信息，使它能担当更大的数据存贮任务，具有更多更强的检索和鉴定功能。本文介绍了为探讨微型机在质谱数据存贮和检索方面的应用所建立起来的一个试验性系统。

1983年3月21日收。

技 术 部 分

设备和配置：

系统建立在CBM公司较旧的PET2001台式计算机上(相当于当前CBM 4000 系列机)。在原有的8KB内存上扩展了4KB的存贮器。外部设备包括一台COMMODORE2040 双软盘驱动器（使用单面双密度5.5英寸软盘，总数据存贮量为340KB）和一台HP—9876A 热敏绘图打印机作硬拷贝输出。这样的配置是按实验室最低限度来设计的。系统可以通过IEEE—488标准仪器总线扩展（最多到15台外设或仪器）。

由于微计算机内存少，在系统设计时，软件按功能分成模块，用链接复盖的方法互相调用。系统的管理、维护、应用程序以及各种索引文件存放在零号磁盘上，质谱数据及其辅助信息存放在一号磁盘上，程序和数据的分别存放使得一个管理盘可以管理存放在不同数据盘上的谱图数据。所有的软件都用BASIC 语言编写，每个质谱的数据和辅助信息全部以独立数据文件方式存贮。因而这个系统可以很容易地移植到任何其他微型计算机上，具有较强的通用性。

系统结构：

利用这个系统建立的亚硝胺类化合物的专用质谱库共包括150个不同结构类型的亚硝胺类化合物的电子轰击源低分辨质谱图，谱图信息主要分为两部分：第一是化合物的描述信息，包括化合物的名称、分子式、分子量及峰数。这些信息都以不定长记录格式存放在谱图文件标题部分。第二是质谱图数据信息、质谱峰的数据以质量数-峰强度对的格式按质量数从大到小在文件内存放。质谱数据文件的格式见图1。如前所述，为了保持系统的通用性，以上信息是以顺序文件格式存放于数据软盘上。这样通过微计算机的文件管理系统存取质谱数据比较简单，但是速度受到一定的限制。为了减少磁盘访问次数，缩短检索时间，采取两级检索的方法。

第一级检索或予检索采取关键字直接比较的方法筛选出与样品谱有一定共同特征的一组中选谱图，然后进行第二级精密比较，选出与样品结构最接近的标准谱，达到识别或鉴定的目的。予检索使用的谱图关键字是谱图信息高度压缩后所得到的数字组合。由于质谱的复杂性，一张谱图所包含的数据量很大，以现有的亚硝胺谱图库为例：质量范围在250以内，平均峰数都在50个以上（以0.1%基峰为阀值）。以这个数字为基准，一张谱图的质量和强度信息至少为200字节，100张谱图至少占20K，而我们的微计算机的内存有限（程序加数据共12K），所以采取适当的数据压缩是必要的。

我们采取的压缩方法是以“MOD 14”谱图方法为基础[5]。这个方法是取一个质谱内的质峰强度按14质量的间隔累加起来（见图2a），形成一个只有14个强度值的组合谱图（图2b)。其方程如下：

$$S_i = \sum_{i} I_{(7+i+j\times 14)} \qquad (1)$$

S_i = 第 i 组峰总强度

i = 从 1 到14峰组序号

$I_{(x)}$ = X质量峰强度

j = 要概括谱图质量范围的倍乘数

例如：第一组的总强度为质峰 8 、22、36、……的强度加和

第二组的总强度为质峰 9 、23、37、……的强度加和

由于大部分的有机化合物都包含以-CH_2-为单元的碳链，因此相隔14质量数的质峰系列是常见的，而不同的取代基团可使这种系列有所差异，因此以这种方法得到的压缩谱图（MOD 14质谱）很好的保持了原来谱图的特征[5]，特别是由不同取代基所导致的差异。

为了适应有限的内存和实现快速筛选，我们对MOD 14谱图的信息进一步加以压缩。取最强的 5 个峰组，按组号的大小顺序排列，组合成一个含 5 个整数（10个字节）的关键字。用以上方法对所有的 150 个亚硝胺谱图进行编码，并按谱图目录的顺序建立索引文件。如图 3 （Ⅰ）所示，在进行予检索的过程中，关键字索引以数组形式保留在内存，将样品谱关键字与索引中的标准谱关键字一一对比，关键字完全吻合者为选中。选出的谱图序号保留在中选表的文件中，以备下一步精密检索之用。由于予选过程全部在内存进行，所以全部过程只需几秒钟。

