

基于高光谱分析的淫羊藿药用成分快速检测研究

姜庆虎¹, 刘峰¹, 于东悦^{2,3}, 罗惠^{2,3}, 梁琼^{3*}, 张燕君^{3*}

1. 中国科学院武汉植物园, 中国科学院水生植物与流域生态重点实验室, 湖北 武汉 430074

2. 中国科学院大学, 北京 100049

3. 中国科学院武汉植物园, 中国科学院植物种质创新与特色农业重点实验室, 湖北 武汉 430074

摘要 中药材淫羊藿富含朝藿定和淫羊藿苷等黄酮类化合物, 具有滋阴补肾、提高免疫力等功效, 有较大的药用价值。当前, 面对生产及育种过程中批量样品快速、无损检测需求的增加, 传统的化学分析方法难以满足需要, 而高效、廉价的现代高光谱分析技术备受青睐。但受制于光谱数据谱峰重叠及噪声的干扰, 全波段光谱分析建模存在模型精度不高和运行效率低的问题。利用便携式地物光谱仪器获取淫羊藿可见—近红外光谱数据, 借助遗传算法(GA)特征波段选择方法剔除无关波段, 并与偏最小二乘回归(PLSR)分析建模技术结合, 构建淫羊藿药用组分(朝藿定 A、朝藿定 B、朝藿定 C 和淫羊藿苷)高光谱 GA-PLSR 校正模型, 探讨淫羊藿药用组分含量高效分析预测的可行性, 并挖掘获取淫羊藿品质鉴定的重要光谱响应波段。结果表明: 高光谱分析结合化学计量学在淫羊藿有效药用组分的快速无损检测方面具有相当大的潜力。与全波段 PLSR 校正模型相比, 通过 GA 迭代优化, 参与建模的有效光谱数据得到简化, GA-PLSR 模型的测量精度和稳定性得到明显提升。主要表现在交叉验证的决定系数(R^2_{cv})得到明显提高, 交叉验证的均方根误差(RMSE_{cv})普遍降低。其中, 四种药用组分校正模型的 R^2_{cv} 分别从 0.645, 0.720, 0.718 和 0.642 提升为 0.671, 0.835, 0.782 和 0.796; 同时, 其对应的 RMSE_{cv} 值分别由 2.102, 2.896, 21.069 和 1.221 降为 2.071, 2.230, 18.656 和 0.912。此外, 明确了红边波段 690~740 nm 以及 420 nm 附近波段为淫羊藿药用组分朝藿定 A、朝藿定 B、朝藿定 C 和淫羊藿苷光谱鉴别分析的重要响应波段。该研究为高光谱技术淫羊藿品质准确高效鉴定和光谱传感器的波段设计提供一定的理论依据。

关键词 淫羊藿; 药用组分; 高光谱技术; 遗传算法; 偏最小二乘回归; 重要波段

中图分类号: O657.3 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2022)05-1445-06

引言

淫羊藿是小檗科(Berberidaceae)淫羊藿属(*Epimedium* L.)多年生草本植物, 作为滋补类中药始载于《神农本草经》, 在我国有 2 000 多年的使用历史^[1]。其有效药成分朝藿定 A、朝藿定 B、朝藿定 C 和淫羊藿苷为黄酮类化合物, 临床上具有滋阴补肾、强筋骨、益精气、提高免疫力等多种功效, 是最具开发潜力的大宗常用中药材之一^[2]。目前, 中药材淫羊藿药用成分含量的测量主要通过高效液相色谱法^[3]。该方法准确可靠、灵敏度高, 但需要做较多前处理, 耗时、费力、成本高, 无法满足企业生产及育种过程中批量样品的快速、无损检测。相比之下, 可见—近红外高光谱分析技术具有快速、

廉价、无损、可重复等优点; 利用中药材中 C—H, O—H 和 N—H 等基团的光谱差异, 借助化学计量学方法, 可以实现中药成分的快速分析与鉴定^[4]。利用现代高光谱技术, 学者们在中药材的品质鉴定、蛋白质含量信息获取及其产地来源判定等方面进行了系列研究, 取得了良好的效果^[5-6]。

