

基于近红外光谱的人参与西洋参的快速鉴别研究

黄亚伟¹, 王加华², 李晓云¹, Jacqueline J Shan³, Lei Ling³, 韩东海^{1*}

1. 中国农业大学食品科学与营养工程学院, 北京 100083

2. 许昌学院化学化工学院, 河南 许昌 461000

3. Afexa Life Sciences Inc. Edmonton, T6N 1G1, Canada

摘要 基于近红外光谱分析技术结合模式判别方法建立了一种人参与西洋参鉴别的新方法。收集根状、根须和粉末状的样品共 90 份, 在有聚乙烯包装袋的情况下直接采集近红外光谱, 去除原始光谱中包装袋的显著吸收后进行了 MSC 与一阶导数处理, 然后采用移动窗口偏最小二乘法选择了建模光谱区间, 分别建立了 PLS-DA, PCA-DA 和 SVM 判别模型, 并对 3 种模型作了对比分析, 结果表明 SVM 判别效果最优, 其对预测集的正确判别率为 100%。该方法准确、便捷, 可实际应用于企业原料药材的质量控制, 实现对原料药材的快速筛查。

关键词 近红外光谱; 人参; 西洋参; 移动窗口偏最小二乘法; 模式判别

中图分类号: O657.3, R282.5 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2010)11-2954-04

引言

人参和西洋参为常用名贵药材, 二者组织结构及化学成分十分相似, 肉眼很难分辨。由于来源、价格等因素, 以人参伪充西洋参的现象时有发生, 而且人参和西洋参的性味、功能主治、临床应用互不相同, 伪充产品甚至会造成医疗事故^[1], 因此快速有效的鉴别人参和西洋参就显得尤为重要。

目前的鉴别方法有感官鉴别、显微鉴别、薄层层析、高效液相色谱、红外光谱、同工酶谱及基因指纹等^[2-4]。然而这些方法一方面人为主观性影响比较大, 结果可靠性低, 一方面需要复杂的前处理, 操作起来繁琐、费时费力, 很难满足实际需要。

近红外光谱(NIRS)分析技术是 20 世纪 90 年代以来发展最快、最引人注目的光谱分析技术, 它具有快速、高效、制样简单、无污染等优点^[5]。由于其为全谱信息, 能同时反映物料的理化指标, 符合中医整体观的思想, 近年来在中药研究方面应用越来越多。在人参研究方面也有应用, 如 Woo 等^[6]对高丽参和人参的道地性进行了探讨研究; 杨海雷等^[7]对红参药材进行了鉴别, 取得较好的效果; 何淑华等^[8]成功对吉林人参进行了分类。然而他们都对样品进行了粉碎过筛处理, 破坏了样品, 应用起来也有些不便。目前关于人参和

西洋参的近红外鉴别研究还鲜为报道。

本研究收集根状、根须和粉末状的人参和西洋参样品在有聚乙烯包装袋的情况下直接采集近红外光谱, 结合模式判别的方法, 对不同状态的人参和西洋参进行了判别分析, 以期建立一种快速无损有效鉴别人参和西洋参的新方法。

1 材料与方法

1.1 实验材料

实验所用的 90 个样品由加拿大 Afexa Life Sciences Inc. 公司提供并鉴定, 其中加拿大主产区的西洋参样品 54 个, 中国主产区的人参样品 36 个, 样品状态包含根、根须和粉末, 其具体信息见表 1。将样品分为校正集和预测集, 60 个样品(34 个西洋参, 26 个人参)用于建模, 30 个样品(20 个西洋参, 10 个人参)用于模型预测。

Table 1 Statistics of *Panax ginseng* and *Panax quinquefolium* samples

样品状态	西洋参			人参		
	根	根须	粉末	根	根须	粉末
样品个数	21	16	17	11	10	15

收稿日期: 2009-12-09, 修订日期: 2010-03-12

基金项目: 中加合作研究项目(基于近红外光谱的西洋参质量评价研究)资助

作者简介: 黄亚伟, 1980 年生, 中国农业大学食品科学与营养工程学院博士研究生

e-mail: ywei0371@163.com

* 通讯联系人 e-mail: handh@cau.edu.cn

1.2 光谱采集

使用 ANTARIS (Thermo Nicolet, USA) 傅里叶变换近红外光谱仪, 积分球漫反射方式进行光谱采集, 仪器配备 In-GaAs 检测器及 Result 光谱采集软件。

