基于拉曼光谱技术的桑椹花色素苷快速检测研究

张慧洁，蔡　冲*，崔旭红，张雷蕾

中国计量大学生命科学学院，浙江 杭州　310018

摘　要　花色素苷是一种天然的水溶性黄酮类色素，具有多种药用价值，广泛存在于桑椹中，成为评价桑椹产品品质的重要指标。传统检测方法费时费力，因此实现花色素苷含量的快速检测对于桑椹产品的开发利用至关重要。该研究以桑椹中的花色素苷为研究对象，探索花色素苷与拉曼光谱特性之间的关系及拉曼光谱技术对其定量检测的可行性。对桑椹及3种花色素苷标准品的拉曼光谱进行了分析，其中可将545，634和737 cm^{-1}处的峰位作为桑椹中花色素苷的拉曼特征峰，以此判断桑椹中是否含有花色素苷，并根据其峰值的高低来定性判断花色素苷含量多少。运用多元散射校正(MSC)、基线校正(airPLS)、归一化(Normalized)三种方法及其组合方法进行光谱数据预处理，并结合PLSR筛选最佳预处理方式。比较发现最佳预处理为airPLS＋MSC＋Normalized，其PLSR模型效果较好，建模集决定系数为0.97，$RMSE_c$为2.74，预测集决定系数为0.82，$RMSE_p$为13.69。基于airPLS＋MSC＋Normalized预处理后的光谱，采用竞争性自适应重加权算法(CARS)对光谱进行特征波长筛选，将筛选出的波长变量作为输入变量分别建立了PLSR模型和SVR模型，研究两种模型的预测效果。结果表明经过CARS处理的两种模型均能对花色素苷的含量进行准确预测，其中经过CARS变量筛选建立的SVR模型效果最好，建模集决定系数为0.98，$RMSE_c$为1.92，预测集决定系数为0.94，$RMSE_p$为4.70，预测精度较高。因此拉曼光谱技术可以实现对桑椹中花色素苷含量的快速、准确预测。

关键词　拉曼光谱；花色素苷；桑椹；特征提取；PLSR；SVR

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)12-3771-05

引　言

花色素苷是一种天然的水溶性黄酮类色素，具有保护人体心血管、降血糖、护肝脏、抗癌和刺激视紫红质再生等功能[1]。桑椹因含有丰富的花色素苷而成为食品、保健品和药品的良好加工原料。花色素苷不稳定，在加工和储藏中易受光照、热、酸等影响致使颜色变淡、生物活性降低[2]，给食品加工产品的品质保持造成困难，而某些产品宣称含有丰富的花色素苷以此欺骗消费者。因此建立一种快速、准确的桑椹中花色素苷含量的检测方法对于桑椹产品的品质检测、分级及开发利用具有重要意义。

目前测定花色素苷常用的方法如高效液相色谱法、分光光度法等，检测步骤复杂，耗时长且具有破坏性，难以满足样本中花色素苷快速检测的需求[3]。拉曼光谱技术以拉曼散射效应为基础，光波被散射后频率发生变化，频率位移与发生散射的分子结构有关，从而完成对不同结构分子的检测。拉曼光谱不需要样品前处理过程，样品可通过光线直接测量，方法快速、简单、可重复性强[4]。已经广泛的应用在食品中糖类、维生素、蛋白质、DNA和色素等成分的定性和定量分析中[5-6]。但目前国内外采用拉曼光谱技术对花色素苷的应用研究较少，未见有拉曼光谱技术对花色素苷含量检测的文献报道。本文以桑椹为实验材料，分析花色素苷的拉曼光谱特性，研究桑椹中的花色素苷与其拉曼光谱特性之间的相关性，并建立桑椹花色素苷的定量模型，实现花色素苷的定量检测。

