

光谱数据库软件系统

谢狄霖^② 施伟巧^a 李勇波^a 林积荣 陈忠^b

[福建省医学测试重点实验室(福建省医学科学研究所) 福州市五四路 7 号 350001]

^a(福建省地质测绘院 福州市 350001)

^b(厦门大学化学系 福建省厦门市 361005)

摘 要 介绍了红外光谱及核磁共振波谱信息库系统的设计方案和技术关键。系统数据库包含纯化化合物红外光谱约 9 万多张,高聚物红外光谱 1.2 万张,药品红外光谱 1 千张,以及核磁共振氢谱 6 万多张,碳谱 4 万多张。可按光谱编号、化学名、商品名、原子数、分子式进行查询,还能根据未知物光谱图的谱峰形状进行检索。结果得到未知化合物的相关信息及其标准谱图。

关键词 光谱,数据库,软件。

中图分类号:O657.33;O657.2

文献标识码:A

文章编号:1004-8138(2006)01-0118-05

1 前言

未知样品的光谱与标准光谱图之间的峰-峰匹配比较即“指纹”核对通常是鉴定未知物的有效方法。如今人类已积累了十多万种物质的标准红外光谱图和核磁共振谱,要靠人工进行查询检索是非常艰难的^[1]。因此,我们进行了光谱数据库系统的开发研究。该系统数据库目前包含纯化化合物红外光谱 9 万多张,高聚物红外光谱 1.2 万张,药品红外光谱 1 千张,以及核磁共振氢谱约 6 万多张,碳谱 4 万多张。

系统采用 VB 编程语言编写而成,在 windows 平台上运行。整个系统程序包括光谱信息数据输入建库模块和光谱信息查询模块两大部分,以及光谱信息数据库和光谱图形数据库。各模块内部又成子模块结构,以利于系统功能的扩充和维护。其中光谱信息查询模块包含按光谱编号、化学名、商品名、原子数、分子式进行查询的子模块,以及根据未知物光谱图的谱峰形状进行检索的子模块。检索结果得到未知物的相关信息并显示其标准谱图。

2 光谱信息数据建库输入模块

光谱信息数据建库输入模块用于输入各标准光谱图的各种信息数据,包括光谱编号、化学名、商品名、原子数、分子式以及光谱图像,用于建立光谱数据库以供查询。光谱数据库采用 Microsoft Access 软件进行管理。红外光谱信息数据输入的用户界面如图 1 所示。用户只须键入标准光谱的商品名、化学名、原子数即可,而编号、卷标以及光谱图像文件名均由系统自动生成,分子式则由键入的原子数自行组合而成,从而大大提高用户建库输入的速度和质量。程序的查找、编辑、删除、添加等功能使用户可以方便地对数据库文件进行查阅、删改、增添等操作。用户可利用该模块将本专

① 基金项目:福建省自然科学基金(C0110024)资助项目;国家自然科学基金(20172042)资助项目

② 联系人,手机,013615050830;E-mail,xiedilin@sohu.com

作者简介:谢狄霖(1945—),男,福州市人,硕士,研究员,从事光谱仪器分析测试技术的开发研究。

收稿日期:2005-09-21;接受日期:2005-10-09

业的一些常用化合物的标准光谱的信息以及图谱添加进数据库以供日后查询。

图 1 光谱信息数据库输入的显示界面

3 光谱信息查询模块

光谱信息查询模块可以根据需要选择按光谱编号、中文名、英文名、原子数或分子式等查询条件中的一个进行查询。查询条件选择在“信息查询”界面中进行。查询结果得到中选光谱的各种信息,查询结果的显示界面如图 2 所示。勾选查询结果的显示界面右上角“显示图像”前的选择框即可显示中选的谱图。光谱图形数据库可以储存在硬盘中供调用显示,也可以存于光盘中,使用时插入系统提示卷名的光盘供调用。

图 2 光谱信息查询结果的显示界面

4 谱峰检索子模块

光谱信息查询模块不仅包含按光谱编号、化学名、商品名、原子数、分子式进行查询的子模块,还包括根据未知物光谱图的谱峰形状进行检索的子模块。光谱谱峰检索按峰-峰比较的方式进行,它能够根据未知样品光谱的主要特征峰的峰位和峰强进行检索,找出图谱数据库中与之最相近的若干种化合物的光谱。谱峰检索功能的实现比起按光谱编号、化学名、商品名、原子数、分子式等进行查询要复杂得多,对未知样品的实用意义也更大,故在此以红外光谱谱峰检索为例单独进行介绍。