光谱采集范围为 $4\ 000 \sim 10\ 000\ \text{cm}^{-1}$, 分辨率为 $8\ \text{cm}^{-1}$, 扫描次数为 32, $2\times$ 增益。将袋装样品直接置于积分球附件上, 使样品有一定的厚度确保没有光直接透过, 每个样品分别在 3 个不同部位采集光谱后取其平均光谱作为该样品的原始光谱。90 个样品的原始近红外谱图如图 1。

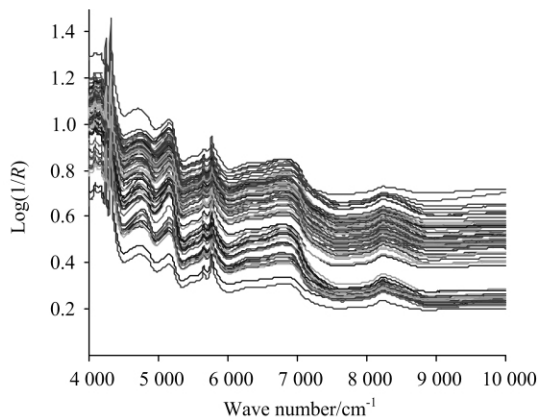


Fig 1 Raw spectra of Ginseng samples

1.3 光谱处理与数据分析

近红外光谱仪所采集的光谱除样品的自身信息外, 还包含了其他无关信息和噪音, 如电噪音、样品背景和杂散光等, 因此在建立模型之前需要对光谱进行预处理消除这些因素的影响^[9]。多元散射校正(MSC)和导数光谱是常用的预处理方法, MSC 主要是消除颗粒分布不均匀及颗粒大小产生的散射影响; 导数光谱可有效地消除基线和其他背景的干扰, 分辨重叠峰, 提高分辨率和灵敏度。

建模波长选择是提取光谱信息的有效途径, 它可以简化模型, 剔除不相关或非线性变量, 提高模型的预测能力和稳健性。本研究采用移动窗口偏最小二乘法(MWPLS)进行建模有效波长选择。

实验数据利用 TQ Analyst V7.2 (Thermo Nicolet, USA) 和 MATLAB7.0 (The Mathworks Inc., Natick, MA) 软件进行分析和处理。

2 结果与讨论

2.1 谱图分析

由图 1 可以看出, 原始光谱分为上下两部分, 上部分为根状和根须状的人参和西洋参样品, 下部分为粉末状的人参和西洋参样品, 物理状态的不同导致基线有一定的偏移。但是所有样品的原始近红外谱图形状极为相似, 虽然西洋参与人参的物理状态和化学组成有很大差异, 也无法直接从谱图上找出特定的吸收峰加以区分。

对比分析图 2 所示的三种类型光谱可以得知, $4\ 300$ 和 $5\ 800\ \text{cm}^{-1}$ 附近主要为聚乙烯包装袋的 C—H 基团的倍频与

合频吸收, 为了避免包装袋的影响, 选取包装袋吸收极其微弱的 $4\ 500 \sim 5\ 500\ \text{cm}^{-1}$ 与 $6\ 000 \sim 10\ 000\ \text{cm}^{-1}$ 处的光谱作为研究用数据。然后对选取的光谱分别进行 MSC 和一阶导数预处理, 经过预处理后的光谱用于下一步的数据分析。