1　实验部分

1.1　材料与仪器

收稿日期：2020-11-02，**修订日期**：2021-02-14

基金项目：浙江省自然科学基金项目(LY21C150007，LY17C150003)，浙江省重点研发计划项目(2019C02074)，国家自然科学基金项目(31401923)资助

作者简介：张慧洁，女，1994年生，中国计量大学生命科学学院硕士研究生　　e-mail：1650970552@qq.com

*通讯作者　　e-mail：ccjacn@cjlu.edu.cn

实验选用“大 10”品种桑椹，购买于杭州下沙超市，并用冰盒运输至实验室。挑选大小均匀的无机械损伤桑椹 510 个，去除果柄，用去离子水清洗干净，每取 10 个桑椹作为一个样本采集完光谱之后立即榨汁并过滤，共计 51 个样本，将制得的样本溶液暂存于 4 ℃冰箱中用于花色素苷的理化检测。

标准品矢车菊素-3-O-葡萄糖苷(Cyanidin-3-O-glucoside，C3G)，矢车菊素-3-O-芸香糖苷(Cyanidin-3-O-Rutinoside，C3R)，天竺葵素-3-O-葡萄糖苷(Pelargonidin 3-O-glucoside，P3G)购买于 Macklin 试剂公司，纯度均大于 95%，分别配制成 2 mg·mL^{-1}的水溶液，并模拟桑椹中的花色素苷含量按矢车菊素-3-O-葡萄糖苷，矢车菊素-3-O-芸香糖苷，天竺葵素-3-O-葡萄糖苷为 45%，45%和 10%的比例配制成 2 mg·mL^{-1}混合标准溶液，用于花色素苷的拉曼光谱分析及桑椹中花色素苷拉曼光谱特征峰的提取。

拉曼光谱仪为实验室自行搭建，主要包括 QE-Pro 光谱仪(Ocean Opticis 公司)、Laser-785 nm 激光器、传输光纤、拉曼检测探头和置物台五个部分；紫外分光光度计(UV-1800，岛津)。

1.2 拉曼光谱采集

拉曼光谱的激发波长为系统默认 785 nm，波长检测范围为 200～2 870 cm^{-1}，光谱采集时选用优化后的参数即激光功率 350 mW、平均次数 2 次、积分时间 3000 ms，采样距离 3～5 mm。为减少荧光产生的干扰，在暗室环境下进行光谱采集；每个桑椹取不同部位采集两次，每采集 10 个桑椹作为一个样本进行花色素苷理化测定，最后取均值作为一个样本的原始光谱。

1.3 花色素苷测定

桑椹中总花色素苷含量测定采用 pH 示差法[7]，重复测定 3 次。

1.4 光谱预处理及模型评价

应用 Matlab R2018b 分析软件对光谱进行预处理及定量模型的建立。样本拉曼光谱中含有的冗余信息使得模型的预测性能降低，所以对原始光谱数据采用多元散射校正(MSC)、基线校正(airPLS)、归一化(Normalized)及其组合方法进行预处理。确定最佳预处理后采用竞争性自适应重加权算法(CARS)提取全光谱中的有效波数，并且根据有效波数建立偏最小二乘回归(PLSR)模型[8]和支持向量机回归(SVR)模型[9]，取一种预测精度更高的方法来建立较为准确的预测模型。模型性能采用决定系数(R^2)和均方根误差(RMSE)进行评价。R_c^2，R_p^2 和 $RMSE_c$，$RMSE_p$ 分别表示建模集和预测集的决定系数与均方根误差。

2 结果与讨论

2.1 桑椹花色素苷的拉曼光谱分析

测定 C3G，C3R 及 P3G 三种花色素苷的拉曼光谱，如图 1，C3G，C3R 和 P3G 的拉曼光谱这与 Merlin 等[10]的研究结果一致。由于花青素在结构上存在相同的苯并吡啶部分，仅因苯环上的取代方式不同而有所区别；花色素苷的可见生色团主要位于苯并吡啶部分，而不是苯环上，因此不同的花色素苷具有很大的相似性[11]。花色素苷在 1 400～1 650 cm^{-1}之间的拉曼信号可归因于苯并吡啶部分和苯环的环状拉伸振动，1 335 cm^{-1}附近主要是由苯环取代引起的环间键拉伸，低光谱范围 500～900 cm^{-1}主要与糖基化模式有关[10]。