4.1 谱峰数字化方法

红外光谱是连续量的曲线谱,而且特征信息集中,光谱数字化比较困难。我们采用谱峰编码的方法对光谱进行数字化处理。首先把整个图谱以 2100cm^{-1} 为界分成高波数区和低波数区两个部分,分别以 200cm^{-1} 和 100cm^{-1} 为间隔划分区段,然后利用各区段内最强吸收峰的波数的十位数字编制其代码,每个区段一个代码,并用该波数的千、百两位数字作为本区段代码的序号进行排列。若某区段内没有强吸收峰存在,则以“.”作为代码表示之。由于高波数区的区段间隔是 200cm^{-1} ,若某区段内最强峰的波数的百位数字为奇数,则应在十位数字的代码之后再加一个零,以区别于偶数的百位数字。这样,每张红外光谱都可以用 21 个码元素来表示。显然,这样编制的索引码既包含了光谱特征峰的峰位信息,又隐含了它们的相对峰强信息,反映了光谱的主要特点^[2]。除了谱峰的码元素之外,代码记录中还包含有该光谱在图谱集中的序号,化合物的熔点或状态,分子式等信息,组成一个代码条。它们共占 64bytes 的存贮单元。

数据库中各图谱来源不同,其光谱的强度无一定标准,且基线位置也各不相同,所以每张光谱在进行编码前都要进行数学变换,使其透过率介于 5% 与 95% 之间^[3]。

4.2 光谱代码数据库

代码数据文件按判别树结构进行组织,以图谱的最强吸收峰的波数值(四舍五入,精确到十位)为分叉。每个最强峰下的所有图谱数据单独组成一个数据文件,并利用该最强峰的波数值命名。则全部图谱代码条数据分为 96 个子数据库,组成系统的图谱代码数据库。图谱代码条数据以随机文件的方式储存,每两条图谱数据占据随机文件的一个记录段,共 128bytes。

系统设计有光谱代码数据库维护程序,可以让用户方便地对数据库中的代码条进行查阅、删改,用户也可以增添本专业的光谱的代码条和谱图,供日后检索比较用。

4.3 未知物光谱检索

4.3.1 样品谱峰数据输入

待检索的样品的谱峰在开始检索之前利用样品谱峰数据输入程序由用户按照图 3 所示用户界面的提示在键盘上输入各主要吸收峰的峰位和峰强等数据,组成一个样品峰表文件存于文件夹中供检索程序调用。

序号	谱峰波数	透过率
1	3125	44
2	2822	36
3	1788	24
4	1445	22
5	1284	11
6	1020	24

图 3 样品谱峰数据输入的界面显示

4.3.2 检索比较

通过比较未知物光谱代码与相应的某些标准光谱代码之间的相似性,可以实现未知物光谱的自动检索。红外光谱的谱线形状受样品的来源、制样技术、仪器状态、环境因素等影响很大,为提高检索的质量,系统除了要对样品光谱进行规范化数学变换,使之符合红外谱图规范的要求外,还要在光谱数据检索比较时对样品光谱进行模糊化处理,即容许各谱峰的波数值有 $\pm 10\text{cm}^{-1}$ 的偏差,透过率值有 $\pm 3\%$ 的偏差。

由于数据库包含的谱图数量很大,系统采用了“筛选器”技术对数据库的谱图进行粗筛,只有最强峰波数值和熔点状态筛选合格者才进行相似性比较。检索比较分高、低波数区分别进行,低区中选者才进行高区的比较。检索比较的相似性得分 G 按下列经验公式计算:

$$G = \text{INT}[10 - k - MT - |A_i - B_i| \cdot (2 - T_i)^3]$$

式中 INT 表示取整函数; k 表示最强峰的顺序号; A_i 与 B_i 分别代表样品光谱和库谱的第 i 个代码值; T_i 代表样品光谱第 i 区段内特征峰的透过率; MT 是决定于样品谱与库谱的熔点间差异大小的数值函数^[4]。

通过高、低波数区分别进行比较评分的检索方案,系统能对波数等间隔的光栅谱和早期波长等间隔的棱镜谱实现兼容检索。

4.3.3 检索结果

检索结果如图 4 所示,结果列出图谱数据库中 with 未知样品光谱最相近的若干种化合物的图谱高、低波数区的相似性得分,图谱序号,熔点或状态,分子式等。检索结果按低波区得分由高到低顺序排列,低波区得分相同者再按高波区得分高低顺序排列。用鼠标双击检索结果的某行,即可显示该行所示的标准光谱的图形供用户比较确认。