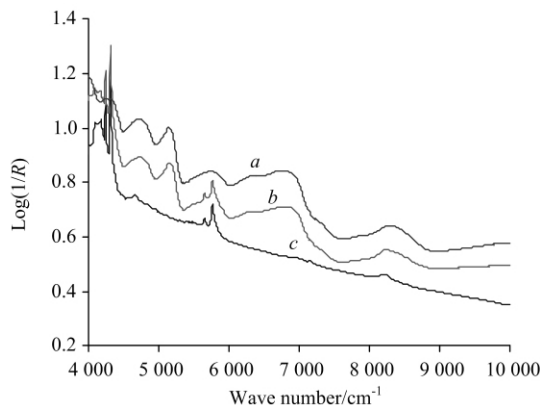


Fig 2 Analysis of spectral characteristics of polyethylene packing film

a: Sample without package; b: Sample with package; c: packing film

2.2 建模波长选择

移动窗口偏最小二乘法(MWPLS)是 Jiang 等^[10]提出的一种建模波长区间优选方法。对于有 n 个数据点的光谱, 该方法构造一个从第 u 光谱点到第 $(u+w-1)$ 光谱点的窗口, 其窗口大小为 w 。窗口从第 u 个光谱点开始移过整个光谱区间, 整条光谱上有 $n-w+1$ 个窗口, 在每个窗口下使用不同因子数构建 PLS 模型, 并计算该窗口的残差平方和 (sums of squared residues, SSR), 在这里将获得一系列残差线, 每条线包含某一因子下的 SSR 值和对应的窗口位置。在一系列残差线中, 选择最小 SSR 值, 其对应的窗口位置即为所选择信息区间。

窗口大小 w 的选择很重要, 较小的窗口不能完全涵盖有效的信息, 较大的窗口则会带来冗余的数据。本文将建模集样品实际类别设为化学值向量, 即 1 代表人参, -1 代表西洋参, 窗口大小设定为 $20 \sim 180$, 间隔为 10, 最大因子数为 10, 在光谱区间 $4\ 500 \sim 5\ 500\ \text{cm}^{-1}$ (a) 与 $6\ 000 \sim 10\ 000\ \text{cm}^{-1}$ (b) 分别进行 WMPLS 区间优选, 作出窗口大小与之对应的 SSR 的曲线, 依据该曲线确定最佳窗口大小。如图 3 所示, 在光谱 a 和 b 区间内窗口大小分别为 100 和 140 时, SSR 值最小, 因此确定 WMPLS 在 a 和 b 两段光谱的窗口大小分别为 100 和 140。

图 4 为 WMPLS 光谱区间优选图, 图 4(a) 和 (b) 中分别有 10 条残差线, 每条线包含两个信息, 即某一个因子数及在该因子数下窗口所在位置区间 PLS 模型的残差和。从图 4(a) 可以得出, 当窗口位于 $5\ 118 \sim 5\ 500\ \text{cm}^{-1}$ 时, PLS 模型获得最小残差值; 图 4(b) 中, 当窗口位于 $6\ 665 \sim 7\ 201\ \text{cm}^{-1}$ 时, PLS 模型残差值最小, 而且这两处主要为 C—O, C—N 和 N—H 键的倍频与合频吸收, 从化学组成上看主要为人参和西洋参的特征成分吸收, 因此把光谱区间 $5\ 118 \sim 5\ 500\ \text{cm}^{-1}$ 和 $6\ 665 \sim 7\ 201\ \text{cm}^{-1}$ 作为模式判别的输入数据。

