基于多模式识别结合指纹图谱的三叶青产地鉴别比较研究

李士敏[1]，李 强[2]，孙崇鲁[1]，彭 昕[1*]

1. 浙江医药高等专科学校，浙江 宁波 315100
2. 艾美卫信生物药业（浙江）有限公司，浙江 宁波 315600

摘 要：目的 比较不同模式识别方法结合化学指纹图谱对不同产地三叶青 *Tetrastigma hemsleyanum* 的鉴别效果，提出三叶青产地鉴别的新方法。**方法** 收集浙江、云南和贵州 3 个产区的 72 批三叶青样品，采集 HPLC 指纹图谱，标注 18 个共有峰，比较主成分分析、正交偏最小二乘法-判别分析和随机森林算法在处理不同产地样品复杂数据中的差异效果。**结果** 72 批三叶青样品经主成分分析只能分为 2 类，正交偏最小二乘法-判别分析的结果优于主成分分析，随机森林法能将 3 个产区的样品完全分开。利用随机森林算法结合指纹图谱能有效地对不同产区的三叶青进行鉴别和区分。**结论** 本研究可作为三叶青产区和质量控制的有效方法，为多指标的复杂指纹图谱鉴别不同产地药材提供有效的参考。

关键词：三叶青；高效液相色谱法；指纹图谱；随机森林；化学模式识别；绿原酸；虎杖苷；山柰酚-3-*O*-芸香糖苷；紫云英苷；白藜芦醇；山柰酚

中图分类号：R282.6 **文献标志码：**A **文章编号：**0253 - 2670(2020)01 - 0197 - 07

DOI: 10.7501/j.issn.0253-2670.2020.01.026

Comparative study on multiple chemical pattern recognition combined with fingerprint of *Tetrastigma hemsleyanum* from different habitats

LI Shi-min[1], LI Qiang[2], SUN Chong-lu[1], PENG Xin[1]

1. Zhejiang Pharmaceutical College, Ningbo 315100, China
2. AiMei Vacin Biopharmaecutical (Zhejiang) Co., Ltd., Ningbo 315600, China

Abstract: Objective To compare the identification effects of different pattern recognition methods combined with chemical fingerprints on *Tetrastigma hemsleyanum* from different habitats, and propose a new identification method for the habitats of *T. hemsleyanum*. **Methods** A total of 72 batches of *T. hemsleyanum* samples were collected from Zhejiang, Yunnan and Guizhou. HPLC fingerprints were collected, 18 common peaks were marked, and the difference of principal component analysis (PCA), orthogonal partial least squares-discriminant analysis (OPLS-DA) and random forest (RF) in processing complex data of samples from different habitats was compared. **Results** A total of 72 batches of *T. hemsleyanum* samples can only be divided into two categories by PCA, the results of OPLS were better than PCA, and the RF can completely separate the samples from three habitats. The RF combined with fingerprint can effectively identify and distinguish *T. hemsleyanum* from different habitats. **Conclusion** This study can be used as an effective method for the quality control of *T. hemsleyanum* from different habitats and provide an effective reference for multi-index complex fingerprint identification from different habitats.

Key words: *Tetrastigma hemsleyanum* Diels et Gilg; HPLC; fingerprint; random forest; chemical pattern recognition; chlorogenic acid; polydatin; kaempferol-3-*O*-rutinoside; astragalin; resveratrol; kaempferol

三叶青为葡萄科崖爬藤属三叶崖爬藤 *Tetrastigma hemsleyanum* Diels et Gilg 的块根，是我国特有草本植物，分布在浙江、福建、云南等长江以南的大部分地区，以块根入药，传统医学认为三叶青具有清热解毒、消肿散结、活血止痛之功效[1]。民间广泛用于抗肿瘤、调节免疫力以及治疗小儿高热等[2]。由于其临床疗效显著，市场需求不断增大，野生资源已经无法满足药材需求，近年来各地均开展人工栽培驯化。

中药材产区与其质量关系密切，不同产区的三

收稿日期：2019-04-09

基金项目：浙江省公益技术应用研究项目（2017C32075）；浙江省基础公益研究计划项目（LGN18B020001）；宁波市领军和拔尖人才培养工程择优资助科研项目（NBLJ20180101）

作者简介：李士敏（1982—），男，讲师，硕士，研究方向为药物分析与质量控制。Tel: (0574)88222693 E-mail: shiminlee@ymail.com