近年来, 在淫羊藿药用成分光谱分析领域, 科研人员主要利用傅里叶变换红外光谱仪并配有积分球等测样器件获取数据, 并以此为基础建立了多种化学计量模型^[4,7]。然而, 随着淫羊藿育种、栽培技术的发展, 推升了田间原位、实时数据的观测需求。受测试环境和光谱仪器差异的干扰, 获取的光谱数据无法避免的包含诸如样品背景、杂散光等无关噪音信息, 且在波长范围、分辨率等指标上存有较大差异, 构建的模型的适用性必将受到很大的限制。

收稿日期: 2021-04-20, **修订日期:** 2021-08-17

基金项目: 国家自然科学基金面上项目(32071675, 31670346), 湖北省研发创新平台建设项目(2018AFA164)资助

作者简介: 姜庆虎, 1986 年生, 中国科学院武汉植物园副研究员 e-mail: jiang8687@163.com

* 通讯作者 e-mail: yanjunzhang@wbcas.cn; qiongl@wbcas.cn

为了获取预测精度高、稳健性强的高光谱分析模型,人们普遍关注多种化学计量方法和建模策略之间的比较研究。例如,基于获取的全波段光谱信息,运用多元线性回归(MLR)、偏最小二乘回归(PLSR)、人工神经网络(ANN)、支持向量机(SVM)等多种算法建立中药材药用成分的光谱反演模型,并比较其预测精度和运行效率^[4, 8-9]。淫羊藿光谱数据是其所含多种化学成分光谱的综合叠加,波长信息变量多,谱峰重叠非常严重,如果直接将全波段光谱用于建模分析势必造成模型精度和运行效率的降低^[9]。因此,在用化学计量学方法建模的同时,有必要采用适当的方法剔除无关波段变量,以期得到预测能力更强、稳健性更好的校正模型。

为了改善现有模型的局限性,提升淫羊藿药用成分的光谱预测效率,以便便携式地物光谱仪器数据获取载体,借助遗传算法(genetic algorithm, GA)选取淫羊藿药用成分的特征光谱波段,并评估该算法与传统的 PLSR 模型相结合的实用性。

1 实验部分

1.1 样品

实验材料采自中国科学院武汉植物园同质园引种的淫羊藿野生品种,涵盖了我国的 35 个品种。采集同一植株无损成熟叶片,经烘干研磨后,过 80 目筛,分别供光谱和黄酮类有效成分的测定。其中,淫羊藿有效成分朝霍定 A、朝霍定 B、朝霍定 C 和淫羊藿苷含量通过高效液相色谱法测定。

1.2 光谱采集

采用便携式地物光谱仪(ASD Field Spec 3)进行淫羊藿粉末光谱数据的测定,光谱波段范围 350~2 500 nm。采样间隔为 1.4 nm(350~1 000 nm)和 2 nm(1 000~2 500 nm)。整个过程在暗室中进行。处理好的淫羊藿样品置于培养皿中,并将其表面刮平;50 W 卤素灯作为光源,光源距样品表面 35 cm,入射角度为 45°;采用 10°视场角裸光纤探头,距样品表面垂直上方 10 cm 处;每测量 10 组样品光谱进行一次白板校正。每个样品重复测量获取 10 条光谱曲线,算术平均后得到淫羊藿样品的反射光谱数据。

为了减少仪器背景随机噪声、样本自身和光散射等因素对光谱信号的干扰,对原始光谱进行了一系列预处理,包括平滑(savitzky-golay convolution smoothing, SG)、一阶求导(first derivative, FD)、标准正太变量变换(standard normal

variable transformation, SNV)和均值中心化(mean center, MC)。相关处理过程在 Matlab 2013a 软件中进行。

1.3 特征波段选取与模型构建

应用遗传(GA)算法结合偏最小二乘回归(PLSR)方法筛选淫羊藿药用组分朝霍定 A、朝霍定 B、朝霍定 C 和淫羊藿苷的特征波段,以建立抗外界因素的稳健分析模型。其中,遗传算法是 Holland 于 1975 年提出的一种模拟生物自然选择遗传进化过程的智能优化方法^[10]。具体操作步骤如下:首先按照遗传算法建模框架,划分适宜的光谱区间大小,并将每个光谱区间作为一个基因进行 0/1 字符二进制编码,随机生成的不同组合构成初始种群;接着采用偏最小二乘交互验证中的预测值和标准值的相关系数以及均方根误差作为评价指标,并变换得到适应度函数;在此基础上,通过全局搜索,采用染色体信息随机交叉、变异等操作,实现适者生存优胜劣汰,最终得到最佳的建模波段。相关操作基于 Matlab 2013a(Math Works Inc., U.S.)平台,使用 PLS-Toolbox 7.8.1 工具包实现。相关参数设置如下:种群大小 64,窗口大小 20 nm,最大繁殖代数 100,交叉概率 0.5,变异概率 0.005。