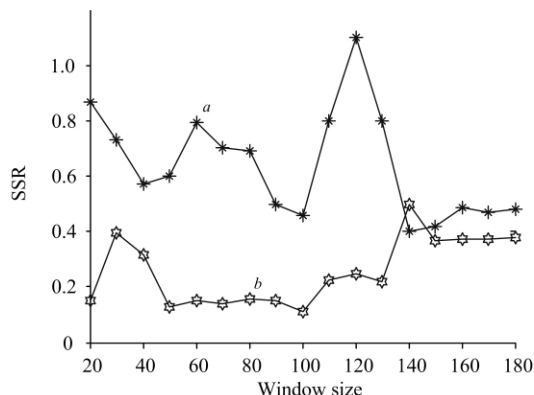


Fig 3 Relationships of window size to its lowest SSR in the range of *a* and *b*

a: 4 500~5 500 cm^{-1} ; *b*: 6 000~10 000 cm^{-1}

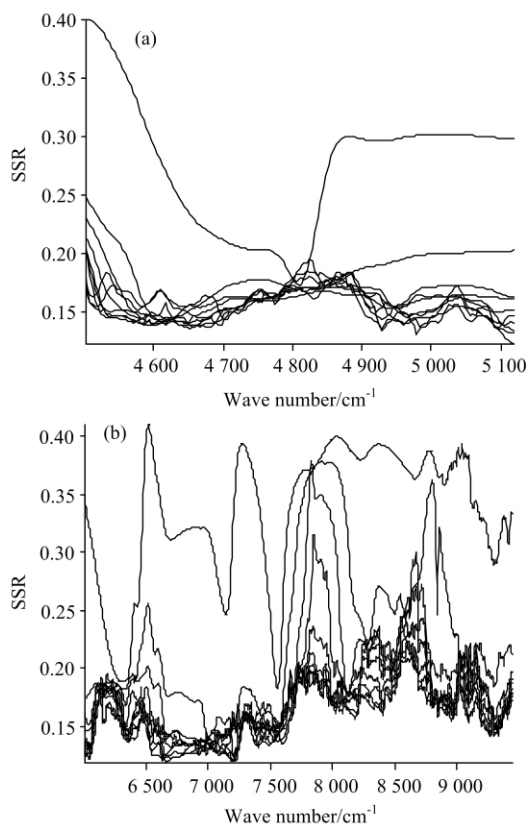


Fig 4 Informative regions chosen by WMPPLS in range of 4 500~5 500 cm^{-1} (a) and 6 000~10 000 cm^{-1} (b)

2.3 模式判别分析

2.3.1 PLS-DA 判别分析

偏最小二乘判别分析 (partial least squares discriminant analysis, PLS-DA) 法是对不同种类样品进行数值变量标定, 采用偏最小二乘法 (PLS) 结合近红外光谱特征建立校正模型, 对样品进行定性判别。

本研究利用 TQ Analyst V7.2 软件分别对所选波段进行模型构建, 采用交互检验预测残差平方和 (PRESS) 确定最佳主因子数为 5。所建模型对预测集的判别结果如表 2, 人参的正确判别率为 100%, 西洋参有 2 个被误判, 其正确判别率

为 90%; 西洋参和人参的总判别率为 93.3%。

2.3.2 PCA-DA 判别分析

主成分判别分析 (principal component analysis discriminant analysis, PCA-DA) 法首先对 NIR 光谱矩阵进行主成分降维处理, 再利用其主成分得分来计算马氏距离进行判别。主成分分析是多元统计中的一种数据挖掘技术。在不丢失主要光谱信息的前提下选择为数较少的新变量来代替原来较多的变量, 解决了由于谱带的重叠而无法分析的困难^[11], 这样既能利用主成分分析消除大量重叠的冗余信息, 最大限度的保留有效数据, 又能利用马氏距离进行定量判别。

Table 2 Accurate identification rate for validation set by different analysis methods

判别方法	建模波段 / cm^{-1}	人参判别率/%	西洋参判别率/%	总判别率/%
PLS-DA	5 118~5 500, 6 665~7 201	100	90	93.3
PCA-DA	5 118~5 500, 6 665~7 201	100	95	96.7
SVM	5 118~5 500, 6 665~7 201	100	100	100

本研究采用 TQ Analyst V7.2 数据处理软件对上述波段进行主成分判别分析, 对预测集的判别结果如表 2, 人参的正确判别率为 100%; 西洋参有 1 个被误判, 其正确判别率为 95%; 西洋参和人参的总判别率为 96.7%, PCA-DA 判别方法略优于 PLS-DA 判别方法。

2.3.3 支持向量机判别分析

支持向量机 (support vector machine, SVM), 是近几年产生的机器学习算法, 其核心思想就是把数据非线性映射到高维特征空间, 在高维空间构造具有低维最优分类超平面使分类风险上界最小, 其优点在于同时兼顾了经验风险和置信范围这两个最小化风险的泛函问题, 能较好地解决非线性和高维数等实际问题, 训练得到的结果通常也是全局最优解, 其在分类、非线性模型识别方面得到了广泛的应用^[12,13]。

本研究在 MATLAB7.0 中对所优选光谱区域进行主成分分析, 提取累计贡献率为 98.7% 的前 10 个主成分作为 SVM 的输入变量, 选取径向基核函数, 通过多次试验、逐步搜寻的方式确定最佳惩罚系数 C 和径向基系数 γ 分别为 1 和 2, 然后进行模型训练, 训练后对预测集的判别结果如表 2, 人参和西洋参判别率均为 100%。结果表明 SVM 对两类人参的判别效果最好。