第二级精密检索是在样品谱与予检索中选出的标准谱间进行。如图 3 （Ⅱ）所示，通过予检索中选表提供的谱图序号，从目录文件中找到标准谱文件名，然后通过文件管理系统从软盘中逐一取出谱图数据与样品谱作精密比较，并按一定数学模式计算出匹配因子。计算匹配因子的方法很多，本系统采取全谱距离分析法：质谱图被看作一个由N个线性互不相关的矢量（质峰）组成的一个矢量。

$$\vec{MS} = (c_1 m_1,\ c_2 m_2,\ c_3 m_3,\ \cdots,\ c_i m_i,\ \cdots c_n m_n)$$

$\{m_i\}$ = 质量峰（单元矢量）的基集

c_i 为对应每个质谱的强度系数〔一般情况为质峰的相对丰度（%）〕

两个质谱图的差别可以通过计算在质量峰的矢量空间中两个矢量的距离来表示。计算多维矢量距离的常用公式有以下两种[5]：

$$\alpha_{RMS} = \sqrt{\sum_i (c_i^2 - c_i^{*2})^2} \qquad (2)$$

$$\alpha_{abs} = \sum_i | c_i - c_i^* | \qquad (3)$$

c_i—样品谱强度系数，c_i^*—标准谱强度系数。

系统采取方程（3）作为计算矢量距离的方法。显然两个矢量的距离越短，则谱图匹配越好，完全吻合时距离为零。由于质谱数据的特点，强度信息一般只是以基峰为准的相对值，为便于比较，在计算匹配因子时需引入样品谱的总离子强度作为归一化因子。在系统中使用的匹配因子计算方程如下：

$$DF = \sum_I |SB(I) - RB(I)| / \sum_I SB(I)$$

DF = 匹配因子（差异因子）

SB(Ⅰ) = 样品谱第Ⅰ个质量峰的相对强度

RB(Ⅰ) = 参考谱第Ⅰ个质量峰的相对强度

Ⅰ = 质量数

样品谱逐个与中选的标准谱比较，计算出匹配因子并按从小到大的顺序排列，即得到基本反映结构接近度的最终检索结果。在谱图库内没有与样品相同的化合物时，也可通过比较

匹配因子的数值和相应质谱的分子结构获得有助于了解未知物结构的信息和提示。

软件结构与操作

这套实验性的质谱数据管理和检索系统共包括八个应用程序，全部采用BASIC语言，存贮在一个软磁盘上。其中主要为监控程序、检索程序、编辑程序、谱图编码程序。

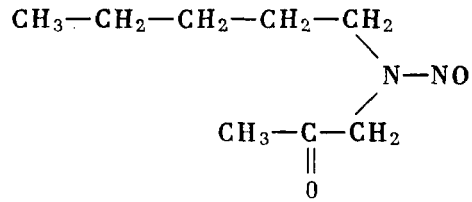
全部系统操作由监控程序“MS—MONITOR”监督。图4为监控程序框图。为便于非计算机人员使用，此程序通过人机对话方式控制执行系统的三个主要功能，执行时，利用PET操作系统的链接功能，调用功能模块，复盖在原有的程序上，以克服微计算机内存量小的缺点，模块间的数据交换都通过软盘数据文件来实现。使用这种方法还可以通过改变功能模块方便地改变预检索和精密检索所使用的数学模型。

图5为系统功能之一——谱图数据更新程序框图。系统在“MS—MONITOR”控制下执行，用户可通过人机对话方式建立小型专用质谱数据包，（表1，2）并可随时查阅，增加和修改样品质谱图数据及其辅助信息（表3）。

图6为系统功能之二——谱图检索程序框图。用户可以从键盘输入被检索样品的谱图信息（表4），系统自动调进编辑程序进行样品图数据编码，取关键字，然后调入MS—SEARCH程序，先进行一级检索，选出预选图，再将预选图进一步精密检索，计算匹配因子，且打印或萤光屏显示出检索结果。（表5）