使用优选后的光谱波段,运用 PLSR 方法,分别构建淫羊藿药用成分朝霍定 A、朝霍定 B、朝霍定 C 和淫羊藿苷含量的高光谱 GA-PLSR 校正模型,并以全光谱波段 PLSR 校正模型作为参照。受样品量大小的限制,模型性能用舍一交叉验证的均方根误差(RMSE_{CV})和决定系数(R_{CV}^2)进行评价。

2 结果与讨论

2.1 样品化学组分与光谱数据分析

建模所用的淫羊藿药用组分含量的化学值及其统计特征如表 1 所示。统计显示,样品药用组分的含量变动范围较广,其中:朝霍定 C 的平均含量最高,变化范围为 0.11~159.35 mg·g⁻¹;朝霍定 A 的变幅为 0.11~11.7 mg·g⁻¹;朝霍定 B 的变幅为 0.02~19.97 mg·g⁻¹;淫羊藿苷的变化范围为 0.09~7.74 mg·g⁻¹。此外,样品药用组分含量存在很大的变异性特征,其变异系数为 89.47~110.78 mg·g⁻¹。上述统计结果表明,研究样品较大的数据分异性有助于鲁棒性模型的建立,为探索淫羊藿药用组分含量光谱反演的可行性提供了良好的数据支持。

表 1 样品中朝霍定 A、朝霍定 B、朝霍定 C 和淫羊藿苷含量统计结果

Table 1 Summary statistics of epimedin A, epimedin B, epimedin C, and icariin contents

药用组分	最小值/ (mg·g ⁻¹)	最大值/ (mg·g ⁻¹)	中位数/ (mg·g ⁻¹)	均值/ (mg·g ⁻¹)	标准差/ (mg·g ⁻¹)	变异系数/ %
Epimedin A	0.11	11.7	3.20	3.78	3.55	93.92
Epimedin B	0.02	19.97	2.81	5.01	5.55	110.78
Epimedin C	0.11	159.35	20.85	38.75	40.26	103.90
Icariin	0.09	7.74	2.41	2.28	2.04	89.47

经光谱预处理之后的淫羊藿反射光谱曲线如图 1 所示。其中,图 1(a)为经平滑处理后的光谱,可见淫羊藿样品集的

反射光谱波动趋势大致相同,但不同样品的光谱反射峰强度不同,说明样品中药用组分含量不同。图 1(b)和(c)为经一

阶导数变换和标准正太变量变换后的光谱，经预处理后可以显著减少样品表面散射和光程变化的干扰，同时消除了部分噪声和基线漂移对定标结果的影响，使得光谱分布更为集中，且光谱的吸收峰得到了明显增强。图 1(d)为经上述预处理后进行了均值中心化的光谱，经过均值中心化，样品光谱之间的差异得到增强，从而可以提高后期光谱建模的稳健性和预测能力。

2.2 样品药用组分光谱反演建模分析与对比

运用 PLSR 和 GA-PLSR 方法分别建立淫羊藿药用组分的光谱反演模型，结果如表 2 所示。其中，基于全波段的 PLSR 校正模型建模效果良好，模型的 R^2 为 0.801~0.999。然而，受本研究中样本数量的限制，且样品间药用组分含量

差异较大的影响，模型交叉验证的 R_{CV}^2 值显著降低(0.642~0.720)，同时 $RMSE_{CV}$ 值明显增加，证明模型的稳健性较低。与全波段模型相比，在 GA 算法特征波段选择的基础上，经过剔除冗余与不相关的波段变量，可以有效减少参与建模的波段数量(筛选出的朝霍定 A、B、C 和淫羊藿苷光谱建模波段数量分别为 456, 675, 275 和 375 个)，简化模型提高运行效率的同时，各药用组分校正模型的预测性能得到明显提升。具体表现在：朝霍定 A 模型的 R_{CV}^2 从 0.645 提升为 0.671，朝霍定 B 模型的 R_{CV}^2 从 0.720 提升为 0.835，朝霍定 C 模型的 R_{CV}^2 从 0.718 提升为 0.782，淫羊藿苷模型的 R_{CV}^2 从 0.642 提升为 0.796；同时，其对应的 $RMSE_{CV}$ 值分别由 2.102, 2.896, 21.069 和 1.221 降为 2.071, 2.230, 18.656