桑椹中的花色素苷主要为矢车菊素-3-O-葡萄糖苷(C3G)、矢车菊素-3-O-芸香糖苷(C3R)、天竺葵素-3-O-葡萄糖苷(P3G)和天竺葵素-3-O-芸香糖苷(P3R)，且矢车菊类花色素苷占总花色素苷的 90%及以上[12]。本文模拟桑椹中各种花色素苷的含量，将 C3G，C3R 及 P3G 三种花色素苷按照 45%，45%和 10%的比例均匀混合，并测定混合标准液的拉曼光谱，如图 2 所示。

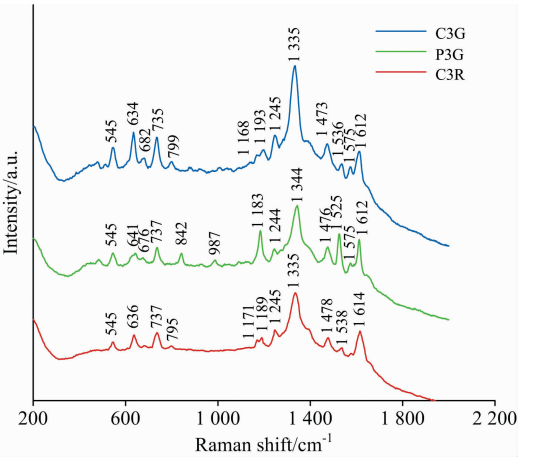

图 1 三种花色素苷标准溶液的拉曼光谱

Fig. 1 Raman spectra of three anthocyanin standard solutions

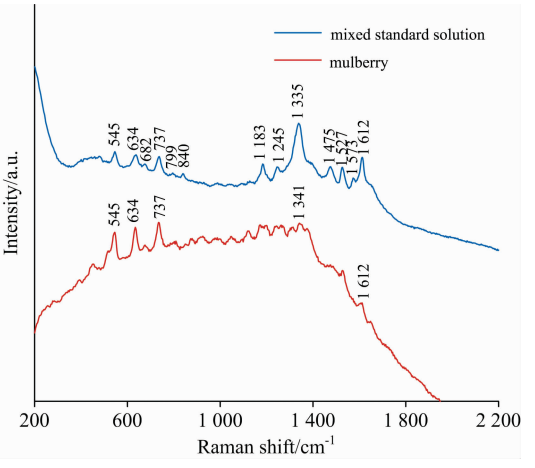

图 2 混合标准液和桑椹原始拉曼光谱

Fig. 2 Original Raman spectra of mixed standard solution and mulberr

矢车菊素-3-O-葡萄糖苷、矢车菊素-3-O-芸香糖苷、天竺葵素-3-O-葡萄糖苷及混合花色素苷标准溶液在波数 545，634，737，1 335 和 1 612 cm^{-1}附近均存在较强的拉曼峰，分别归结于 545 和 634 cm^{-1}处的 C—C 面内弯曲[10]，737 cm^{-1}处的 C—C—O 面内弯曲[11]，1 335 和 1 612 cm^{-1}处的内环

C—C 拉伸[10-11]。对比桑椹的原始拉曼光谱，如图 2 所示，由于桑椹所含成分较多，桑椹的拉曼光谱谱峰较多，各种成分之间相互影响，某些特征峰的波数与混合花色素苷相比发生了偏移，偏移均在 10 cm^{-1}之内，其在 545，634 和 737 cm^{-1}处有较强的拉曼特征峰，1 341 和 1 612 cm^{-1}处的峰强较弱，因此选择波数 545，634 和 737 cm^{-1}处的峰作为桑椹花色素苷的拉曼特征峰，通过桑椹拉曼光谱中这 3 处特征峰强度的高低即可定性判断桑椹中总花色素苷含量的多少。