图7为系统功能之三——标准谱图编码程序框图。执行时，可将谱图按MOD14方程简化，组成关键字，并由用户决定是否纳入系统的关键字索引中（表6）。在此之前用户可以在屏幕上检查生成简谱以及关键字数据。

检索效果如表7所示。库中含有的与样品相同的化合物谱图都在首位检出，而与样品同类结构的化合物谱图都出现在输出表的前端。表5是一个库内没有的样品谱的检索结果。样品为N—NITRSO—N—(1—METHYL—2—OXOPROPYL)—PENTYLAMINE[6]，其结构式为：



虽然库内没有相同结构的化合物谱图，但排在中选表前端的几个标样都含有与结构（I）侧链碳数相同或相近的烷基，或含有羰基侧链的子结构。尽管从匹配因子数值看出中选谱与检索目标谱差异较大，但通过系统的检索仍能得到样品的部分结构信息。

结　　束　　语

本文介绍的在商品台式计算机上自行编制的质谱数据管理系统，其特点是简单易用，全部以对话方式操作，非计算机专业人员可以很快掌握使用方法，建立与自己有关的质谱图库，并使用编辑功能进行修改和进行未知物检索。由于所采用的是最低限度的配置，内存和外存有限，因此存贮谱图量不多（目前只能存150个）。这一限制可以通过扩展内存和增加磁盘驱动器去解决。但在另一方面，由于采取的硬件结构简单，方法有较大的适应性，可以很容易的移植到其他类型的微型机上，具有较广泛的推广意义。使用解释性的BASIC语言编写检索程序，也使检索速度受到限制，虽然通过予筛选已将需要进行比较的范围大大的缩小，但最后精密比较时，每张谱图仍需20—30秒。（如中选表中有15个谱图，则整个检索过程要7

表 1

```
DO YOU WANT TO MODIFY THE LIBRARY ('Y' OR 'N')? Y
OPTIONS AVAILABLE?
 1 = APPEND NEW SPECTRUM
 2 = MODIFY OLD SPECTRUM
 3 = RESTART
OPTION REQUESTED? ( 1 OOR 2 OR 3 )
```

表 2

```
        (OPTION 1): FILENAME?
                    DISK NO.?
                    CHEMICAL NAME?
                    MOL. FORMULAR?
                    MOL. WETGHT?
        ENTER 'MASS NO.', 'REL. ABUNDANCE'
        IN DECENDING ORDER; ENTER -1, -1 TO END
        PEAK NO.1  MASS NO., RELATIVE ABUNDANCE
```

上为从键盘输入内容

表 3

```
        (OPTION  2):  MODIFY SPECTRUM
                      SPECTRUM TO BE EDITTED?
                      PRINT REPORT?
        (ANSWER: Y)   REPORT OPTIONS?
                       1 = SCREEN OUTPUT
                       2 = PRINTER OUTPUT
                       3 = EXIT
                      OUTPUT REQUEST?
        (ANSWER: N)   EDITOR OPTIONS?
                       1 = CHANGE PARAMETER
                       2 = CHANGE SPECTRAL DATA
                       3 = DELETE MASS PEAK
                       4 = EXIT
                      OPTION REQUTSTED?
```

表 4

```
SEARCH SAMPLE? Y
SAMPLE I. D.? XXXX
EXPECTED MOL. WT.? 9999
ENTER 'MASS NO.','REL. ABUNDANCE'
IN DECENDINE ORDER; ENTEB -1, -1 TO END
PEAK NO. 1
?  MASS 1, REL. ABUNDANCE 1
?  MASS 2, REL. ABUNDANCE 2
```

表 5

```
LIST OF MATCHES FOR N-NITROSODIMETHYL UREA
SPECTRUM NO.: 42   C4. H9.N3.02.                0
N-NITROSODIMETHYL UREA
SPECTRUM NO.: 45   C5. H11.N3.02.               0.2279
N-NITROSO-1; 1-DIMETHYL-3-ETHYLUREA
SPECTRUM NO.: 44   C6. H13.N3. 02.              0.8884
N-NITROSO-1-METHYL-3, 3-DIETHYL UREA
SPECTRUM NO.: 108  C6. H10.N2.03.               1.6791
N-NITROSO-BIS(2-OXOPROPYLL)AMINE
PRESS 'SPACE BAR' TO CONT
```