***通信作者** 彭 昕，女，教授，硕士生导师，研究方向为天然产物分析及应用。Tel: (0574)88223132

叶青外观性状、药理活性及价格均差异悬殊。课题组前期调研发现，我国西南地区的三叶青生长周期短，1～2 年即可采收，产量较大，价格低；而东南沿海等地的三叶青生长周期长、3～5 年才可采收，产量低，价格较高。《备急千金要方》记载：“医者用药必依土地，所以治十得九”。研究发现，不同产区三叶青有效成分含量及药理活性均差异悬殊，总黄酮含量最高差 7 倍[3]；个别产区药材部分黄酮成分几乎检测不到[4]；浙江产区三叶青提取物在解热[5]、抑制肝癌细胞增殖[6]等方面都有明显的优势。然而目前对三叶青的产区鉴别多依赖于外观性状评定[7]，课题组前期已报道应用红外光谱技术对不同产地的三叶青进行鉴别研究[8]，但以上方法均缺乏客观的量化依据，不足以全面而准确地反映不同产区三叶青品质差异的主要变量。

区分药材产地是中药道地性及品质形成研究的重要基础之一。化学指纹图谱是一种从整体上研究复杂物质体系的技术，已成为国际公认的控制天然药材质量的最有效手段[9]。近年来，多元统计分析应用的迅速发展为中草药的指纹图谱研究提供了前所未有的应用前景，主成分分析（principal component analysis，PCA）与偏最小二乘法-判别分析（partial least squares discriminant analysis，PLS-DA）的化学模式识别方法已被广泛应用于色谱法获得的复杂中药化学指纹图谱进行特征提取。PCA 可以直观地区分出多组多元变量间的差异程度，而 PLS-DA 可以对比 2 组多元变量之间的差异并计算出对差异贡献较大的因素，寻找主要的差异成分，为药材质量的全面控制及其含量测定指标的选择提供科学依据。然而，这些算法都存在一定的局限性，如 PCA 对离群点较敏感，PLS-DA 容易产生过拟合现象[10]。随着指纹图谱数据复杂性不断增加，一些更先进的算法应用于多维复杂数据分析，随机森林算法（random forest，RF）作为一种操作方便、结果可靠的基于分类回归树集成的机器学习方法，具有较高的分类准确率，相对于其他的分类算法，RF 能较好地克服噪音，分析结果不易过拟合以及可以泛化误差，是一种利用多个分类树对数据进行判别与分类的方法，在许多领域取得了广泛的应用[11]。

前期课题组对三叶青的地上部分和地下部分进行了化学成分分析与鉴定，并对不同产地三叶青中 27 种矿物元素的组成及含量进行了综合评价[12-14]。本研究以云南（YN）、贵州（GZ）和浙江（ZJ）3 个产区的 72 批三叶青药材为材料，分析比较了 PCA、OPLS-DA 和 RF 等化学模式识别方法的产地鉴别结果，为三叶青药材产地溯源系统的建立打下基础，为三叶青药材质量控制标准制定提供参考，也为今后其他道地药材产地鉴别和质量评价提供新思路。

1 仪器与试剂

Agilent 1260 高效液相色谱仪（美国 Agilent 公司）；Agilent SB-C_{18} 色谱柱（250 mm×4.6 mm，5 μm，美国 Agilent 公司）；KQ-3200DE 超声波清洗器（昆山市超声仪器有限公司）；Cascada-BIO 超纯水仪（美国 Cascada 公司）；分析天平 XSE105DU（瑞士梅特勒-托利多公司）；QB-3B 手提式中药粉碎机（温岭市大德中药机械有限公司）。

对照品绿原酸（批号 18100210，质量分数≥99.39%）、虎杖苷（批号 18100216，质量分数≥99.78%）、山柰酚-3-*O*-芸香糖苷（批号 18100411，质量分数≥98.0%）、紫云英苷（批号 18022087，质量分数≥98.0%）、白藜芦醇（批号 18021134，质量分数≥99.95%）、山柰酚（批号 18100261，质量分数≥98.0%）购自于上海源叶生物科技有限公司；甲醇、乙腈为色谱纯（美国 Tedia 公司）；所检测的样品（表 1）经浙江医药高等专科学校彭昕教授鉴定为葡萄科崖爬藤属三叶崖爬藤 *Tetrastlgma hemsleyanum* Diels et Gilg 的块根。

