一种新型交替迭代算法用于农药 GC-MS 信号快速分析

梅桢 蔡文生 邵学广*

(南开大学化学学院 天津 300071)

摘要 化学计量学算法为重叠 GC-MS 信号解析提供了有效手段. 免疫算法(IAs)可根据标准样品的信号(色谱或质谱)对重叠 GC-MS 信号进行解析并得到重叠信号中各组分的信息. 但是标样信号与实际测量信号之间的差别会造成解析结果出现偏差. 针对标样信号与测量信号不一致的问题, 作者提出了一种交替迭代算法用于重叠 GC-MS 测量信号的直接解析. 该方法在不提供标样信号的情况下可以解析出组分的质谱和色谱信息. 计算过程中, 采用随机产生的质谱作为初始输入, 利用最小二乘和 IA 算法交替计算质谱和色谱, 直到满足终止条件. 采用所建立的方法对 40 种组分的农药混合物进行了分析, 使用快速升温程序在 10 min 保留时间内得到了全部组分的色谱和质谱信息.

关键词 气相色谱-质谱联用; 免疫算法; 重叠峰解析; 农药混合物

Rapid Analysis of Pesticide Mixture by Gas Chromatography–Mass Spectrometry with a New Alternative Iterative Algorithm

Mei, Zhen Cai, Wensheng Shao, Xueguang*

(College of Chemistry, Nankai University, Tianjin 300071)

Abstract Chemometric methods have been proved to be a powerful tool for resolution of overlapping signals. Immune algorithm (IA) is one of the chemometric approaches for analyzing multi-component GC-MS signals. The method extracts the information of the components by iteratively eliminating standard information (chromatogram or mass spectrum) from overlapping GC-MS signals. In the primary IA, however, the standard signal of each component possibly contained in the mixture must be provided by measuring the standard or by theoretical simulation. When there is difference between the measured signals of the standards and the mixture, distortion and negative values will appear in the resolved chromatograms. In order to conquer the problem, an algorithm based on an alternative iteration of least squares fitting and IA was proposed in this work. In the method, the measurement of GC-MS was achieved with a very fast temperature program to make the analytes to elute within a short retention time period, and then the chromatographic and mass spectral information of the components in the overlapping signal is calculated with the proposed algorithm. In the calculation, the algorithm takes random mass spectra of the components as the starting input, and then the information of each component is obtained with an alternating iterative process. In the iteration, the mass spectra are calculated by using the least squares fitting and the chromatographic profiles are resolved by IA. Furthermore, the non-negative and unimodality constraints are adopted in the calculation for improving the resolved results. The iteration stops when the remaining signal does not change. The feasibility of the method was validated by using a simulated GC-MS data matrix of a three-component mixture, and the practicability of the method was proved by resolving the GC-MS data of the 40-pesticide mixture. The results show that both the mass spectra and the chromatographic information of the components were extracted from the overlapping signals, and the pesticide mixture was analyzed within 10 min elution with the help of the proposed method.

Keywords gas chromatography-mass spectrometry; immune algorithm; resolution of overlapping peak; pesticide mixture

1 引言

复杂样品分析一直是分析化学中具有挑战性的课题. 化学计量学算法为复杂体系的定性定量分析提供了有效的手段. 其中, 傅里叶变换、卡尔曼滤波、小波变换在重叠信号分辨等方面得到广泛应用^[1~3]. 这类算法的缺点是它们只能用于一维信号的处理. 为了解析二维信号, 特别是联用仪器信号, 提出了很多基于多元统计

的化学计量学算法, 如化学因子分析(CFA). 这些算法已广泛应用于多组分重叠信号的解析^[4~6]. 结合二维数据和交替最小二乘的优势, 提出了多元曲线分辨-交替最小二乘(MCR-ALS)算法, 该算法通过迭代运算得到混合信号中各组分的信息^[7~9]. 尽管该算法已经得到了许多成功应用, 但在有未知干扰存在的情况下往往无法得到正确的解析结果. 为了解决这一问题, 提出了直接三线性分解(DTD)、平行因子分析(PARAFAC)和交替三

* E-mail: xshao@nankai.edu.cn

Received January 9, 2013; published March 8, 2013.

