第五卷 质 谱 学 杂 志 Vol. 5
第三期 JOURNAL OF CHINESE MASS SPECTROMETRY SOCIETY №3

计算机辅助质谱解析的谱图匹配和模式识别方法

郭传杰

（中国科学院化学研究所）

〔摘要〕本文综述了计算机辅助质谱解析中常用的谱图匹配及模式识别方法的近期进展。扼要阐述了有关方法的原理，对某些具有代表性的系统作了简单介绍和评价。

近十几年来，计算机辅助解析质谱方面的活跃研究，导致发展了大量的方法和系统[1,2]。横观这些方法，本质大体相同。它们都试图让机器模拟人们解释质谱这一活动的基本过程，部分地代替人去推断未知谱图所代表的分子结构。因此，从方法论的角度看，建立各种系统的思维基础，不外乎归纳和演绎两个范畴。

通常，这些方法分为三类：谱图匹配、模式识别和人工智能。前两者均属于归纳，人工智能则属于演绎方法。鉴于已有专文[3]对人工智能的方法和系统作过详述，本文拟只讨论前两类方法和系统的近期进展和存在问题。

一、谱图匹配方法与系统

利用谱图—结构的相关性，借助计算机可迅速处理大量数据的能力，将未知谱和参考谱图集对照以获取未知谱的结构信息，是谱图匹配方法的出发点。到目前为止，大部分已获不同程度应用的系统，都是使用谱匹配的方法，方法各异，但都以信息论为其理论基础。

（一）主要研究的问题：

匹配系统主要追求的目标有三个：检出速度快，存贮空间省，可靠程度高。围绕这些目标，通常研究工作涉及五个方面。每个方面对目标的贡献各有不同，有时甚至相互矛盾。它们的相互关系如图 1 所示：

1984年 1 月27日收

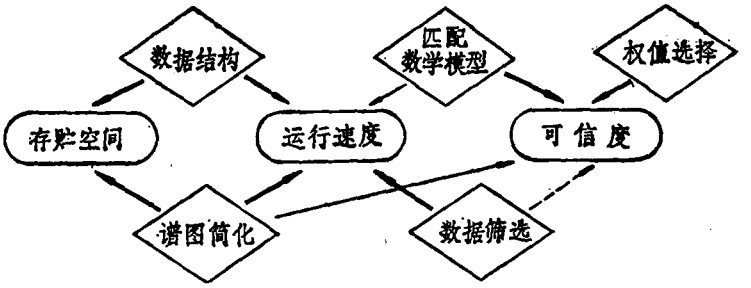

图1．匹配系统中各因素的相互关系
⟸贡献大，⟵有贡献，←…负贡献

研究表明，利用数学方法或质谱规律从全谱中抽提若干特征峰后形成的简约形式的谱图进行匹配检索，不仅为速度和存贮要求所必须，而且准确性更高。因为简化过程降低了冗余信息的干扰，提高了系统“容错性”。简化方法有四种。选择n个最强峰方法研究得最多最早[4]。二进制编码方式可有效减少存贮空间[5]，但最近研究证明[6]，从信息论角度看，这种编码方式不包含为检索目的而需要的足够信息。通过对全谱计算Khinchine函数[2]，用其单值表示一个谱图的方法，也进行过研究。直接利用质谱裂解规律选择特征峰的方法，受到越来越多的重视。

信息科学研究表明，适当运用加权方法，对提高数据集合中的信息检出可靠性，非常必要。质谱信息系统中，关于用总离子流标准化、强度平方根的Σ加和再标准化和局部区域再标准化等谱峰加权方法都进行过研究。PBM系统[7]通过对大量谱峰的统计学研究，获得质峰的权值。Dromey[8]建议了一种用恒定的分级因子或质量依赖因子进行动态调整的加权方法。

通过适当的数学模型，获得未知谱与参考谱的相似性判据，是检索系统的核心。大部分系统的模型都是涉及特征峰等参数的经验函数。

多级检索或适当的筛选程序对检索速度影响很大。谱库编序方法[9]、系列位移指数法[10]等都是有效的筛滤方法。最近，Shindo[11]等应用奇偶质量离子等七种筛选器，将最终匹配的参考谱减少到10%。但值得注意的是，不慎使用预选程序，会对结果可靠性带来负贡献。

（二）质谱数据库：

近年来，已产生了不少计算机可读的大型谱库，如NIH/EPA/MSDC、Registry of MS Data等。前者含有从七万谱图中优选出来的38805个不同化合物的质谱[12]，McLafferty等编汇的Registry of MS Data最近扩充了[13]一倍，拥有67510个不同化合物的79525张谱图，是目前国际上最大的质谱库。此外，许多实验室还开发了一系列适于自己需要的专业性谱库，文献[14]提供了一个查寻有关数据源的指南。