2 方法

2.1 对照品溶液的制备

分别精密称取绿原酸、虎杖苷、山柰酚-3-*O*-芸香糖苷、紫云英苷、白藜芦醇和山柰酚对照品适量，用甲醇溶解并定容。置于 4 ℃冰箱中保存，经 0.22 μm 滤膜滤过，备用。

2.2 供试品溶液的制备

精密称取三叶青样品粉末 2.0 g，加 80%甲醇 25 mL，超声（300 W，40 ℃）提取 45 min，滤过，滤渣用 80%甲醇 25 mL 重复提取 1 次，合并提取液，蒸干，转移至 10 mL 量瓶中，用 80%甲醇稀释至刻度，置 4 ℃，临用前经 0.22 μm 滤膜滤过。

2.3 色谱条件

色谱柱为 Agilent SB-C_{18} 柱（250 mm×4.6 mm，5 μm）；流动相为乙腈-0.2%磷酸溶液，梯度洗脱，洗脱条件为 0 min，10%乙腈；0～30 min，10%～30%乙腈；30～40 min，30%～95%乙腈；40～45 min，

表 1 不同产地三叶青样品信息

Table 1 Information of *T. Hemsleyanum* from different habitats

样品编号	产地	批号	样品编号	产地	批号
S1	浙江镇海	180603	S37	云南宣威	181102
S2	浙江镇海	180705	S38	云南玉溪	180702
S3	浙江镇海	180805	S39	云南玉溪	180705
S4	浙江镇海	180907	S40	贵州兴义	180303
S5	浙江武义	180910	S41	贵州兴义	180503
S6	云南曲靖	180721	S42	贵州兴义	180505
S7	云南曲靖	180725	S43	贵州兴义	180910
S8	云南曲靖	181005	S44	浙江武义	180303
S9	浙江宁波	180505	S45	浙江武义	180503
S10	浙江宁波	180527	S46	浙江武义	180505
S11	浙江宁波	180713	S47	浙江武义	190110
S12	浙江武义	180404	S48	浙江台州	190121
S13	浙江武义	180501	S49	浙江台州	190123
S14	浙江武义	180503	S50	浙江台州	190124
S15	云南昭通	180302	S51	浙江台州	190125
S16	云南昭通	180303	S52	浙江台州	190301
S17	浙江武义	180706	S53	浙江台州	190302
S18	浙江武义	180709	S54	云南宣威	170910
S19	浙江武义	180711	S55	云南宣威	170915
S20	贵州安顺	180303	S56	云南宣威	170917
S21	浙江镇海	181009	S57	云南玉溪	170910
S22	浙江镇海	181104	S58	云南玉溪	171004
S23	浙江镇海	181129	S59	云南玉溪	171005
S24	云南玉溪	180303	S60	云南曲靖	180307
S25	云南玉溪	180503	S61	云南曲靖	170917
S26	云南玉溪	180505	S62	云南曲靖	170910
S27	云南昭通	180503	S63	云南昭通	171014
S28	云南昭通	180505	S64	云南昭通	171007
S29	云南昭通	180706	S65	云南昭通	171121
S30	贵州安顺	180508	S66	浙江镇海	181205
S31	贵州安顺	180509	S67	浙江镇海	190107
S32	贵州安顺	180709	S68	浙江镇海	190205
S33	贵州安顺	180712	S69	浙江镇海	190307
S34	浙江宁波	180504	S70	云南玉溪	181220
S35	浙江宁波	180507	S71	云南玉溪	181221
S36	云南宣威	180303	S72	云南玉溪	181222

95%乙腈；45～60 min，95%～10%乙腈；体积流量 0.8 mL/min；检测波长 320 nm；柱温 25 ℃；进样量 10 μL。

2.4 方法学考察

2.4.1 重复性试验 按“2.2”项下方法制备样品 6 份，按“2.3”项下色谱条件检测，计算各共有峰的相对保留时间和相对峰面积，结果表明各共有峰的相对保留时间的 RSD＜1.0%，各共有峰的相对峰面积的 RSD＜1.2%。表明本方法重复性好。

2.4.2 稳定性试验 按“2.2”项下方法制备供试品溶液。取放置 0、2、4、6、8、12 h 的样品按“2.3”项下色谱条件检测，计算各共有峰的相对保留时间和相对峰面积，结果各共有峰的相对保留时间的 RSD＜1.2%，各共有峰的相对峰面积的 RSD＜