Project supported by the National Natural Science Foundation of China (No. 21175074).

项目受国家自然科学基金(No. 21175074)资助.

线性分解(ATLD)等方法^[10~12]. 这类算法具有二阶优势, 能够在未知干扰物存在下提取重叠信号中的有用信息.

免疫算法(IAs)是一类用于解析重叠信号的算法^[13,14]. 最初的 IA 必须提供混合物中所有被测物标样信息(色谱或者质谱), 不适用于未知体系的解析. 为了解决标样提供的问题, 提出了自适应免疫算法(AIA)^[15]. AIA 无需提供标样信息, 而是采用独立成分分析(ICA)直接从测量数据中提取被测物信息. 但是, 某些情况下, 重叠信号不满足独立性要求, 会导致 IA 得到的质谱出现负值, 色谱出现扭曲. 因此又提出了不同的改进算法, 比如平均场-独立成分分析(MF-ICA)、基于化学知识的后旋转策略(PMBK)以及窗口独立成分分析(WICA)等^[16~18]. 为了实现复杂体系高通量分析, 还建立了非负免疫算法(NNIA), 实现了混合信号中某一组分信息的提取^[19]. 非负免疫算法与目标因子分析(TFA)结合用于重叠信号中某些特定组分信号的识别与解析^[20]. 该算法首先利用 TFA 和标样质谱库检验出实际体系中存在的组分, 然后利用 TFA 计算得到的质谱取代原有的标样质谱用于解析.

免疫算法的基本原理是从重叠化学信号中逐步扣除标样信号, 扣除量由重叠信号在标样信号上的投影决定, 直到该投影值趋向于零. 根据免疫算法的原理, 标样信号的提供至关重要. 只有标样信号与测量信号一致, 才能正确地解析出色谱或光谱. 但是无论是实验得到的标样信号, 还是数据库(如质谱库)中的信号都可能与实际测量信号有所差异, 造成解析出结果出现偏差. 为了解决这一问题, 本文提出了一种重叠 GC-MS 测量信号的直接解析方法. 该方法无需提供标样信号, 由随机产生的质谱开始, 采用迭代运算得到各组分的色谱和质谱信息. 算法的初始输入是一组随机产生的质谱, 质谱个数等于实际组分数, 然后利用交替迭代过程解析出重叠信号中所有组分的质谱和色谱. 在每次迭代过程中, 组分的质谱由最小二乘计算得到, 利用得到的质谱作为标样信号, 由 IA 计算得到色谱. 解析过程中, 对质谱进行非负限制, 并做归一化处理, 同时用非负和单峰对迭代所得的色谱进行限制^[21].

2 结果与讨论

2.1 模拟数据解析过程

为了检验本方法的原理和可行性, 模拟了一组 3 组分的 GC-MS 数据. 图 1 是模拟数据的 GC-MS 图. 其中, 图 1(a)中黑色曲线是模拟的总离子流图, 红色曲线 1、蓝色曲线 2 和紫色曲线 3 分别是模拟组分的色谱图, 图 1(b1), (c1)和(d1)是模拟组分的质谱图. 模拟的色谱乘以质谱得到模拟的 GC-MS 数据. 最后随机产生一个最高强度为信号强度 5% 的随机噪声矩阵添加到数据矩阵中作为噪声.

图 1 模拟和解析的色谱、质谱图. 模拟和解析的色谱图(a), 模拟组分的质谱图(b1), (c1)和(d1), 模拟组分的解析质谱图(b2), (c2)和(d2)
Figure 1 Simulated and resolved chromatograms and mass spectra. Simulated and resolved chromatograms (a), simulated mass spectra (b1), (c1) and (d1), resolved mass spectra (b2), (c2) and (d2)

为了说明迭代算法的解析过程, 图 2 是第 2 个组分在不同迭代次数时解析得到的色谱图, 左上角的小图是色谱曲线顶部的局部放大. 图中黑色曲线是模拟色谱曲线, 暗黄色的点划线是最终解析结果, 曲线 1, 2, 3, 4, 5, 10, 15, 20 分别代表不同迭代次数所得色谱曲线.