关于谱图质量的评价和控制问题，Speck[15]、Heller[12]等相继提出了使用若干质量因子的算法，并已用于各自的系统中。但值得指出的是，虽然许多人进行过长期的努力，但在谱库的质量、数量方面还有不少问题，在一个拥有数万张谱图的大库中发现少数质量低劣、甚至错误的谱图，是不足为怪的，这只能说明谱库扩充和质量控制工作的艰巨性。

（三）某些有代表性的系统：

七十年代初期，活跃的研究工作产生了大批实验室检索系统。在这些工作中，佼佼者是

MIT的Biemann的工作[16]。它的重要贡献在于，谱图简化时，不是任意挑选n个最强峰，而是利用了同系物的质谱基础，每14个amu质量区间内保留两个最强峰，这在一定程度上克服了质量歧视问题的影响。这一思想对后来工作影响颇大，被称为MIT—KB方法。他们的系统综合考虑了匹配峰数目、强度及权重因子，建立计算相似指数的模型：

$$SI=\frac{1}{n}\sum_{i=1}^{n}[(Rn/\overline{R}_{>10\%})W]/(F+1) \qquad (1)$$

我所建立的质谱信息系统[17]中，谱图检索子系统采用了西德煤炭所研究的SISCOM[18]方法。这一方法在考虑—CH_2同系列的同时，还利用强度参数考查了同系列邻近离子的情况，抽提特征峰的能力更强。系统规定，在经过^{13}C同位素等一系列预处理后，入选的特征峰必须超过上、下邻质峰（m±14）强度的算术平均值。系统先求算B值，再用式（3）得出的S值从150张粗选谱中确定最后输出结果：

$$B=Nc/(aN_R+bN_S+C) \qquad (2)$$

$$S=F_1(N_C)+F_2(P_C)+F_3(N_R)+F_4(I_R,I_S) \qquad (3)$$

式中，F_1是与参加匹配的特征峰数目、强度、图象相关性有关的权重经验函数。与K·Biemann系统相比，SISCOM更进一步减小了质量歧视效应，在检出相似结构化合物方面，该系统也有相当强能力，具有一定解析功能。

上述两系统都属于正向检索。它们一般难胜任混合物组份谱检出问题，因为“杂质”峰的存在会大大降低S值。为此，用参考谱同未知谱匹配的逆检索方法，也受到了重视。

运用逆检索原理建立的系统中，获最广泛应用的当推康奈尔大学的McLafferty小组的PBM[7,19]。基于某些离子峰的出现几率大大低于另一些峰这一基本质谱事实，通过对大量谱峰数据进行统计分析，给每一谱峰的质量和强度分别确定其唯一性值U和A，按下式计算简化后谱图的匹配置信度K：

$$K=\sum_{j=1}^{n}(U_j+A_j-D+W_j) \qquad (4)$$

式中，W、D分别是考虑杂质和峰值波动范围的修正值。改进后的系统综合考虑K值大小以及标记峰数目、分子峰是否参与匹配、ΔK值大小和倾斜函数等多参数影响，求出可靠性，作为输出判据[20]。系统经一系列改进之后，已实现实时联机运行[21]。最近，通过对近八万谱图的大库进行峰分布统计分析[13]，我们发现，U、A值与以前基于三万谱图的统计结果基本相同，这证明PBM抽提特征峰的方法有较可靠的统计基础和质谱基础。因此，近来发展的一些系统，常使用该法的基本原理[11]。不过，由于PBM的设计思想只针对鉴定性检索，因此，解析能力比SISCOM差。

近年来，鉴于正、逆检索各具优缺点，又出现了一些兼具正、逆检索功能的小型系统。例如，Terwilliger[22]等发展的STAR系统，匹配时用未知谱的n个最强峰与参考谱的八峰进行比较，通过对n的数目由大到小动态调整，系统巧妙地由正检索自动过渡到了逆检索。

目前，各种不同系统虽然发展了很多，但真正获得较好应用的系统则不多见。而且，对各系统的严格评价仍是个有待解决的课题。McLafferty[23]、Rasmssen[24]等都提出过评价方法，但实施相当困难，因为不同系统的适应范围和操作特性与数据库规模、谱库构成、预检索条件等等许多因素有关。至今尚未出现过能满足全部需要的“万能”系统，这也可能