2.1%。表明供试品在 12 h 内稳定。

2.4.3　精密度试验　取“2.1”项下混合对照品溶液，按“2.3”项下色谱条件检测，重复进样 6 次，记录色谱图。计算指纹图谱中各共有峰的相对保留时间和相对峰面积，结果表明各共有峰相对保留时间的 RSD＜0.9%，各共有峰相对峰面积的 RSD＜1.2%，提示仪器精密度良好。

2.5　指纹图谱构成与数据处理

将不同产区三叶青样品按“2.3”项色谱条件进样，采集数据生成色谱图。将 72 批样品色谱图导入“中药色谱指纹图谱相似度评价系统”（2012A 版），采用平均数法谱峰多点矫正生成三叶青叠加图谱。比较各样品的色谱图，确定 18 个共有主要特征峰。将 72 批三叶青样品指纹图谱 18 个共有峰峰面积输入 SIMCA（14.0）软件，建立 PCA 和 OPLS-DA 模型并进行分析。

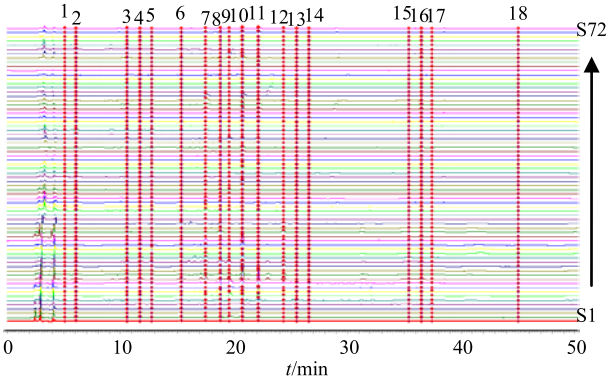
PCA 是一种无监督的多元统计分析算法，其通过数学降维的原理，从原数据中提取几个综合变量来代替原来众多的变量，综合变量的提取尽可能地代表原来的变量信息，且彼此之间互不相关。偏最小二乘法（PLS）是一种可以同时实现多元线性回归、主成分分析的数据分析方法，PLS 在建模时可能会在自变量 X 中引入与因变量 Y 无关的自变量信息或噪音。OPLS 是通过正交信号校正法（OCS）将原始数据 X 中与因变量 Y 不相关的信息去除，减少对 Y 的预测误差，从而改善 PLS 的算法模型。OPLS 能使组别间的差异性最大化，提高模型的分类精确度。RF 是一类基于分类回归树集成算法，其在进行数据聚类分析的同时能够得到各变量对于聚类的贡献度。RF 算法主要参数有 2 个：树的规模 nTree 和属性特征的子集大小 mtry。nTree 表示 RF 模型中决策树的数目，mtry 表示 RF 算法节点分裂时选择的分裂属性的个数，mtry 通常远小于总属性的数量，且建树过程中保持不变。

本研究从全部三叶青样品中随机选出 80%作为训练集，20%的产品作为独立测试集。RF 模型的建立使用 MATLAB（R2019a，The mathworks，USA）数学建模软件。

3　结果与分析

3.1　HPLC 指纹图谱的建立

将 72 批样品色谱图导入“中药色谱指纹图谱相似度评价系统”，采用平均数法谱峰多点矫正生成三叶青叠加图谱，确定了 18 个共有峰。通过对照品比对，在指纹图谱共有峰中共指认出 6 个成分，分别为绿原酸（峰 3）、虎杖苷（峰 9）、山柰酚-3-*O*-芸香糖苷（峰 11）、紫云英苷（峰 12）、白藜芦醇（峰 16）、山柰酚（峰 17）。72 批三叶青药材色谱峰图见图 1。



3-绿原酸　9-虎杖苷　11-山柰酚-3-*O*-芸香糖苷　12-紫云英苷　16-白藜芦醇　17-山柰酚

3-chlorogenic acid　9-polydatin　11-kaempferol-3-*O*-rutinoside　12-astragalin　16-resveratrol　17-kaempferol

图 1　72 批样品的 HPLC 指纹图谱

Fig. 1　HPLC fingerprint for 72 batches of samples

3.2　PCA

以 18 个共有峰峰面积为变量，采用 SIMCA（14.0）软件对 72 批三叶青样品进行 PCA，结果提取到 2 个具有最大特征值的主成分（PC），总贡献率为 61.9%，其中 PC1 贡献率为 52.6%，贡献率最大；PC2 贡献率为 9.3%。由前 2 个主成分建立坐标系，得到 72 批三叶青的 PCA 得分图（图 2），由图可见，大部分样品在 95%可信区间能够清晰分为 2 类，说明提取到的前 2 个主成分已能反映出 2 大不同产区三叶青的主要特征，提示浙江和云贵产区的三叶青在化学成分含量上存在一定的差异。

