气相色谱-质谱法和模式识别技术用于雷公藤和昆明山海棠的分类鉴定

张 亮 张正行 盛龙生 安登魁

(药物分析学研究室)

摘 要 在样品色谱分离较完全的条件下,依据一定原则,将样品的成分质谱叠加,形成总质谱。用总质谱表征每一样本,并以总质谱的质量信道编码作指标。经 Shannon 信息理论特征选取,采用模式识别方法对雷公藤去皮根和根皮进行了分类,并对同属植物昆明山海棠作了定性预报。模式识别方法是 SIMCA 和 LDA 法,总预示率均为 100.0%。结果表明:雷公藤去皮根心和根皮差异明显,可分为两类;昆明山海棠和雷公藤亲缘关系密切,被划为同类。本法灵敏、快速,不仅为联用技术中信息的综合应用,而且为从化学本质上研究中药质量提供了依据。

关键词 气相色谱-质谱法;谱图叠加;Shannon 信息理论;模式识别;雷公藤;昆明山海棠

目前,中药分类、鉴定和质量分析的方法仍是以形态学特征鉴定为主,辅以少量现代理化分析(薄层色谱法居多)。高分辨气相色谱-质谱联用技术不仅具有色谱的高分离效能,而且兼备了质谱鉴定的高灵敏性和准确性,是研究物质分子的有效手段,也是现代分析中不可缺少的工具之一。本文采用气相色谱-质谱法和模式识别技术以中药雷公藤为例对药用植物质量分析进行了研究,首次提出谱图叠加技术,将每一样本经色谱分离后的各成分质谱,依据一定原则叠加,形成样本总质谱,从而在利用二维信息的基础上获得了模式分类器可按受的数值特征;其次将总质谱用于表征样本,并以总质谱的质量信道的编码作指标,经 Shannon 信息理论特征提取,从化学成分角度,对中药雷公藤去皮根心、根皮和昆明山海棠作了模式识别分类研究。

1 实验与结果

1.1 仪器与试药

惠普 5988A GC/MS 联用仪。分类用样本部分购自产地药材公司,部分系作者采集。见表 1,所有药材由江西庐山植物园赖书坤研究员鉴定。其它化学试剂为分析纯。

Tab 1. Samples used in classification

Species	No.	Collection places	Collection time
T. wilfordii	1	Jiling, Tonghua	1990.7
	2	Guangxi	1990.7
	3	Fujian	1990.7
	4	Zhejiang, Jinghua[a]	1991.7
	5	Jiangxi, Nanchang[a]	1991.4
	6	Guangzhou	1991.7
	7	Fujian, Shanghang[a]	1991.7
	8	Zhejiang, Xinchang	1991.7
	9	Sichuan[a]	1989.3
	10	Jiangxi, Jingdezhen[a]	1990.4
	11	Jiangxi, Lushan[a]	1990.2
	12	Jiangxi, Dexing[a]	1990.7
	13	Jiangxi, Jinggangshan[a]	1991.7
	14	Anhui	1991.7
	15	Fujing, Taining	1991.7
T. hypoglaucum	16	Yunnan, Midu	1991.7
	17	Kunming	1990.7
	18	Yunnan 1	1990.7
	19	Yunnan 2	1990.7
	20	Jiangxi, Lushan	1991.7
	21	Jiangxi, Sanqingshan	1990.3

[a]indicates root and bark

1.2 样品预处理

取样品粉末 10 g,准确称定,置索氏提取器中,用氯仿提取至完全,氯仿提取液减压浓

收稿日期 1993-02-18

缩至小体积，转移到 60 ml 分液漏斗中，加入 1% HCl 溶液 10 ml，振摇 5 min，分去水层。上述酸处理连续操作两次，然后，用水 10 ml 洗涤氯仿提取液，并用无水硫酸钠脱水，中性氧化铝脱色，最后，将氯仿提取液浓缩至 1 ml，从中吸取 50 μl，转移到 1 ml 反应瓶中，氮气流吹去氯仿，加无水乙醚溶解残渣后，通入重氮甲烷（氢氧化钾＋乙酰氨基脲），直至黄色不消失。用氮气流吹干，加吡啶 50 μl 和 TMSIM（三甲基硅基咪唑）50 μl 振荡混匀后，90℃保温 30 min，供样品分析。