2.2　桑椹花色素苷定量模型的建立

2.2.1　数据集样本划分

由于桑椹全光谱中存在较多的荧光背景以及噪声干扰，且花色素苷的光谱信息主要在 400～1 800 cm^{-1}波段之间，所以选择该波段光谱进行分析。采用 KS 算法将 51 个样本以约 4∶1 的比例划分为建模集和预测集。样本集的统计信息如表 1 所示。

表 1　桑椹样本集的统计信息

Table 1　Statistics of the mulberry sample set

	最大值/(mg·L^{-1})	最小值/(mg·L^{-1})	平均值/(mg·L^{-1})	标准差	个数
建模集	100.86	37.57	65.27	15.96	40
预测集	101.03	31.48	64.21	19.12	11

2.2.2　光谱预处理方法筛选

为了消除无关信息和噪声的影响，采用多元散射校正(MSC)、基线校正(airPLS)、归一化(Normalized)及其组合方法对桑椹样品原始拉曼光谱进行预处理。多元散射矫正能够有效地消除光谱散射的影响，增强与成分含量相关的光谱信息[13]；基线校正能够消除背景噪声以及基线漂移[14]；归一化的作用是消除数据量纲的影响，提高模型的运行速度。

结合 PLSR 对光谱预处理效果进行评价，各种预处理方法的预测结果如表 2 所示。

表 2　不同预处理方法的 PLSR 建模效果

Table 2　PLSR modeling effects of different preprocessing methods

预处理方法	建模集		预测集	
	R_c^2	$RMSE_c$	R_p^2	$RMSE_p$
无预处理	0.40	12.89	0.39	13.44
airPLS	0.82	6.77	0.70	15.91
MSC	0.46	12.03	0.41	13.98
Normalized	0.47	11.95	0.54	11.95
MSC+Normalized	0.47	11.35	0.46	12.14
airPLS+MSC	0.93	3.94	0.76	14.62
airPLS+Normalized	0.87	5.94	0.76	8.10
airPLS+MSC+Normalized	0.97	2.74	0.82	13.69

从表 2 中花色素苷 PLSR 模型的评价指标结果可以看出，与原始光谱相比，三种单一预处理方法有效的消除了基线漂移、光谱散射等产生的影响，建模集和预测集的决定系数均有不同程度的提高，其中经过 airPLS 预处理的模型决定系数达到 0.7，但其模型预测集的均方根误差较大。进一步研究了三种预处理方法组合的建模效果，研究发现不同组合顺序的建模效果一致(未在表中列出)，并且经过 airPLS+MSC+Normalized 处理后所建立的 PLSR 模型效果较好，建模集 R_c^2 为 0.97，$RMSE_c$ 为 2.74；预测集 R_p^2 为 0.82，$RMSE_p$ 为 13.69，较原始光谱模型有很大改善，但是预测集的 RMSE 值仍然较大，说明模型预测值与实际值误差较大，模型预测准确度欠佳。

2.2.3　基于 CARS 特征波长提取的定量模型

由于拉曼光谱中变量信息较多，变量之间存在较多冗余及无用信息，降低了模型的精度及速度，为了进一步提高预测集的预测精度，基于 airPLS+MSC+Normalized 处理后的桑椹拉曼光谱，研究了 CARS 特征波长提取方法的 PLSR 和 SVR 两种不同模型的建模效果。