3.3　OPLS-DA

依据 PCA 结果，对不同分布区域的 2 组样品 Gr1（ZJ）和 Gr2（YN、GZ）进行 OPLS-DA 分析。该 OPLS-DA 模型，R^2Y (cum)＝0.769，Q^2 (cum)＝0.673，均大于 0.5，说明模型稳定可靠，可用于不同产区样品的区分。从图 3 可以观察到，数据的分类算法中 OPLS-DA 的效果较好，可使 3 类样本点完全被分开，相互之间没有样本出现交叉的情况，少部分样品位于 95%可信区间之外。浙江产区，云南产区和贵州产区样本分别聚集在各自所在区域，贵州产区与云南产区样品分布靠近。

OPLS-DA 载荷图见图 4，结合变量重要性投影

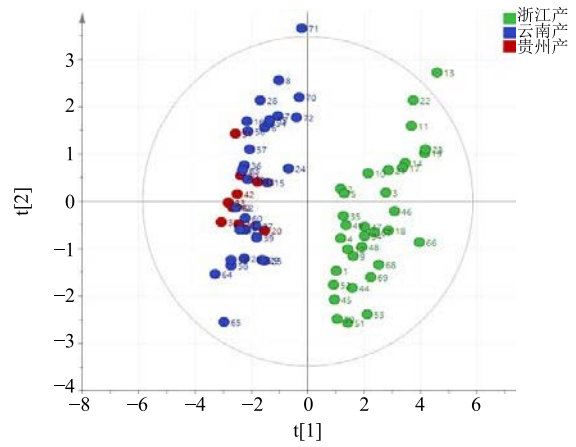


图 2 PCA 得分

Fig. 2 Score scatter plot for PCA

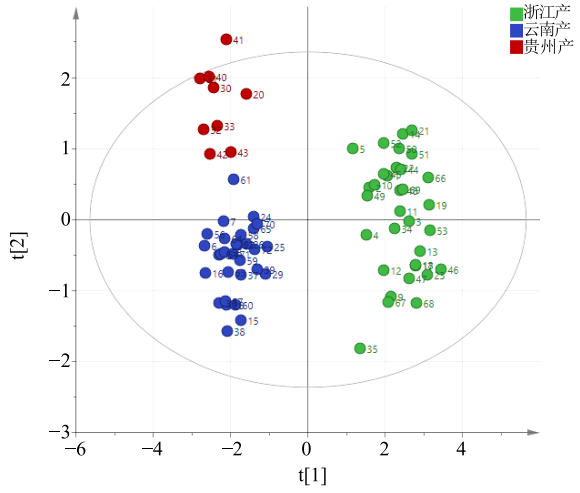


图 3 不同省份 72 批三叶青样品的 OPLS-DA 得分图

Fig. 3 OPLS-DA score plot of 72 batches of *T. hemsleyanum*

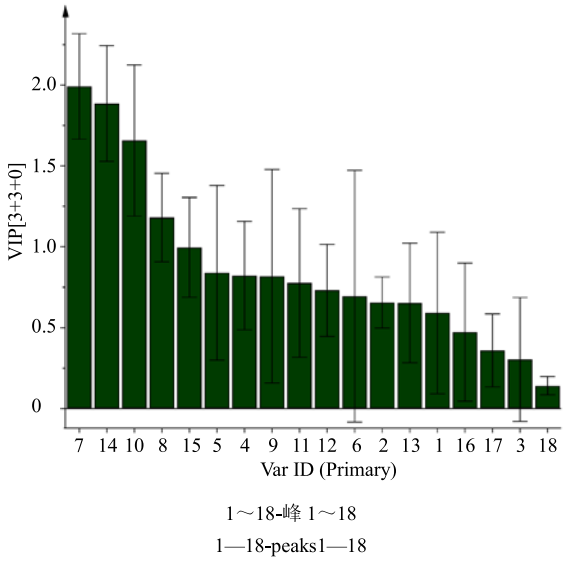


图 4 三叶青 18 个共有峰的 VIP 值

Fig. 4 VIP of 18 common chromatographic peaks of *T. hemsleyanum*

值（variable importance in the projection，VIP），拥有较大 VIP 值的变量（至少大于 1）对分类的贡献越大。以 VIP＞1.0 作为标准筛选出对模型上 2 组间分类贡献较大的 4 个变量，依次为峰 7、14、10、8，这些成分是引起三叶青不同产区间成分差异的主要标志性成分，其余峰 VIP 值小于 1，对样品的区分影响较小。