是不现实的。

二、模式识别方法与系统

用数学语言来说，模式就是具有相同属性的某些现象的集合。模式识别是研究用机器模拟人对不同模式进行描述、分类、识别等有关的理论和方法。七十年代中，这一方法引入质谱解析领域，开展了大量研究工作[25,26]。它在确定未知谱的某些结构特征、提供分类信息方面，有一定作用。

质谱解析中应用的模式识别方法，可分为三种：聚类分析、因子分析和学习机方法。这些方法的基本步骤相同：将参考谱分类，分别表示为多维空间中的点，基于这些空间点发展各种分类器，最后用相应的分类器处理未知谱图，获得有关结构信息。

（一）聚类分析方法：

质谱解析中的聚类分析法实际与匹配检索相似。其根据是，当质谱表示为多维空间中的点时，相似结构化合物将集中于该空间的相同区域。K-最近邻域法（KNN）是聚类分析中的典型方法[26]。它计算未知谱与每参考谱间的距离，检出K个最相近的参考谱，未知物与它们同类型。与别的聚类分析法相比[27]，K-NN法在处理不同源谱图方面，有较好的容错性能。Ziemer[28]等应用KNN法鉴定二肽甲基酯衍生物，确定C-端基和N-端基氨基酸的正确率分别达到97%和100%。McGill[27]对预测氧存在与否的许多方法进行比较考察，认为KNN法较优，其缺点是需较长的分析时间。

另一种属于聚类分析的技术是非线性映照（nolinear mapping）[29]。鉴于人对高维空间的图象缺乏直观认识，于是研究了用计算机将谱图的这些点集合投影到低维空间（二维或一维），并使这些集合依然保留在高维空间时的相对特征（距离、方向等），再对低维空间进行作图分类处理。有人[30]研究用此法确定核苷酸系列结构。不过，Rasmnssen[31]研究发现，用非线性映照不可能实现完全的分类。

（二）因子分析方法：

因子分析法用于质谱解析[32]，是基于解析时无须谱图中的全部信息这一事实。它将同类化合物的谱图，表示为n×m的数据矩阵。经过数据变换，算出相关矩阵m×m。对该矩阵进一步分析，抽出因子，直至剩余方差接近实验误差。这些因子然后就作为该类化合物谱的简约表式。通过对22个$C_{10}H_{14}$烷基苯异构体的分析，Rozett[23]等发现了三个基本因子。研究表明，因子分析用于判断核苷酸的存在[34]以及从GC-MS数据中抽提纯组份谱图[35]是有一定意义的。

（三）学习机方法：

为使计算机进行学习，先要确立一个“训练集”[26]。将表示为多维空间中点集的参考谱，依其是否具有某结构特征，分为二类。希望在此特征空间中找到一个超平面，使这二类化合物分居于平面两侧，使用线性判别函数G：

$$G(\underline{X}) = W_1X_1 + W_2X_2 + \cdots\cdots + W_dX_d + W_{d+1} \qquad (5)$$

式中，$\underline{X}$是由$X_1, X_2 \cdots X_d$和$W_1, W_2 \cdots W_{d+1}$定义的质谱矢量。试定义一权矢量W，使其用于上述二类谱图时，分别得正值或负值。反复调整权矢量中的每一权值，直至获得全部参考谱的正确分类。如果训练集内的谱图是线性可分的，过程将很快收敛，得到一恰当的权矢量，可用于未知谱分类。否则，过程将不收敛。七十年代早期，线性学习机曾有许多应

用[2]，但其后使用渐少，这是因为预报效果与训练集大小有较大关系。有人证明[36]，训练集中的点数必须远大于表示每个元素的维数。另一原因是该法要求的前提是数据的线性可分性，而实际情形往往并非如此。虽用付里叶变换或自动相关技术[27]等可获得一些改进，但效果不十分明显。

上述方法是一种二元分类器，只可检测单一结构特征是否存在的问题。连续使用多个二元分类器，可形成一决策树，能进行多元分类。有些作者[39]已研究过用决策树确定化合物分子式、元素组成等问题。各种学习机方法的共同优点是，一旦完成训练之后，分类过程进行极快。然而，完成训练是件比较困难的事。

上面所有这些基于模式识别原理建立的系统，都未获得真正的实际应用。原因可能是这些系统一般只考虑数学基础，很少利用质谱本身的经验和理论。从功能上讲，Mclafferty小组研究的STIRS[38]（"自训练解析与检索系统"）也属于模式识别范畴。由于它的解析能力，是模式识别方法中获得最广泛应用的一个成功系统，与PBM系统一起，通过TYMNET网络[39]，已在北美、西欧、日本的250多个实验室获得应用。