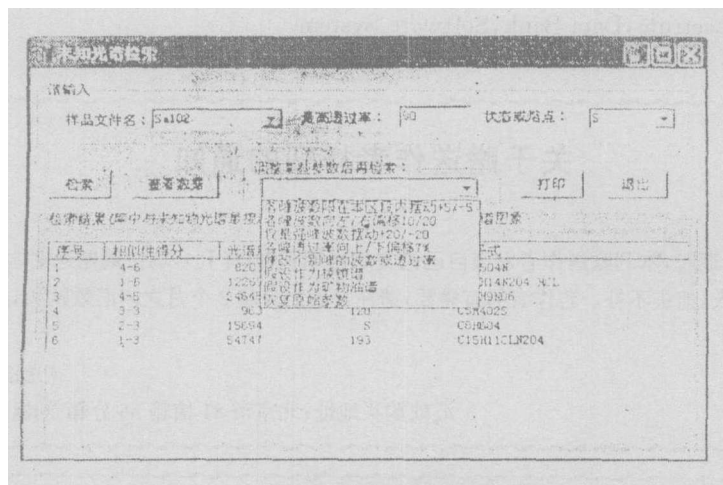


图 4 光谱谱峰检索结果的显示界面

4.3.4 调整参数后再检索

如果对检索结果不满意,程序提供各种再检索的调整方案菜单列表于检索结果界面中央供用户选择(见图 4),这些调整方案可单独选用,也可组合使用,并可多次反复进行,从而大大提高了系统的检出率。用户只需按照系统的提示,根据未知光谱可能存在的误差因素选击修正方案,系统就会自动开始重新检索,如此反复,直至得到满意的结果。

参考文献

- [1] 甘峰, 杨家红, 梁逸曾, 李晓宁. 用于结构解析的波谱智能数据库系统[J]. 计算机与应用化学, 2000, 17(2): 141--142.
- [2] Sadtler Research Laboratories. *IR Grating Collection Cumulative Spec-Finder Index*[M]. Philadelphia, U. S. A: 1995. 1--2.
- [3] 国家药典委员会. 药品红外光谱集[M]. 北京: 化学工业出版社, 2000. 9--10.
- [4] 谢秋霖, 林积荣, 张水冰. 红外光谱微机数据库[J]. 化学通报, 1989, 10: 53--55.

Spectrum Data Bank Software System

XIE Di-Lin SHI Wei-Qiao^a LI Yong-Bo^a LIN Ji-Rong CHEN Zhong^b

(Medical Science Research Institute of Fujian, Fuzhou 350001, P. R. China)

^a(Fujian Mapping Institute, Fuzhou 350001, P. R. China)

^b(Department of Chemistry, Xiamen University, Xiamen, Fujian 361005, P. R. China)

Abstract This paper introduces a spectrum data bank software system including more than one hundred thousand infrared spectra and more than ten thousand of nuclear magnetic resonance spectra of pure compounds. The data banks can be searched according to their serial number, chemical name, commercial name, amount of each atoms, or molecular formula, as well as their spectrum peak appearances. Programs for spectrum information inputting, spectrum information search and spectrum peak-peak match, and banks of spectrum information data, spectrum peak code data and spectrum figure data were attached to the system. System program was written by visual basic, and run under Windows system. The spectrum information data bank and spectrum figure data bank were administrated by Microsoft Access.

Key words Spectrum, Data Bank, Software System.

关于赠送作者样刊的通知

各有关作者:

从 2006 年第 1 期起, 本刊赠送作者发表自己论文的当期刊物(样刊), 均按篇赠送 2 本样刊, 用普通印刷品邮寄给作者联系人, 遗失不补。若作者另有需要, 请在发表之日起 2 个月之内汇款购买, 逾期不再办理。

特此通知

《光谱实验室》编辑部

汇款购买地址: 北京市 81 信箱 66 分箱 刘建林, 邮编: 100095

作者姓名更正启示

《光谱实验室》2005 年第 6 期题为《导数-比导数紫外分光光度法同时测量水样中硝酸根和亚硝酸根的含量》一文, 由于第一作者的粗心, 投稿时将第三作者黄溢谊误写为黄益溢, 在此特提出更正, 并向黄溢谊表示歉意。

作者: 任金响

2005 年 12 月 30 日