3.4 RF

利用 RF 对训练集进行分类预测，其 13 折交互验证的预测准备率为 100.0%，说明随机森林具有较强的分类能力，能够有效区分不同产区的三叶青样品。基于构建的训练模型，对剩下的 20%的独立测试集进行分类预测，其预测准确率为 100.0%。各产区分类预测见表 2。

表 2 各产区分类预测

Table 2 Predicted results of each chemometric methods

分类	样品类型	样品个数	准确预测个数
训练集	浙江	26	26
	贵州	7	7
	云南	24	24
独立测试集	浙江	7	7
	贵州	2	2
	云南	6	6

样品分布结果（图 5-A）可见，3 个产区三叶青样品均得到有效的区分。浙江产区样品处于其他 2 类的另一个方向，而云南和贵州产区样品距离较近，说明云南和贵州产区样品尽管较为相似，但依然存在区别，在 RF 中均得到有效区分。18 个共有峰经 RF 分析热图（图 5-B）可得，峰 14、15、2 对不同产地区分重要性较大，可作为鉴别不同产三叶青的指标性成分。尤其是峰 14 代表的成分对不同产地样本的分类有最突出的贡献，说明在区分浙产三叶青和云贵三叶青时，峰 14 代表的化学成分是非常重要的指标。

4 讨论

4.1 三叶青产地鉴别方法比较

化学成分是中药产生生物效应的物质基础。作为三叶青中重要的抗癌物质，黄酮类是最具有应用价值的有效成分，能抑制癌细胞的增殖，促进癌细胞的凋亡；多糖类、氨基酸类是生物组成的重要物质，具有免疫调节、抗氧化、抗病毒等诸多作用[15]。《本草纲目》记载“性从地变，质与物迁，未尝同

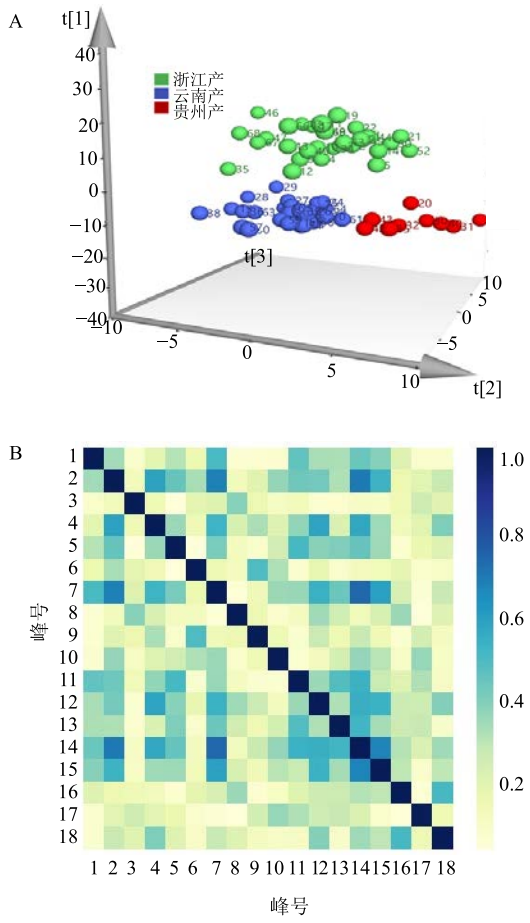


图 5　72 批三叶青样品的 RF 得分图 (A) 和热图 (B)