表 6

```
ENCODE REFERNCE SPECTRUM
**HOMOLOG SUM EVALUTION**
MASS SPECTRUM TO BE TREATED? (MS XXXX)
OPTION TABLE:
 1 = DISPLAY HISTOGRAM
 2 = PRINT
 3 = STORE SEARCH PARAMETERS
 4 = EXIT
YOUR CHOICE?
```

表 7

LIST OF MATCHES FOR:N-NITROSO-N-(1-METHYL-2-OXO-PROPYL)-PENTYLAMINE			
SPECTRUM NO.:	2	C10.H22.N2.0	.9781
DIPENTYL NITROSAMINE			
SPECTRUM NO.:	45	C5.H11.N3.02	1.1913
N-NITROSO-1; 1-DIMETHYL-3-ETHYLUREA			
SPECTRUM NO.:	42	C4.H9.N3.02	1.2158
N-NITROSOTRIMETHYL UREA			
SPECTRUM NO.:	90	C12.H26.N2.0	1.2186
DI-N-HEXYL NITROSAMINE			
SPECTRUM NO.:	108	C6.H10.N2.03	1.2951
N-NITROSO-BIS(2-OXOPROPYL)AMINE			
SPECTRUM NO.:	95	C6.H12.N4.02	1.3005
N;N′-DINITROSO-2; 6-DIMETHYLPIPERAZINE			
SPECTRUM NO.:	102	C19.H40.N2.0	1.3716
N-NITROSOMETHYLOCTADECYLAMINE			
SPECTRUM NO.:	19	C14.H30.N2.0	1.3726
DIHEPTYL NITROSAMINE			
SPECTRUM NO.:	44	C6.H13.N3.02	1.4426
N-NITROSO-1-METHYL-3; 3-DIETHYL UREA			
SPECTRUM NO.:	142	C6.H18.N2.0	1.4563
PROPYLPENTYLNITROSAMINE			

分钟）。这比人工查找固然快得多，但从计算机的角度来说是非常缓慢的。不过检索速度在微型机系统不是严重问题，因为整个计算机资源都由用户支配，时间安排可以灵活处理，也不存在机时费问题，这是与用大型计算机有所不同的。使用大多数人都熟悉的 BASIC 语言编写程序，使用户能对检索程序进行改造，以增加其功能，或适应其特殊要求。而解释性语言在调试检错过程中的优越性是众所周知的。我们利用此系统建立了亚硝胺类化合物的专用质谱库，并通过更换程序模块的方法对不同的检索方案进行研究试验。这种实验在商品数据系统中是无法进行的。上述的质谱检索系统除了作为一个能推广的实用系统的模型外，同时也是一个开发研究质谱检索方法的工具。

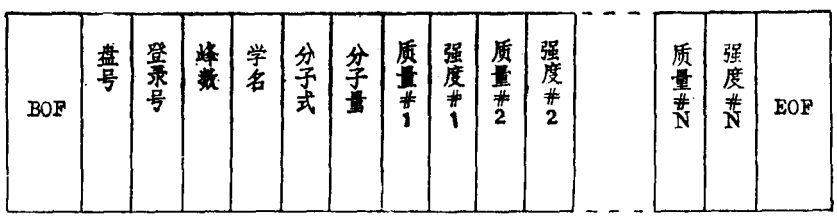
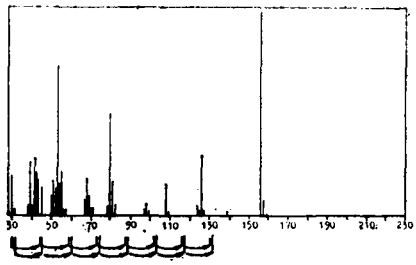