图 2 第 2 组分在不同迭代次数时解析得到的色谱图, 左上角是色谱顶部曲线的局部放大图
Figure 2 Resolved chromatograms of the 2nd component obtained by different number of iterations, the amplificatory figure of the top of resolved chromatograms in the upper left corner

从图 2 可以看出, 最初的几次迭代解析得到色谱从初始值开始快速接近真实值, 然后解析结果在真实值附近震荡, 经过多次迭代(通常在 1000 次左右)可以达到终止条件(剩余信号中各组分的色谱强度均小于 10^{-6})后, 结果趋向于真实值. 解析结果绘制于图 1 中, 其中 1(a)中用不同颜色的点划线代表三种组分的色谱解析结果, 图 1(b2), (c2)和(d2)是解析得到的质谱. 从图 1(a)中可以看出, 解析得到的色谱和模拟的色谱几乎完全重合. 解

析得到质谱与模拟质谱的匹配率分别是 987%, 994%和 991%。这些结果可以说明本文方法用于 GC-MS 重叠信号解析的可行性。

2.2 农药混合物 GC-MS 数据解析

40 种农药混合物的 GC-MS 信号采用快速升温程序测量, 所有物质在 10 min 之内全部洗脱。由于洗脱快速, 所以存在数个重叠色谱峰区域。本文选择保留时间为 4.53~4.63 min 和 5.30~5.37 min 的区域为例讨论重叠信号的解析过程。为了避免其他组分的干扰, 计算过程中只采用了所选择时间区间的信号。

图 3 是保留时间为 4.53~4.63 min 区域的解析色谱图和质谱图, 其中图 3(a)中的黑色曲线是测量信号的总离子流图, 不同颜色的五星点划线(1~4)代表采用 NIST05 质谱库中的质谱作为标样信息解析得到的色谱, 而暗黄色五星点划线代表解析后的剩余信号。从图中可以看出, 解析得到的各组分色谱信号发生了扭曲, 如组分 1 的解析色谱出现了扭曲和拖尾现象。同时, 将各组分信号加和可以得到解析结果的总离子流曲线, 与测量信号的总离子流曲线相比可以发现解析出的信号总强度与测量信号有明显不同。这是因为质谱库中质谱与测量信号中该组分的质谱信息不同, 造成 IA 的解析结果与测量信号出现偏差。图 3(a)中的曲线 1', 2', 3'和 4'分别

图 3 保留时间为 4.53~4.63 min 区域的农药混合物 GC-MS 数据的解析色谱图(a)和质谱图(b)~(e)

Figure 3 Resolved chromatograms (a) and mass spectra (b)~(e) of the components from the 4.53~4.63 min retention time region of the experimental GC-MS data

代表本文方法解析得到的色谱, 橙色圆点点划线(5)是解析得到的总离子流曲线。可以看出, 解析结果与测量信号具有较高的一致性。同时, 各组分的色谱峰也具有较好的对称性。根据样品的组成和保留指数可知, 该重叠区域中含有 4 种物质, 分别是虫螨威(Carbofuran), 莠去津(Atrazine), 乐果(Dimethoate)和六氯苯(HCB)。图 3(b)~(e)是本文方法解析得到的质谱图。采用 NIST05 库中质谱进行匹配, 匹配率分别是 766%, 785%, 524%和 895%。显然, 匹配率较低的原因是质谱库中的质谱与测量信号存在较大差异。为了验证结果的正确性, 采用本文的仪器和实验条件对虫螨威、莠去津和六氯苯的质谱进行了测量, 然后与解析结果进行了匹配率计算。结果显示三个组分的匹配率分别为 931%, 933%和 893%。