1.3　GC/MS 条件

色谱柱 HP-1（25 m×0.2 mm，I.D.，膜厚 0.33 μm，熔融二氧化硅毛细管柱）。柱前压 15 Psi，载气为高纯氦（99.999%），初始柱温 150℃，停留 5 min，升温速率 5 ℃/min，终温 290℃，保持 20 min。分流进样（20∶1），进样口温度 360℃，离子源温度 280℃，接口温度 290℃，溶剂延迟 2.5 min，离子化方式 EI，离子能量 70 eV，质量扫描 45～500 a.m.u.。

1.4　数据采集

吸取样品溶液 0.1 μl，直接进样分析，得样品的总离子流色谱图。采用数据编辑系统，归一化法计算各峰峰面积。同时，打印各峰的质谱图。

1.5　谱图叠加

通过上述方法得到的质谱图往往不纯，来自漏气，残留样品（记忆效应），或色谱柱流失所产生的本底的异常峰包含其中，同样，进样口的热分解或催化分解也会产生类似情况。因此，在谱图叠加前，需要除去一些常见的杂质峰。(1)水(18)、氮气(28)、氧气(32)和二氧化碳(44)等；(2)超越分子离子峰及其合理同位素组以上的质量区的怪峰；(3)不合理的中性碎片；(4)质荷比为 207 的杂质峰（来自进样口垫片）。

去除杂质峰后的每一张质谱图，以分辨率为 1 a.m.u. 进行编码，阈值设定为基峰丰度的 1%，即每一质量信道，丰度超出阈值，用 1 表示，低于阈值用 0 表示。按下式对应叠加，形成样本的总质谱。

$$X_j^q = \sum_{i=1}^{n^q} S_i^q A_{ij}^q \qquad (1)$$

$i=1,2,3,\cdots,n^q$，$j=1,2,3,\cdots,m$，$q=1,2,3,\cdots,K$

式中 A_{ij}^q 为第 q 样本第 i 成分峰第 j 质量信道的编码；n^q 为第 q 样本参加叠加的成分峰总数；S_i^q 为第 q 样本第 i 成分峰的峰面积；X_j^q 为第 q 样本经叠加后的第 j 质量信道的编码值。

1.6　特征选择

经式(1)处理，每一样本构成一个 390 维特征向量。就微型计算机而言，外理高维样本比较困难。同样，就模式分类而言，也应该去掉对分类无效或反而易造成混淆的那些特征，尽量保留对分类特别有效的特征，即必须进行特征维数压缩。但是，维数（特征数）及其对谱图的分辨率是相互矛盾的两个方面。本研究采用 Shannon 信息理论进行特征选取，获满意结果。

Shannon 信息理论是研究从数量上定量描述信息的方法[1]。基本出发点是以被消除的不确定性来表示信息量，从而提供了评价方法与结果的定量标度。信息量的大小主要取决于事件发生的几率和事件间的相互关系，可用下式表示：

$$I_{(j)} = -\sum_{i=1}^{m} P_j(i)\log_2 P_j(i) \qquad (2)$$

式中 m 为第 j 质量信道的不连续的特征值；$P_j(i)$ 为第 j 质量信道特征值为 i 的样本对所有数据集样本发生的几率；$I_{(j)}$ 为所有样本第 j 质量信道的信息量之和。本研究中计算了两种情况下的 $I_{(j)}$ 值：一是 $S_i^q=1$ 时的质谱叠加；二是 S_i^q 等于峰面积时的质谱叠加。对所有 29 个样本（总质谱），每一质量信道的信息量分别在 0.150～2.44 bit 和 0.330～4.58 bit 之间。质量信道的信息量分布见表 2，由表可见，在 151 以下的低质量区含 50%以上的信息量。两种叠加信息量分布趋向一致。本