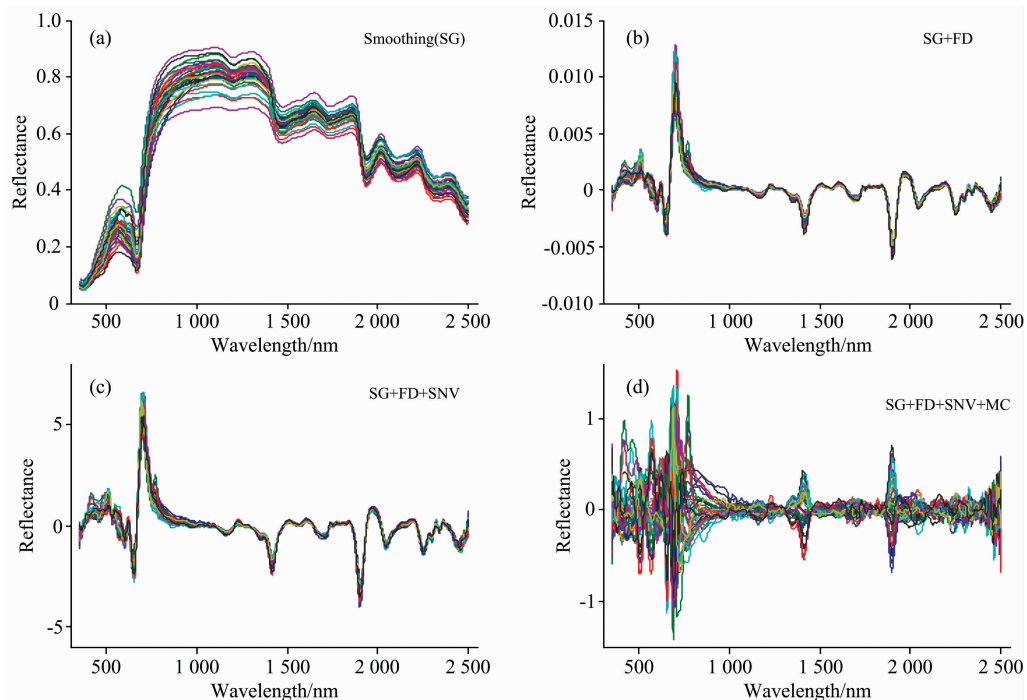


图 1 预处理后的淫羊藿反射光谱曲线

(a)~(d)分别经平滑、平滑/一阶求导、平滑/一阶求导/标准正太变量变换、平滑/一阶求导/标准正太变量变换/均值中心化处理后的淫羊藿光谱反射曲线

Fig. 1 Spectral reflectance curves after pretreatment with (a) SG; (b) SG and FD; (c) SG, FD and SNV; (d) SG, FD, SNV and MC

表 2 模型构建与交叉验证

Table 2 Model construction and cross-validation

药用组分	建模方法	波段数量	光谱预处理最佳方案	校正模型		交叉验证模型	
				R^2	RMSE	R_{CV}^2	$RMSE_{CV}$
Epimedin A	PLSR	2 151	SG+FD+SNV+MC	0.991	0.336	0.645	2.102
	GA-PLSR	456	SG+FD+SNV+MC	0.985	0.425	0.671	2.071
Epimedin B	PLSR	2 151	SG+FD+SNV+MC	0.999	0.090	0.720	2.896
	GA-PLSR	675	SG+FD+SNV+MC	0.995	0.381	0.835	2.230
Epimedin C	PLSR	2 151	SG+MC	0.801	17.684	0.718	21.069
	GA-PLSR	275	SG+MC	0.867	14.479	0.782	18.656
Icariin	PLSR	2 151	SG+FD+MC	0.893	0.656	0.642	1.221
	GA-PLSR	375	SG+FD+MC	0.980	0.284	0.796	0.912

和 0.912。因此,通过 GA 算法迭代优化,能够更好地对淫羊藿药用组分对应的光谱区域进行解释,从而建立抗外界因素干扰的光谱分析模型,以提高模型的预测能力和稳定性。