图 4 是保留时间为 5.30~5.37 min 区域的解析色谱图和质谱图, 其中图 4(a)中的黑色曲线是总离子流图, 不同颜色的五星点划线(1~2)代表采用 NIST05 质谱库中的质谱作为标样信息解析得到的色谱, 而暗黄色五星点划线代表解析后的剩余信息, 而曲线 1'和 2'分别代表本文方法解析得到的色谱, 橙色圆点点划线(3)是解析得到的总离子流曲线, 图 4(b)和(c)是本文方法解析得到的质谱。显然, 与采用质谱库中质谱作为标样信息解析

图 4 保留时间为 5.30~5.37 min 区域的农药混合物 GC-MS 数据的解析色谱图(a)和质谱图(b 和 c)

Figure 4 Resolved chromatograms (a) and mass spectra (b and c) of the components from the 5.30~5.37 min retention time region of the experimental GC-MS data

得到的总信号强度相比,采用本方法解析得到的总信号强度更接近测量信号。同样,采用本文方法得到的各组分解析色谱具有较好的对称性。经确认这两个组分分别是杀扑磷(Methidathion)和苯线磷(Fenamiphos),其匹配率分别是 771%和 809%。其中杀扑磷匹配率较低的原因也是由于质谱库中的质谱与测量信号存在较大差异。

由于分离实验采用了快速的升温程序,除个别组分外大部分组分的色谱都出现了与其他组分的重叠。采用本方法对所有组分的信息进行解析的结果表明,除三氯杀螨醇(Dicofol),由于信号强度过低没有检出外,其他成分的信号都得到了较好的解析。此外,本文方法对于质谱非常相似的同分异构体也无法进行识别,需要借助保留指数信息进行进一步核实。

3 结论

针对标样信号与测量信号不一致的问题,本文建立了一种用于混合物 GC-MS 快速分析的迭代算法。该方法采用随机产生的质谱作为算法输入,通过最小二乘和 IA 算法交替迭代得到各组分的质谱和色谱。利用模拟数据和 40 组分农药混合物的 GC-MS 数据检验了方法的可行性和实用性。但对于质谱高度相似的组分,该方法仍需要借助保留指数等其他信息。

4 实验部分

4.1 模拟数据

为了说明本文方法的原理和可行性,模拟了一组 3 组分混合物的 GC-MS 数据。其中 3 个组分的色谱用指数修正的高斯公式(EMG)^[22]模拟,由 100 个采样点构成,质谱由 20 个质荷比通道构成。因此所模拟的 GC-MS 数据为 100×20 的数据矩阵。模拟数据中还增加了 5% 的随机噪声。

4.2 实验数据

样品为 40 种农药组成的混合物,浓度均为 10 ppm。采用 GC-MS 分析仪(PolarisQ 2000, Finnigan)进行测试,色谱柱为弱极性柱(CP-Sil8CB, 30 m 长,内径 0.32 mm,液膜厚度 0.25 μm , VARIAN, USA),载气为氦气(BOC, 99.999%),流速为 1.0 $\text{mL}\cdot\text{min}^{-1}$ 。初始柱箱温度设置为 50 $^{\circ}\text{C}$,保持 2 min,然后以 100 $^{\circ}\text{C}\cdot\text{min}^{-1}$ 的梯度升到 300 $^{\circ}\text{C}$,并保持 6 min。GC 进样口温度为 280 $^{\circ}\text{C}$ 。电子轰击源设置为 70 eV,质量检测范围从 50 到 650 mu 。为了保持质荷比与数据序号的一致性,将前 49 个质谱通道数值设置为 0,因此分析结果的数据为 685×650 的矩阵。