Fig. 5　Scatter plot (A) and heat map (B) of 72 batches *T. hemsleyanum*

也。”不同产地的气候条件和生态环境，会通过影响药用植物的生长发育、能量代谢、物质合成及气体交换等生理作用，继而影响其有效物质（次生代谢产物）的变化，从而导致药材质量及临床疗效产生差异[16]。目前报道的关于三叶青产区鉴别方法包括性状鉴别、显微鉴别、化学反应鉴别、紫外鉴别和红外光谱鉴别等，通常人们通过药用植物的生长环境、药材的性状判断出中药质量的优劣，但对于炮制、加工后的三叶青饮片或粉末，很难从外观、性状等方面评价质量优劣，一些销售商为了赚取更多的利益，以次充好，造成了药材市场的混乱。目前常见的产地鉴别方法中，性状鉴别不适用于炮制、加工后的样品；显微鉴别前处理复杂，准确度低；理化鉴别专属性差；红外鉴别法噪音检测次数增加而增大，重复性差；紫外鉴别法可能存在漏检或过度检出[17]。HPLC 可定性定量分析三叶青样品，但其缺乏整体性，且不能反映样品间关系。构建具有良好的特异性和灵敏度的评价方法显得尤为重要。本研究表明，所采用的 HPLC 指纹图谱结合化学识别模式可有效阐明中药样品间的细微差别，为不同产区中药材质量评价提供参考。

4.2　不同化学识别模式方法的应用比较

药用植物中所包含的各种成分信息无法直观地通过个体之间的差异展现，借助指纹图谱结合化学计量学方法进行数据处理和解析并提取有用信息，能够解释测量值与体系状态之间的联系。目前，PCA、PLS、OPLS 和 RF 等方法是最常用于药用植物产地鉴别的化学计量学方法。

PCA 是一个无监督的学习方法，其依靠样品间的相似性进行分析。依据 PCA 结果，浙江产三叶青和云贵产区的三叶青清晰的分为 2 类，这表明 2 个产地的三叶青化学成分含量上存在显著差异。PCA 不能将来自云南和贵州 2 个产地的三叶青区分开来。可能的原因之一是这 2 个产区的三叶青化学成分含量确实很相似，因此造成了这 2 类数据的差异性很小，使用 PCA 不容易区分开。另一个原因是由于 PCA 算法是一种线性分类器，其本质是通过线性变换把数据重新组合成原始数据的线性组合，但复杂的化学指纹图谱数据包含的非线性因素越来越多，简单地把原始数据看作一个线性模型并作线性变换可能会丢失非常多数据的原始信息。

PLS-DA 目前已应用于中药材的来源分析及差异标记物筛选等方面，在夏枯草[18]等多种中药材上质量评价方面取得了较好的表现。PLS-DA 可能会在自变量 X 中引入与因变量 Y 无关的自变量信息而产生噪音。OPLS 能将原始数据 X 中与因变量 Y 不相关的信息去除，提高了 PLS 模型的分类精确度并使类与类之间的差异性得到最大化。OPLS 方法作为一种广泛使用的分类学习器，能够有效提高预测准确率，和 PLS 的分类结果相比，OPLS 的结果中类内聚集性提高。结果显示浙江产区，云南产区和贵州产区样本分别聚集在各自所在区域，贵州产区与云南产区样品分布靠近，这与两地地理距离相近，生长环境相对接近有关。这 2 个产区与浙江产区样品分布相距较远，说明浙江产区样品的生长环境与云贵产区差异较大。分类算法中 OPLS 可使 3 个产区样本完全分开，相互之间没有出现交叉的情况，但少部分样品分布超出了 95%可信区间。

RF 算法结合了 Bagging 方法和随机子空间算法的思想，通过自助采样，获取数据子集来构建多棵相互独立的决策树，然后组合这些决策树得到 RF

模型。Beriman[19]通过实验证明了 RF 算法能够比较好地克服过拟合问题，且在分类和回归问题上都能取得较好的效果。RF 算法的分析结果说明相比于 PCA 和 OPLS 算法，RF 在复杂数据或者指纹相似样品的处理和分类上具有显著的优势，具有较好的效果。经 RF 分析发现浙江产区和云贵产区三叶青色谱图中峰 14、峰 15 和峰 2 代表的化合物是引起差异的主要成分。浙江产区三叶青指纹图中峰 14、峰 15 和峰 2 代表的化合物含量普遍高于云贵产区。下一步课题组将对不同产区的三叶青药材进行抗肿瘤等活性研究，评价不同产区三叶青活性差异。在此基础上，完成峰 14、峰 15 和峰 2 代表的化合物的分离和结构鉴定，并结合生物活性实验确认其是否能作为不同产地三叶青鉴别和质量评价的标志物。

综上所述，中药指纹图谱结合化学模式识别研究手段应用最普遍的是聚类分析、PCA、以及判别分析等，大部分研究中均能用上述算法进行较好的数据处理。随着分析技术的进步和中药材质量控制的需要，复杂指纹图谱出现的越来越多，常规的化学模式识别算法不能满足对复杂数据的分析需求。本研究结果表明相比于 PCA 和 OPLS-DA 算法，RF 在复杂数据处理和分析上具有明显的优势，本研究对中药材质量分类和产地溯源方面具有重要的意义。

参考文献

[1] 国家中医药管理局《中华本草》编委会. 中华本草 (第 4 册) [M]. 上海: 上海科学技术出版社, 1999.