STIRS具有完全不同的训练基础，正如质谱本身是用各种软、硬离子化手段将分子裂成碎片，再通过对碎片离子的鉴定而达到确认整体分子结构的过程一样，STIRS可以说是"质谱的质谱"。它通过计算机将质谱数据分为特征离子、同系列、中性丢失及其相互复盖等18类数据，分别建库，然后按下述公式计算不同数据类的匹配因子，根据对结果的分析及WLN的提示，预测未识谱的结构信息。

$$MF_2 = 1000\sum_{k=1}^{M} R_k\,(I_i + I_j)_k / \left(\sum_{i=1}^{m} I_i + \sum_{j=1}^{n} I_j\right) \qquad (6)$$

$$MF_{11.0} = \frac{f_i MF1 + f_c \sum_i N_i MF_i^{c} + f_n \sum_j N_j MF_j^{n}}{f_i + f_c \sum_i N_i + f_n \sum_j N_j} \qquad (7)$$

改进后的系统具有预测分子量[40]、元素组成[41]、环加双键等功能。最近[42]，我们将STIRS专用数据库的化合物扩充了一倍，同时，按随机抽样模型

$$P_{(J)} = \frac{N!}{J!(N-J)!}\rho^{J} \cdot (1-\rho)^{N-J} \qquad (8)$$

对18类数据分别预测597种亚结构的可靠性进行了概率统计分析，进一步提高了系统预测未识谱图结构信息的能力。

结　束　语

计算机辅助质谱解析领域经十余年大量的研究，已经取得许多开创性的成就，MIT-KB、MSSS[12]、SISCOM、PBM、STIRS等系统各在不同程度上获得了国际范围的应用。斯坦福大学研究的著名DENDRAL[43]人工智能系统，无论是在设计思想还是程序设计技术方面，都博得了质谱学界和计算机科学界的赞誉。

值得补充说明的是，一些质谱数据检索系统，名称上虽也叫"数据库检索"，实则与计算机科学中的"数据库检索"不同。首先是数据结构不一样。质谱数据系统通常使用比较简单的顺序或随机文件存贮方式，不用数据库的数据结构。其次是功能不一样。设计比较先进的质谱检索、解析系统，如SISCOM、STIRS等，不仅能单纯对应地找出库中某一个或某

一批数据，而且，通过对多种特征的分析比较，能向用户报告某些经过推测的可能结果，也就是说，具有某种程度的“专家系统”功能。只是因为“专家系统”、“演绎数据库”、“知识工程”等名词在七十年代末期才出现，所以质谱数据系统还袭用着原来的“检索”一词。

当然，计算机解析质谱只能是人们解析质谱的一个助手。而且，由于各种方法本身的局限，目前所有这些系统都各有其缺点和困难，近年的研究工作也不如七十年代中活跃。谱图匹配对没有参考谱的未知物解析能力虽不断提高，但终究受到限制；模式识别方法目前只能在结构分类方面有意义，作为完整结构解析工具，至少前景不明；人工智能系统有着诱人的前景，但各种不同化合物质谱的复杂性，使它至少在目前的技术条件下难以获得普遍应用。不过，所有这些问题和困难正好说明这一研究领域继续存在巨大潜力和机会，新思想的挑战和新技术的引入，可望引起新的突破。