4.3 计算

本文方法的计算原理是从随机产生的质谱信息出发,经迭代得到组分的质谱和色谱信息。在每次迭代中,经过最小二乘拟合计算得到质谱信息,然后再利用免疫算法计算色谱。为了保证计算的收敛性和加快收敛速度,在计算质谱时增加了非负限制,并做归一化处理,而在计算色谱时采用了非负和单峰限制^[21]。因此,本文的计算过程可描述为:

(1)随机产生各组分的质谱,质谱数为假设的组分数。

(2)采用最小二乘拟合计算各组分的质谱信息,实施非负限制,并作归一化处理。

(3)利用计算得到的质谱作为免疫算法的输入计算各组分的色谱曲线,并实施非负和单峰限制。

(4)重复(2)和(3)步骤的计算直到免疫算法计算中剩余信号不再发生变化。

由于实际测量信号的复杂性,计算过程中将数据矩阵按保留时间顺序划分为 14 个保留时间区间依次进行了解析。同时,为了减小背景对计算的干扰,本文采用了自适应迭代加权惩罚最小二乘(airPLS)法^[23]对实验信号中的背景进行了扣除。

References

- [1] Shen, Y. F.; Zhao, R.; Belov, M. E. *Anal. Chem.* **2001**, *73*, 1766.
- [2] Chen, J. C.; Rutan, S. C. *Anal. Chim. Acta* **1996**, *335*, 1.
- [3] Shao, X. G.; Leung, A. K. M.; Chau, F. T. *Acc. Chem. Res.* **2003**, *36*, 276.
- [4] Maeder, M. *Anal. Chem.* **1987**, *59*, 527.
- [5] Malinowski, E. R. *J. Chemom.* **1992**, *6*, 29.
- [6] Kvalheim, O. M.; Liang, Y. Z. *Anal. Chem.* **1992**, *64*, 936.
- [7] Vives, M.; Gargallo, R.; Tauler, R. *Biopolymers* **2001**, *59*, 477.
- [8] Li, H.; Zhang, F.; Havel, J. *Electrophoresis* **2003**, *24*, 3107.
- [9] Terrado, M.; Barcelo, D.; Tauler, R. *Anal. Chim. Acta* **2010**, *657*, 19.
- [10] Gui, M.; Rutan, S. C.; Agbodjan, A. *Anal. Chem.* **1995**, *67*, 3293.
- [11] Harshman, R. A. *UCLA Working Papers Phonetics*. **1970**, *16*, 1.
- [12] Wu, H. L.; Shibukawa, M.; Oguma, K. *J. Chemom.* **1998**, *12*, 1.
- [13] Shao, X. G.; Yu, Z. L.; Sun, L. *TrAC-Trend. Anal. Chem.* **2003**, *22*, 59.
- [14] Shao, X. G.; Liu, Z. C.; Cai, W. S. *TrAC-Trend. Anal. Chem.* **2009**, *28*, 1312.
- [15] Shao, X. G.; Wang, G. Q.; Wang, S. F.; Su, Q. D. *Anal. Chem.* **2004**, *76*, 5143.
- [16] Wang, G. Q.; Cai, W. S.; Shao, X. G. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 137.
- [17] Wang, W.; Cai, W. S.; Shao, X. G. *Sci. China, Ser. B: Chem.* **2007**, *50*, 530.
- [18] Liu, Z. C.; Cai, W. S.; Shao, X. G. *J. Chromatogr. A* **2008**, *1190*, 358.
- [19] Liu, Z. C.; Cai, W. S.; Shao, X. G. *J. Chromatogr. A* **2009**, *1216*, 1469.
- [20] Miao, L. N.; Cai, W. S.; Shao, X. G. *Talanta* **2011**, *83*, 1247.
- [21] Goicoechea, H. C.; Olivieri, A. C.; Tauler, R. *Analyst* **2010**, *135*, 636.
- [22] Foley, J. P. *Anal. Chem.* **1987**, *59*, 1984.
- [23] Zhang, Z. M.; Chen, S.; Liang, Y. Z. *Analyst* **2010**, *135*, 1138.

(Cheng, B.; Lu, Z.)