研究选取质谱叠加(S_i^q=峰面积)中 26 个 I 值在 4.0 以上的质量信道的编码作分类指标(见表 3)。26 个质量信道中有 21 个在 $S_i^q=1$ 时也最富含信息量。

Tab 2. Distribution of information content for mass channel

Mass	$\bar{S}_i$	Distribution, %	$\bar{P}_i$	Distribution, %
51-151	2.679	52.34	1.0037	50.66
152-252	1.328	25.95	0.6380	31.17
253-353	0.7770	15.18	0.2760	13.48
354-440	0.3340	6.52	0.0960	4.68

$\bar{S}_i$: information content mean when S_i^q equal to peak area; $\bar{P}_i$: information content mean when S_i^q equal to 1

Tab 3. Most information mass channel for data set

Mass channel	I(bit)	Mass channel	I(bit)	Mass channel	I(bit)
55	4.44	121	4.42	97	4.02
57	4.21	145	4.32	105	4.09
69	4.37	171	4.58	109	4.17
73	4.31	199	4.28	131	3.10
81	4.51	67	4.06	133	4.05
85	4.28	75	4.00	147	4.10
95	4.35	79	4.14	159	4.18
107	4.21	83	4.18	185	4.16
119	4.35	91	4.17		

1.7 模式识别方法

模式识别方法大致可分为两个过程:学习训练过程和分类识别过程。在学习过程中,将已知类别的模式样本在进行特征提取后,按设想的分类判决的数学模型进行分类,并将分类结果与已知类别的输入模式作比较,不断修改,制定出错误率最小的判别规则,最后,将未知模式特征拟合到该判别数学模型中完成分类识别过程。本文采用了 SIMCA 法。SIMCA 法属于模型化方法[2],是运用已知类别的训练集样本所构造的主成分数学模型去进行样本分类,其公式如下:

$$Y_{ik}^q=\alpha_i^q+\sum_{a=1}^{A_q}\beta_{ia}^q\theta_{ik}^a+e_{ik}^q \qquad (3)$$

式中 Y_{ik} 为数据矩阵 Y 的矩阵元;e_{ik} 为残差;α,β 为使 e_{ik} 达到极小的参数;$q=1,2,3,\cdots\theta$(θ 为已知样本的类别数);$i=1,2,3,\cdots,M$(M 为变量数或特征数);$a=1,2,\cdots,A_q$(A_q 为主成分数);$K=1,2,\cdots,n_q$(n_q 为类 q 中样本数)。

除 SIMCA 法外还采用了 LDA 法,所有程序出自 PARVUS 软件包,并在 386 微机上运行。

1.8 数据结构显示

常用的数据结构显示有线性映射和非线性映射法。由于本研究中谱图数据多为非线性的,故选用非线性映射技术,将多维空间中的样本点映射在便于观察判别的二维平面上。由图 1 可见雷公藤去皮根心和根皮分类明显,昆明山海棠 1 和 4 落在根皮区,其余在根心区。

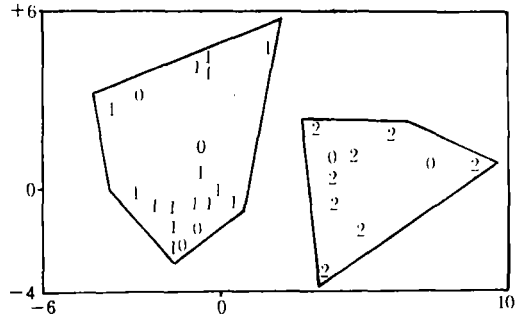

Fig 1. Non-linear mapping plot of 29 objects
1. indicates *T. wilfordu* root; 2. indicates *T. wilfordu* bark; 0. indicates *T. hypoglaucum* root

1.9 分类结果

本文试验了雷公藤的两个不同部位:去皮根和根皮。每次从中任取一样本作试验集,其余样本作为训练集,重复运算,直至每个样本都执行一次,最后计算正确识别率,待预报样本为昆明山海棠去皮根。分类结果见表 4。由表可见,雷公藤去皮根心和根皮差异明显,可分为两类,昆明山海棠去皮根与雷公藤去