2.3 样品药用组分重要光谱波段提取与分析

淫羊藿药用组分的重要光谱响应波段用 PLSR 模型的回

归系数(b 系数)和变量重要值(variables importance in the projection, VIP 值)联合确定^[11]。如图 2 中灰色部分所示,当 VIP 值大于 1,且 b 系数大于其标准差(σ)时,该波段被识别为药用组分朝霍定 A、B、C 和淫羊藿苷的重要光谱响应波段。

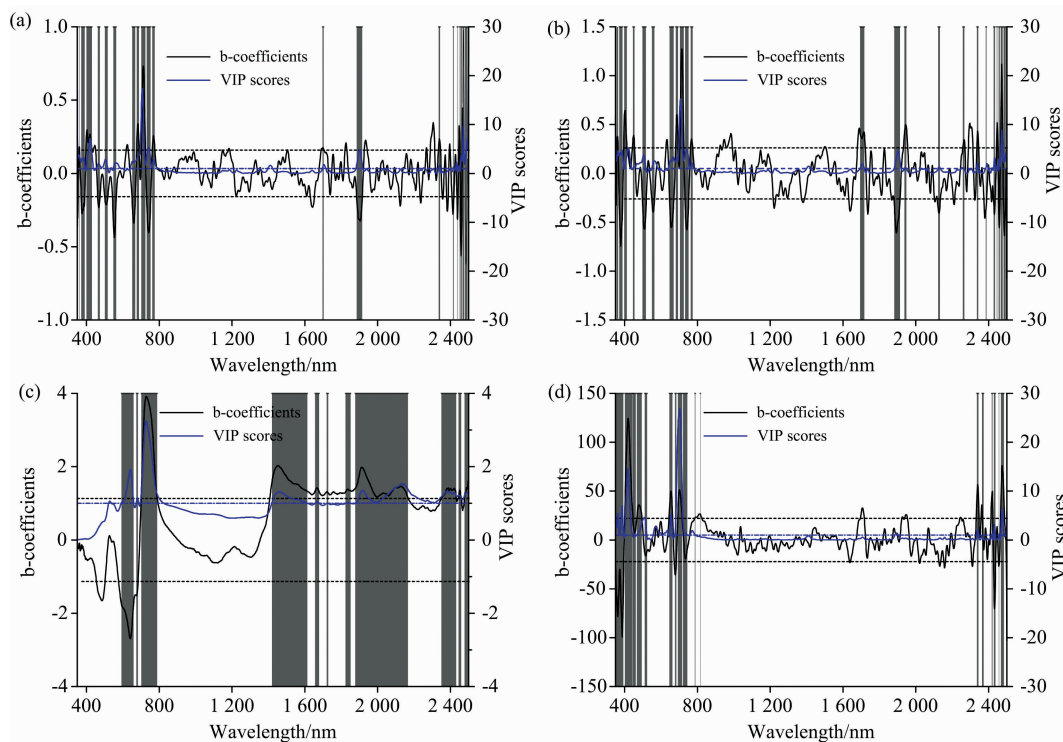


图 2 基于可见-近红外光谱的朝霍定 A(a)、朝霍定 B(b)、朝霍定 C(c)和淫羊藿苷(d)PLSR 校正模型交叉验证的 VIP 值和 b 系数曲线图

Fig. 2 VIP scores (blue line) and b -coefficients (black line) associated with the cross-validation of PLSR model for Epimedin A (a), Epimedin B (b), Epimedin C (c) and Icarin (d) predictions using Vis-NIR spectroscopy

图 2 为基于全波段 PLSR 模型的淫羊藿各药用组分的重要光谱响应波段。其中,提取的朝霍定 A、B、C 和淫羊藿苷的重要的光谱响应波段数量分别为 300, 341, 841 和 222 个(即图中灰色阴影部分涉及的波段数量)。值得注意的是,灰色部分在可见光波段(380~780 nm)和短波红外的中长波段(1 400~2 500 nm)区域内较为集中。这是由于淫羊藿有效成分为黄酮类化合物,其分子化学结构存在一定的相似性,受低能电子跃迁和含氢原子团伸缩振动(如 C—H 伸展, O—H 变形, CH₂ 变形)的倍频和合频吸收作用,从而在上述波段范围内存在相似或近似的光谱响应。此外,在红边(690~740 nm)区域内,朝霍定 A、B 和 C 光谱反演模型的 VIP 值和 b 系数值均达到最大峰值,因此该区间光谱的反射特征有望成为判定淫羊藿药材中朝霍定含量高低的重要依据。对于淫羊藿苷来说,光谱反演模型的 b 系数在 420 nm 波段附近达到最大值,同时 VIP 值也处于较高的峰值,因此 420 nm 附近波段有望成为判定淫羊藿苷含量高低的重要依据。