[2] Sun Y, Li H Y, Hu J N, *et al*. Qualitative and quantitative analysis of phenolics in *Tetrastigma hemsleyanum* and their antioxidant and antiproliferative activities [J]. *J Agric Food Chem*, 2013, 61(44): 10507-10512.

[3] 范世明, 林 婧, 许 文, 等. 不同产地三叶青中总黄酮含量的比较 [J]. 福建中医药大学学报, 2013, 23(3): 44-45.

[4] 许 文, 傅志勤, 林 婧, 等. UPLC-MS/MS 法同时测定三叶青中 10 种黄酮类成分 [J]. 药学学报, 2014, 49(12): 1711-1717.

[5] 杨 雄, 王翰华. 不同产地三叶青提取物解热作用及对大白鼠下丘脑 5-羟色胺、去甲肾上腺素、多巴胺含量的影响 [J]. 长春中医药大学学报, 2014, 30(3): 393-395.

[6] 林 婧, 黄泽豪, 许 文, 等. 不同产地三叶青总黄酮含量及对肝癌细胞增殖的抑制率比较 [J]. 福建中医药大学学报, 2014, 24(5): 40-41.

[7] 黄 真, 胡瑛瑛, 王庆秋, 等. 浙江三叶青与广西三叶青的生药学鉴别 [J]. 浙江中医药大学学报, 2007, 31(6): 759-760.

[8] 赖添悦, 蔡逢煌, 彭 昕, 等. 核密度估计算法结合近红外光谱技术鉴别三叶青产地 [J]. 光谱学与光谱分析, 2018, 38(3): 794-799.

[9] 杨冉冉, 姬 蕾, 李二文, 等. 鸡血藤的 HPLC 指纹图谱及模式识别研究 [J]. 中草药, 2017, 48(21): 4530-4536.

[10] 柯朝甫, 武晓岩, 侯 艳, 等. 偏最小二乘判别分析交叉验证在代谢组学数据分析中的应用 [J]. 中国卫生统计, 2014, 31(1): 85-87.

[11] 翟新房, 赵焕虎, 杨 册, 等. 基于液质联用-模式识别方法分析不同产地的绞股蓝皂苷 [J]. 中草药, 2019, 50(13): 3193-3199.

[12] 张煜炯, 彭 昕, 吉庆勇, 等. 聚类分析和主成分分析法研究三叶青氯仿部位 HPLC 指纹图谱 [J]. 中成药, 2016, 38(3): 607-612.

[13] 孙崇鲁, 吴 浩, 楼天灵, 等. UPLC-Q-TOF-MS 法分析三叶青地上部分化学成分 [J]. 中成药, 2018, 40(6): 1424-1429.

[14] 吴 浩, 常欣桑, 旭 峰 等 不同产地三叶青中 27 种矿质元素的综合评价 [J]. 中成药, 2018, 40(11): 2475-2480.

[15] Ding L, Zhang L X, Qiu Y, *et al*. Chemical constituentsin chloroform extraction of *Tetrastigma hemsleyanum* Diels et. Gilg and their antitumor activities [J]. *Chin Pharm J*, 2015, 50(21): 1857-1860.

[16] 唐仕欢, 杨洪军, 黄璐琦. 论自然环境因子变化对中药药性形成的影响 [J]. 中国中药杂志, 2010, 35(1): 126-128.

[17] 裴艺菲, 左智天, 赵艳丽, 等. FTIR、ATR-FTIR 和 UV 多光谱鉴别不同产地重楼 [J]. 分析测试学报, 2019, 38(1): 14-21.

[18] 皮胜玲, 胡玉珍, 彭 曦, 等. 野生与栽培夏枯草 HPLC 指纹图谱研究及模式识别分析 [J]. 中国药学杂志, 2017, 52(5): 367-371.

[19] Breiman L. Random forests [J]. *Mach Learn*, 2001, 45(1): 5-32.