参 考 文 献

〔1〕 F. W. McLafferty and R. Venkataraghavan, J. Chrom. Sci., 17, 34 (1979).
〔2〕 J. R. Chapman, “Computers in Mass Spectrometry”, Academic Press, New York, 1978.
〔3〕 朱大模、质谱，1，(1982).
〔4〕 S. L. Grotch, Anal. Chem., 43, 1362 (1971).
〔5〕 G. Van Merlen, A. Dijkstra et al, Anal. Chim. Acta, 112, 143 (1979).
〔6〕 G. Van Merlen, A. Dijkstra et al, Anal. Chem., 51, 420 (1979).
〔7〕 G. M. Yesyna, F. W. McLafferty et al, Anal. Chem., 48, 9 (1976).
〔8〕 R. G. Dromey, Anal. Chim. Acta, 112, 133 (1979).
〔9〕 R. G. Dromey, Anal. Chem., 51, 229 (1979).
〔10〕 R. G. Dromey, Anal. Chem., 48, 1464 (1976).
〔11〕 J. Shindo, A. Yasuhara, H. Ito and T. Mizoguchi, Chem. Letter, 4. 521 (1982).
〔12〕 G. W. A. Milne, S. R. Heller et al, Org. Mass Spectro., 17, 11, 547 (1982).
〔13〕 I. K. Mun, D. B. Stauffer, W. Staedeli, C. -j Guo and F. W. McLafferty, 30th Annual conference on M. S. & Allied Topics, Honolulu, June 6—11, 1982.
〔14〕 D. P. Martinsen and M. A. Grayson, “A Guide to Collections of Mass Spectral Data”, 2nd ed., 1978.
〔15〕 D. D. Speck, R. Venkataraghavan and F. W. McLafferty, Org. Mass Spectrom., 13, 209 (1978).
〔16〕 H. S. Hertz, R. A. Hites and K. Biemann, Anal. Chem. 43, 681 (1971).
〔17〕 梁曦云，刘津琨，陈稚芳，郭传杰，D. Henneberg and B. Weimann，第一届全国计算化学会议，兰州，1984.
〔18〕 H. Damen, D. Henneberg and B. Weimann, Anal. Chim. Acta, 103, 289 (1978).
〔19〕 F. W. McLafferty, R. H. Hertel and R. D. Villwock, Org. Mass Spectrom. 9, 690 (1974).
〔20〕 B. L. Altwater, Ph. D. Thesis, Cornell Uni., 1980.
〔21〕 F. W. McLafferty, C. Sheauchi, K. M. Dully, C. -j Guo et al, Int. J. of Mass Spectrom. & Ion Phys., 47, 317 (1983).
〔22〕 T. Terwilliger, 28th Annual Conference on Mass Spectro. and Allied Topics, New York, May 30-June 2, 1980.
〔23〕 F. W. McLafferty, Anal. Chem., 49, 1441 (1977).
〔24〕 G. T. Rasmussen and T. L. Isenhour, J. Chem. Inf. Comput.Sci., 19, 179(1979).
〔25〕 P. C. Jurs and T. L. Venhour, “Chemical Application of Pattern Recognition”, Wiley Interscience Publication, 1975.
〔26〕 B. R. Kowalski and C. F. bender, J. Am. Chem. Soc., 94, 5632 (1972).
〔27〕 J. R. Mcgill and B. R. Kowalski, J. Chem. Inf. Comput. Sci, 18, 52 (1978).

〔28〕 J. N. Ziemer, S. P. Perone et al, Anal. Chem., 51, 1732 (1979).
〔29〕 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95, 686 (1973).
〔30〕 D. R. Bergard, S. P. Perone and J. L. Wiebers, Biochemistry, 16, 1051 (1977).
〔31〕 G. T. Rasmussen, G. L. Ritter et al, J. Chem. Inf. Comput. Sci., 19, 255(1979).
〔32〕 R. W. Rozett and E. M. Peterson, Anal. Chem. 47, 1301 (1975).
〔33〕 R. W. Rozett and E. M. Peterson, ibid, 47, 2377 (1975).
〔34〕 D. R. Burgard, S. P. Perone and J. L. Wiebers, Anal. Chem., 49, 1444 (1977).
〔35〕 F. J. Knorr and J. H. Futrell, Anal. Chem., 51, 1236 (1979).
〔36〕 N. A. B. Gray, Anal. Chem. 48, 2265 (1976).
〔37〕 W. S. Meisel, M. Jolley, S. R. Heller and G. W. A. Milne, Anal. Chim. Acta, 112, 407 (1979).
〔38〕 K. -S. Kwok, R. Venkataraghavan and F. W. McLafferty, J. Am. Chem. Soc., 95, 4185 (1973).
〔39〕 Office of Computer Services, Cornell Uni., Ithaca, New York 14853.
〔40〕 I. K. Mun and F. W. McLafferty, Anal. Chem., 53, 179 (1981).
〔41〕 I. K. Mun and F. W. McLaffertr, ibid, 49, 1723 (1977).
〔42〕 D. B. Stauffer, M. A. Sharaf, C. -j. Guo and F. W. McLafferty, 31th Annual Conference on Mass Spectrom. & Allied Topics, Boston, May 8-13, 1983.
〔43〕 N. A. B. Gray, R. E. Carhart, D. H. Smith et al, Anal. Chem., 52, 1095 (1980).

Spectra Matching and Pattern Recognition Technique for Computer Elucidation of Mass Spectra

Guo Chuanjie

(The Institute of Chemistry, Academia Sinica)

Received 27, Jan., 1984

Abstract

Recent studies on library search and pattern recognition methods which are commonly used in computer-aided interpretation of mass spectra have been reviewed. The basic principles of them are briefly discussed. The evaluation of performance of some typical systems have also been given.