采用 CARS 提取特征波长时，设定采样次数为 50 次，利用 5 折交叉验证法计算均方根误差(RMSECV)，结果如图 3(a)所示。从图 3(a)可以看出 RMSECV 值随着采样次数的增加呈现出先减小后增加的趋势，当采样次数为 22 时，RMSECV 值最小，此时得到的最优波长集包含 84 个特征波长，提取的特征波长在桑椹原始拉曼光谱中的分布如图 3(b)所示。图中 CARS 提取出的特征波长主要集中在波峰及波谷附近[15]，且在 545，634，737，1 341 和 1 612 cm^{-1}处均有分布，这与对比标准品确定的特征峰一致，由此说明 CARS 算法提取出的特征波长与花色素苷的含量具有高度的相关性，不仅

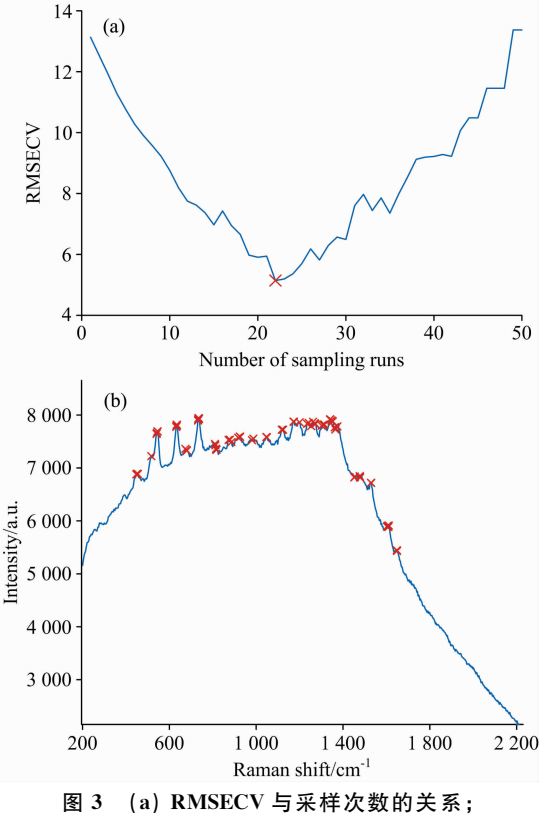

图 3　(a) RMSECV 与采样次数的关系；(b) 提取的特征波长分布

Fig. 3　(a) Relationship between RMSECV and sampling times; (b) Extracted characteristic wavelength distribution

降低了光谱的波长数量，提高模型的预测速度，而且保留了较多的有用信息。

将 CARS 提取出的特征波长作为输入变量，桑椹的花色素苷含量为输出变量分别建立了 PLSR 模型和 SVR 模型。支持向量机回归(SVR)选用 RBF 核函数，反复筛选模型参数，最终选择的最佳参数惩罚因子 C 为 32.0，核系数 g 为 0.001。两种模型的结果如表 3 所示。

表 3 CARS 筛选后 PLSR 及 SVR 模型预测结果

Table 3 Predicted results of PLSR and SVR model after CARS selection

Model	建模集		预测集	
	R_c^2	$RMSE_c$	R_p^2	$RMSE_p$
PLSR	0.97	2.49	0.91	5.23
SVR	0.98	1.93	0.94	4.70

对比 airPLS＋MSC＋Normalized 光谱预处理后的 PLSR 建模结果，CARS 算法提高了 PLSR 模型的预测精度，经过 CARS 筛选后 PLSR 建模集决定系数 R_c^2 为 0.97，均方根误差 $RMSE_c$ 为 2.49，预测集决定系数 R_p^2 为 0.91，均方根误差 $RMSE_p$ 为 5.23。比较 PLSR 和 SVR，经过本方法处理后的两种模型都能够实现对桑椹中花色素苷的含量的测定，SVR 模型的效果最好，预测集决定系数 R_p^2 为 0.94，均方根误差 $RMSE_p$ 为 4.70。证明拉曼光谱能有效的实现对桑椹花色素苷含量的准确、快速的预测。经 CARS 处理后两种模型的预测结果如图 4 所示。