借助相同的重要波段选取方法,获取了基于特征波段的 GA-PLSR 模型淫羊藿药用成分的重要光谱响应波段(如图 3 所示)。相比之下,GA-PLSR 模型确定的重要波段数量显著

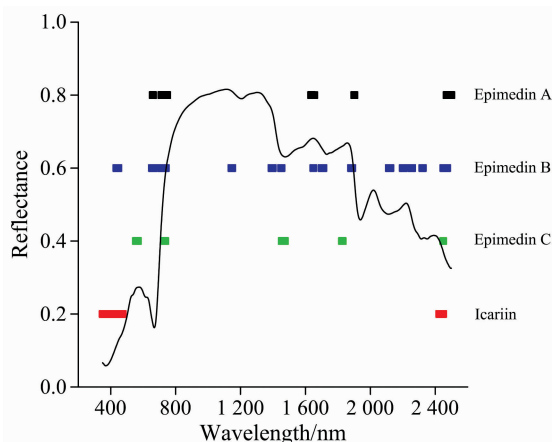


图 3 基于 GA-PLSR 模型的朝霍定 A、朝霍定 B、朝霍定 C 和淫羊藿苷重要光谱响应波段

Fig. 3 The important wavelengths of GA-PLSR model for Epimedin A, Epimedin B, Epimedin C and Icarin predictions

降低, 其中朝霍定 A、B、C 和淫羊藿苷的重要波段数量分别降为 51, 136, 50 和 66 个。然而, 在红边波段(690~740 nm)和 420 nm 波段附近同样提取出淫羊藿药用成分的重要光谱波段。上述重要波段的提取, 可为无人机平台搭载高光谱相机(400~1 000 nm)实时监测大田淫羊藿植株的生长和品质状况, 确定其最佳收获时间提供一定的理论依据。

3 结 论

以便携式地物光谱仪获取的淫羊藿反射光谱数据为基础, 将遗传算法(GA)与偏最小二乘回归(PLSR)结合, 在特

征波段选择的基础上建立了淫羊藿药用成分朝霍定 A、B、C 和淫羊藿苷含量的校正模型。与全波段 PLSR 校正模型相比, GA-PLSR 模型减少建模所用变量的同时, 其交叉验证的决定系数(R_{cv}^2)得到显著提升, 交叉验证的均方根误差(RMSE_{cv})明显降低, 表明模型具有较高的稳定性和测量精度。此外, 本研究提出红边波段 690~740 nm 以及 420 nm 附近波段为淫羊藿药用组分光谱反演的重要波段。有望为今后光谱仪器的波段设计以及淫羊藿品质的高效、准确鉴定提供一定的依据, 并为实现大田生产中的实时动态监测和科学管理带来了可能。

References

- [1] LIANG Qiong, ZHANG Yan-jun, XU Yan-qin, et al(梁 琼, 张燕君, 徐艳琴, 等). Plant Science Journal(植物科学学报), 2013, 31(4): 422.
- [2] YU Dong-yue, WANG Ying, SUN Wei, et al(于东悦, 王 瑛, 孙 伟, 等). World Science and Technology/Modernization of Traditional Chinese Medicine and Materia Medica(世界科学技术-中医药现代化), 2018, 20(11): 2058.
- [3] Ma H, He X, Yang Y, et al. Journa of Ethnopharmacology, 2011, 134(3): 519.
- [4] Yang Y, Wu Y, Li W, et al. Spectrochim Acta A: Mol. Biomol. Spectrosc., 2018, 191: 233.
- [5] LI Ya-hui, LI Yan-xiao, TAN Wei-long, et al(李亚惠, 李艳肖, 谭伟龙, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(12): 3878.
- [6] NIU Li-li, ZHANG Hua-feng, CHEN Le, et al(牛丽丽, 张华峰, 陈 乐, 等). Chinese Bulletin of Botany(植物学报), 2014, 49(5): 611.
- [7] YU Xiao-xue, YI Yin, ZHOU Ning, et al(于晓雪, 乙 引, 周 宁, 等). Chinese Journal of Spectroscopy Laboratory(光谱实验室), 2012, 29(3): 1379.
- [8] Luo Q, Yun Y, Fan W, et al. RSC Advances, 2015, 5(7): 5046.
- [9] Yang Y, Liu X, Li W, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2017, 171: 351.
- [10] Leardi R. J. Chromatogr. A, 2007, 1158(1-2): 226.
- [11] Jiang Q, Chen Y, Hu J, et al. Remote Sensing, 2020, 12(18): 3103.