Tab 4. Classification of 29 samples

	SIMCA		LDA	
	Root ($n=15$)	Bark ($n=8$)	Root ($n=15$)	Bark ($n=8$)
Correct classification	100%	100%	100%	100%
Misidentified into class *T. hypoglaucum*	0	0	0	0
Yunan, Midu	×			×
Kunming	×		×	
Yunan 1	×		×	
Yunan 2	×			×
Jinagxi, Lushan	×		×	
Jiangxi, Shancenshan	×		×	

皮根被划为同类，表明两者亲缘关系密切，分类结果与数据结构显示比较一致。

2 结　论

2.1　雷公藤植物化学成分非常复杂，就目前已知的化学成分而言，几乎大部分成分含羟基或羧基，是一些难挥发性化合物，采用先甲酰化，后硅烷化的方法，大部分成分适于气相色谱分析。衍生化反应温度，实验结果表明以90℃反应30 min，色谱响应信号最强。经衍生化处理后，仍有一些分子结构较大的雷公藤生物碱难以气化，当直接进行GC/MS分析时，峰拖尾严重。本研究在样品衍生化前，利用生物碱的弱碱性，采用稀盐酸(1%)，反复提取，除去了其中的生物碱。

2.2　质谱图处理方法有多种：对峰归一化、对谱强度(丰度)进行对数转换、或进行编码。不同的研究目的应当选择不同的处理方法。本研究采用二值编码，虽然失去了一些强度信息，但保留了化合物的最基本的结构信息。实验证明：质谱的二值码转换是实现类分离的有力手段。

2.3　Shannon信息量是谱图几率分布信息较简便且有效的表达方式，并起到模式识别研究中数据降维作用。本研究中390个不同质量信道被有效地压缩到26个。就SIMCA模型而言，选择高信息量信道，能获得好的分类结果，但并不是所有研究都需要高信息量，有时也会存在例外。

2.4　雷公藤去皮根和根皮虽为同种植物的不同部位，但从分类结果来看，根皮和去皮根明显不同。临床医疗规定入药部位为去皮根，鉴于目前市售雷公藤多为全根，为确保药品的安全有效，应严格控制根皮的混入。昆明山海棠与雷公藤为同属不同种植物，采用本文方法两者的去皮根被划为同类，表明两者成分较为接近(此结论与两者植物化学成分研究的结果比较一致)，能否作为雷公藤的代用品使用可以进一步研究。

2.5　针对丰富的色谱、质谱信息，本研究首次提出谱图叠加技术，形成了现有模式分类器可接受的数值特征，并获满意的分类结果，从而为联用技术中多维信息的综合处理提供了新的思路。同时，本文立足化学成分的质谱，从物质分子水平揭示其内在规律，是现代分析技术研究中药质量的一种新的尝试。

参考文献

1　Donald RS. Determination of chemical classes from mass spectra of toxic organic compounds by SIMCA pattern recognition and information theory. *Anal Chem*, 1986;**58**:881

2　Wold S. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In: Kowalski BR ed. Chemometrics: *Theory and Application*, Washington. 1977: 243-82

Determination of Chemical Classes from Mass Spectra of Medicinal Plants by GC-MS and Pattern Recognition

Zhang Liang, Zhang Zhengxing, Sheng Longsheng, An Dengkui

Department of Pharmaceutical Analysis

Under the complete chromatographic seperation of samples, the sum mass spectra of a set of 29 extraction of medicinal plants were obtained by overlapping of mass spectrum of components (using 1% of the reference peak as the threshold level), examined for information concerning chemical classes. The Shannon information content for each mass channel was calculated for the sum spectra, the 26 mass channels with the highest information were retained as a compressed basic set for SIMCA and LDA pattern recognition. The inherent class structure of the date showed two major classes by NLM; *T. wilfordii* bark and root, *T. hypoglaucum* roots were classified into *T. wilfordii* root class. Classification accuracy was 100% for the above two classes.

Key words　GC-MS; Mass spectrum overlapping; *T. wilfordii*; Pattern recognition