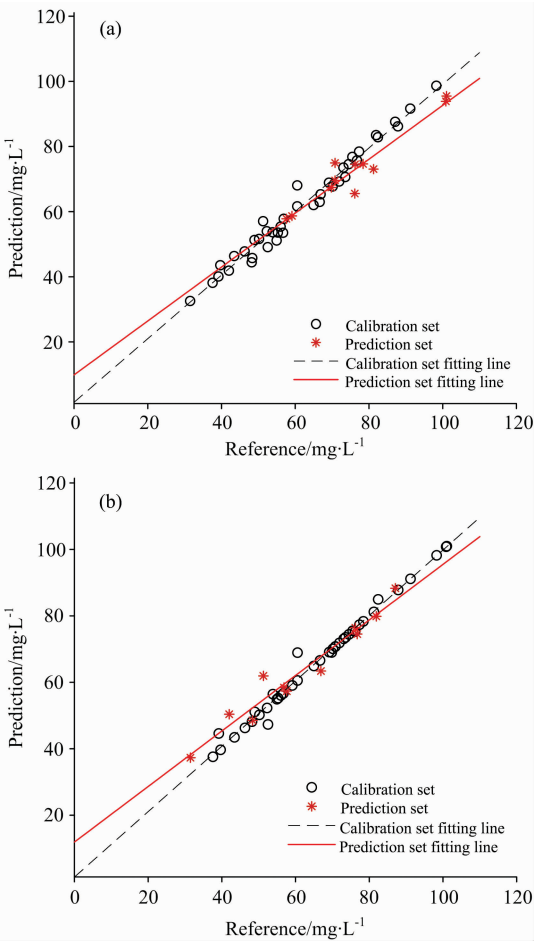

图 4 (a)PLSR 模型；(b)SVR 模型

Fig. 4 (a)PLSR model；(b)SVR model

3 结 论

利用拉曼光谱检测技术对桑椹中的花色素苷进行了原位、准确、快速检测研究。(1)分析了桑椹的拉曼图谱，其中波数 545，634 和 737 cm^{-1} 可作为桑椹花色素苷的特征峰，以此检测桑椹中是否含有花色素苷，并根据特征峰强度的高低来定性判断桑椹样品中的花色素苷含量多少。(2)三种光谱预处理方法中，最佳预处理方式为 airPLS＋MSC＋Normalized，其建立的花色素苷含量的 PLSR 模型效果最好，预测集 R_p^2 和 $RMSE_p$ 分别为 0.82 和 13.69。(3)基于 airPLS＋MSC＋Normalized 处理后的光谱，选用 CARS 算法进行特征波长提取并建立了 PLSR 模型和 SVR 模型，结果表明 CARS 算法不仅减少了模型的输入数量，而且筛选出的波长变量与对比标准品确定的特征峰一致，明显提高了预测精度，且适用于 PLSR 和 SVR 两种模型。其中 SVR 模型预测效果最好，其预测集的 R_p^2 和 $RMSE_p$ 分别为 0.94 和 4.70。研究表明拉曼光谱结合 airPLS＋MSC＋Normalized 预处理及 CARS 波长提取可以为桑椹花色素苷含量的定量分析提供一种快速准确的分析方法。

References

[1] Yousuf B, Gul K, Wani A A, et al. Critical Reviews in Food Science and Nutrition, 2016, 56(13): 2223.

[2] Ali H M, Almagribi W, AlRashidi M N. Food Chemistry, 2016, 194: 1275.

[3] Moldovan B, David L. Foods (Basel, Switzerland), 2020, 9(9): 1266.

[4] LIU Chen, CHEN Fu-sheng, XIA Yi-miao, et al(刘 晨, 陈复生, 夏义苗, 等). The Food Industry(食品工业), 2020, 41(4): 267.

[5] Sebben J A, Espindola J D S, Ranzan L, et al. Food Chemistry, 2018, 245: 1224.