图 1　全谱文件格式　文件名：1：Ms×××



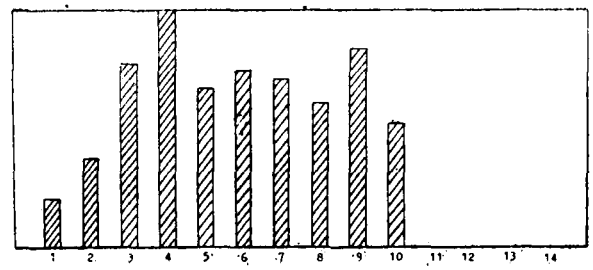


图 2 （a）谱图简化过程示意图

（b）简化谱

N-NITROSOGUVACINE

N-NITROSOGUVACINE

156, C_6, H_8, N_2, O_3,

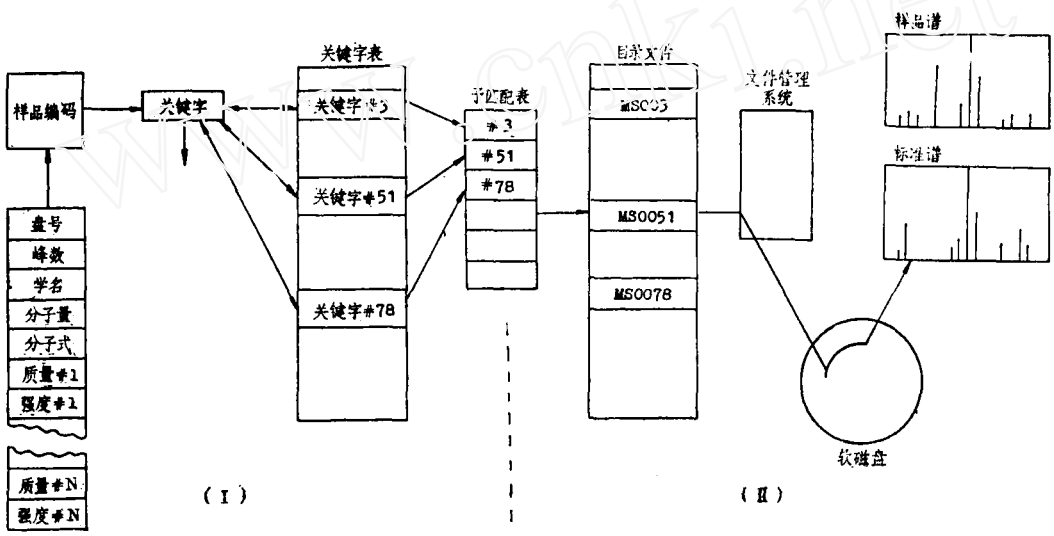


图 3 检索过程

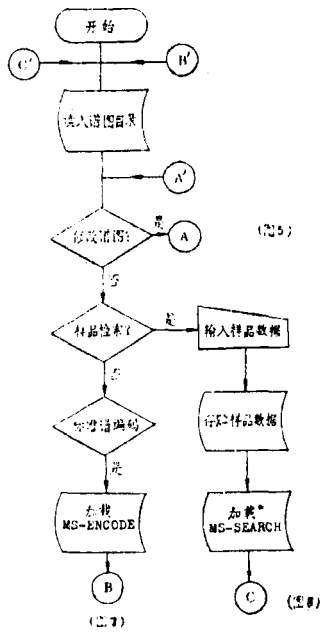


图 4 监督程序框图（包括复盖和链接）

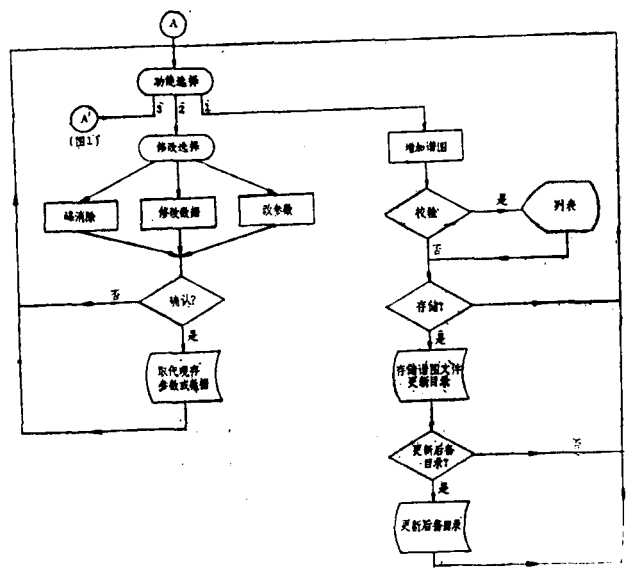


图 5 数据更新程序框图

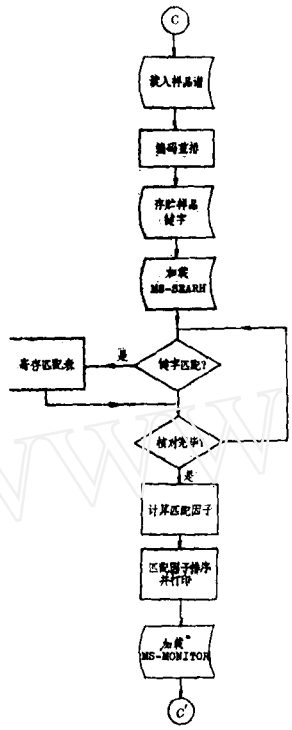


图 6　样品图谱检索程序框图

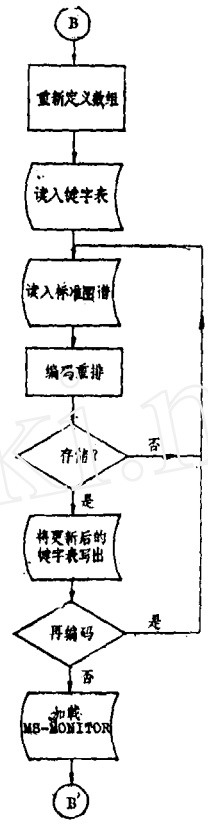


图 7　谱图编码程序框图

参　考　文　献

1. Venrataraghaven R., Klimowskc R.T., F.W.McLafferty, Acc. Chem. Res. 3 158-165 (1970)
2. (a) Herty H. S., Hite R. A., & Biemann K., Anal. Chem. 43 681-691 (1971)
   (b) Hite R.A.and Biemann K., Anal. Chem. 40 1217 (1968)
3. SWeeley C.C., Ray B.D., Wood W. I. and Holland J.F., Anal. Chem.42 1505-1516
4. Grotch S.L., Anal. Chem. 42 1214-1222 (1970)
5. Gregory.T。Raswvssen and T. L. Isenhour, J. Chem. Inf. Comput. Sci. Vol.19.№. 2, 88 (1979)
6. 王光辉、张文信、方一苇、卞则樑，化学学报，38 231-239 (1980)

A Micro-Computer Based Mass Spectral Library Search System

Liang Xiyun, Zhang Meiyi, Chen Zhifeng, Liu Jinkun
(Institute of Chemistry, Academia Sinica)

Abstract