Rapid Measurement of the Pharmacological Active Constituents in Herba Epimedii Using Hyperspectral Analysis Technology

JIANG Qing-hu¹, LIU Feng¹, YU Dong-yue^{2, 3}, LUO Hui^{2, 3}, LIANG Qiong^{3*}, ZHANG Yan-jun^{3*}

1. Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China

Abstract Herba Epimedii contains high amounts of flavonoids, such as epimedin and icariin, which are efficient in tonifying kidney and improving immunity. Nowadays, various chemical analysis methods have been applied to measure the flavonoid content of Herba Epimedii. However, these traditional methods are destructive, time-consuming, and costly and cannot meet the requirements of massive samples analysis in pharmaceutical production and plant breeding. As a rapid and effective tool for quantitative determination and process monitoring, modern hyperspectral analysis technology has earned more and more concerns. However, for the full-range spectra, the existence of insignificant and irrelevant spectral variables can weaken the calibration models' accuracy and efficiency. Therefore, the spectral variables selection is essential to improve the performance of the final models by eliminating the uninformative bands. In this study, the partial least squares regression (PLSR) coupled with the genetic algorithm (GA) variables selection procedure, namely GA-PLSR, was used to estimate epimedin A, epimedin B,

epimedin C, and icariin content in Herba Epimedii. This paper aims to explore the feasibility of hyperspectral analysis technology in the measurement of the pharmacologically active constituents in Herba Epimedii and further explore their important spectral response bands. The results show that the hyperspectral analysis technology combined with chemometrics exhibited considerable potential for rapid and nondestructive assessment of Herba Epimedii. When compared with full-spectrum PLSR models, GA-PLSR models could improve the accuracies and robustness of epimedin A, epimedin B, epimedin C, and icariin content measurements (with R_{CV}^2 values increased from 0.645, 0.720, 0.718, and 0.642 to 0.671, 0.835, 0.782, and 0.796, and with RMSE_{CV} values declined from 2.102, 2.896, 21.069, and 1.221 to 2.071, 2.230, 18.656, and 0.912, respectively). Besides, we found some feature wavelengths, mainly around 690 ~ 740 and 420 nm, which play important roles in detecting pharmacologically active constituents in Herba Epimedii. Given these desirable findings, this study can provide a valuable reference for the rapid and accurate measurement of epimedin A, epimedin B, epimedin C, and icariin contents by hyperspectral technology, can provide a theoretical basis for the design of spectral sensors in qualifying Herba Epimedii.

Keywords Herba Epimedii; Pharmacological active constituents; Hyperspectral analysis technology; Genetic algorithm; Partial least squares regression; Important bands

(Received Apr. 20, 2021; accepted Aug. 17, 2021)

* Corresponding authors

关于《光谱学与光谱分析》调整审稿费收费标准的通知

尊敬的《光谱学与光谱分析》广大作者、读者：我刊自 2018 年 7 月 1 日以后登记的稿件向投稿作者收取审稿费 200 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

1. 没有创新的一般性稿件，请您不要投稿。
2. 没有国家级基金资助的稿件，请您不要投稿。
3. 不是光谱专业的稿件，请您不要投稿。
4. 与其他文章重合率超过 10% 的稿件，请您不要投稿。

所投稿件经初审通过后，作者会收到缴纳审稿费的通知。请作者及时从我刊网站 (<http://www.gpxygpfx.com>) 查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，我刊不再受理。交费后我刊开据增值税电子普通发票，并传至作者提供的电子邮箱，作者可自行打印。

联系电话：010-62181070, 62182998

电子邮箱：chngpxygpfx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2018 年 6 月 30 日