[6] Richardson P I C, Muhamadali H, Ellis D I, et al. Food Chemistry, 2019, 272: 157.

[7] GUO Hao-ran, ZHENG Xin-yi, ZHANG Jing, et al(郭浩然, 郑心怡, 张 静, 等). Science and Technology of Food Industry(食品工业科技), 2020, 41(9): 255.

[8] Bedin F C B, Faust M V, Guarneri G A, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 245: 118834.

[9] Ma Zhixin, Lu Xinghua, Song Xing, et al. International Journal of Computational and Engineering, 2018, 3(3): 21.
[10] Merlin J C, Statoua A, Cornard J P, et al. Phytochemistry, 1993, 35(1): 227.
[11] Zaffino C, Russo B, Bruni S. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 149: 41.
[12] LI Meng-li, MA Jian-yong, LI Chun-mei(李梦丽, 马建勇, 李春美). Food Science(食品科学), 2018, 39(11): 75.
[13] Chen Zeling, Wu Ting, Xiang Cheng, et al. Molecules, 2019, 24(15): 2851.
[14] Zhang Feng, Tang Xiaojun, Tong Angxin, et al. Spectroscopy Letters, 2020, 53(3): 222.
[15] OUYANG Ai-guo, ZHANG Yu, TANG Tian-yi, et al(欧阳爱国, 张 宇, 唐天义, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(6): 1772.

Rapid Detection of Anthocyanin in Mulberry Based on Raman Spectroscopy

ZHANG Hui-jie, CAI Chong*, CUI Xu-hong, ZHANG Lei-lei
College of Life Science, China Jiliang University, Hangzhou 310018, China

Abstract Anthocyanin is a natural water-soluble flavonoid pigment with various medicinal values, which is widely found in mulberry and has become an important indicator for evaluating the quality of mulberry products. Because the implementation of the traditional detection methods could cost a lot of time and effort, it is significant to achieve the rapid detection of anthocyanin content in the development and utilization of mulberry products. In this study, anthocyanin in mulberry was taken as the research object to explore the relationship between anthocyanin and Raman spectral characteristics and the feasibility of quantitative detection of anthocyanin by Raman spectroscopy. The Raman spectra of mulberry and three kinds of anthocyanin were analyzed. The peak positions at 545, 634 and 737 cm^{-1} could be regarded as Raman characteristic peaks of anthocyanin in mulberry, to judge whether there was anthocyanin in mulberry, and the content of anthocyanin could be qualitatively determined as per the peak values. The spectroscopic data were preprocessed with the multiplicative scatter correction (MSC), baseline correction (airPLS), Normalized and the combined methods, and the best preprocessing method was selected by combining PLSR. It could be found that the best preprocessing method was airPLS+MSC+Normalized, and the PLSR model had a better effect. In the modeling set, the coefficient of determination is 0.97 and $RMSE_c$ is 2.74, while in the prediction set, the coefficient of determination is 0.82, and $RMSE_p$ is 13.69. Based on the spectra preprocessed with airPLS+MSC+Normalized, competitive adaptive reweighting sampling (CARS) was adopted to extract the characteristic wavelengths of the spectra. PLSR model and SVR model were established respectively regarding the selected wavelength variables as input variables, and the research into the predicting effects of both models was conducted. As per the results, the two models processed with CARS could predict the content of anthocyanin accurately, and the SVR model established with the screening of CARS variables had the best performance in the prediction accuracy, with the coefficient of the determination being 0.98 and $RMSE_c$ being 1.92 in the modeling set, and the coefficient of the determination being 0.94 and $RMSE_p$ being 4.70 in the prediction set. Therefore, the rapid and accurate prediction of anthocyanin content in mulberry could be achieved by Raman spectroscopy.

Keywords Raman spectroscopy; Anthocyanin; Mulberry; Feature extraction; Partial least squares regression; Support Vector Regression

(Received Nov. 2, 2020; accepted Feb. 14, 2021)

* Corresponding author