We wish to introduce here a mass spectral library search system established on a personal microcomputer. This system allows a user to create his own spectral library containing a specific class of compound. It allows him to append and edit spectral data and descriptive information, as well as the generation and examination of keywords which are used in library searching. Finally, users can key in spectral data of unknown samples and search for matching compounds in the library. The Hardware configuration of the system is extremely simple and basically consists of a CBM/PET 2001 microcomputer connected to a PET-2040 5.5 inch floppy disk subsystem. The spectral library including data and descriptive information is stored in one diskette, each spectrum is represented as an individual file. All the programs for the management, maintenance and search of the data base together with the various index files are stored in another diskette. A modular approach is taken in programming the system and a two steps search technique is utilized in the search process.

中国质谱学会第一届理事会理事及常务理事

季　欧*	向鹏举*	张青莲*	王光辉*	陈国珍*
施士元*	邱纯一*	付桂香	王世俊*	肖桂里*
王永安*	尹翼开*	彭子成	黄知恒	王梦瑞*
刘炳寰	梁曦云	毛存孝	朱良漪*	梁晓天*
付道韫	姜龙飞	西门纪业	王子树	苏焕华
汤汉森	徐永昌	倪宝龄	胡炳森	方家骏
袁希召	汪聪慧	朱育芬	季桐鼎*	卢涌泉
查良镇	刘桂文	陶美娟	刘　琦	沈瑜生

台湾省一名

（有*者为常务理事）