对比分析 3 种判别方法, 结果显示尽管选用相同的光谱数据, 其判别效果差异也较大。其中 PLS-DA 和 PCA-DA 为较常用的判别方法, 一般情况下能取得较好的效果, 但是对于非线性数据不一定能得到最优解, 因此针对特定研究选用合适的建模方法则尤为重要。

3 结论

本研究基于近红外光谱分析技术结合模式判别方法建立

了一种人参和西洋参的快速无损鉴别的新方法。该研究直接采集聚乙烯袋装的各种状态人参和西洋参的近红外光谱,探讨了包装袋的光谱吸收特性并在原始光谱中将其显著吸收峰给予去除,避免了包装袋的影响。剩余的光谱数据经过一阶导数预处理后,运用 MWPLS 方法提取有效建模信息区间,分别建立了 PLS-DA, PCA-DA 型和 SVM 判别模型,3 种模

型的预测总判别率分别为 93.3%, 96.7% 和 100%。结果表明 SVM 判别方法效果最幻。本研究收集的样品较多,而且样品状态广泛,不需经过任何预处理,该方法具有客观、准确、快速、成本低、无污染等优点,可实际应用于企业原料药材的质量控制中。

参 考 文 献

- [1] MA Hai-qin, WANG Zhi-qing, YANG Wei(马海琴, 王志清, 杨 微). Special Wild Economic Animal and Plant Research(特产研究), 2005, 27(3): 58.
- [2] LIU Min, DANG Bo-jun, ZHOU Guo-hua(刘 敏, 党博军, 周国华). Pharmaceutical and Clinical Research(药学与临床研究), 2008, 16(5): 373.
- [3] CONG Deng-li(丛登立). Ginseng Research(人参研究), 2005, 17(1): 4.
- [4] Ngan F, Shaw P, But P, et al. Phytochemistry, 1999, 50: 787.
- [5] LU Wan-zhen(陆婉珍). Modern Near Infrared Spectroscopy Analytical Technology(现代近红外光谱分析技术). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2007.
- [6] Woo Y A, Kim H J, Cho CH, et al. Microchemical Journal, 2002, 73(3): 299.
- [7] YANG Hai-lei, LIU Xue-song, CHENG Yi-yu, et al(杨海雷, 刘雪松, 程翼宇, 等). Chinese Traditional and Herbal Drugs(中草药), 2005, 36(6): 915.
- [8] HE Shu-hua, SUN Rui-yan, et al(何淑华, 孙瑞岩, 等). Acta Scientiarum Naturalium Universitatis Jilinensis(吉林大学自然科学学报), 2001, (1): 96.
- [9] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Progress in Chemistry(化学进展), 2007, 19(16): 528.
- [10] Jiang J H, Berry R J, Ozaki Y, et al. Analytical Chemistry, 2002, 74: 3555.
- [11] HE Yong, LI Xiao-li(何 勇, 李晓丽). J. Infrared Millim. Waves(红外与毫米波学报), 2006, 25(3): 192.
- [12] Olivier Devos, Cyril Ruckebusch, Alexandra Durand, et al. Chemometrics and Intelligent Laboratory Systems, 2009, 96: 27.
- [13] YU Ke, CHENG Yi-yu(虞 科, 程翼宇). Chinese Journal of Analytical Chemistry(分析化学), 2006, 34(4): 561.

Research on Fast Discrimination between *Panax Ginseng* and *Panax Quinquefolium* Based on Near Infrared Spectroscopy

HUANG Ya-wei¹, WANG Jia-hua², LI Xiao-yun¹, Jacqueline J Shan³, Lei Ling³, HAN Dong-hai^{1*}

1. College of Food Science & Nutritional Engineering, China Agricultural University, Beijing 100083, China

2. College of Chemistry and Chemical Engineering, Xuchang University, Xuchang 461000, China

3. Afexa Life Sciences Inc. Edmonton, T6N 1G1, Canada

Abstract Near infrared spectroscopy combined with pattern recognition techniques were applied to develop a method of fast and nondestructive discrimination between Chinese ginseng and American ginseng. A total of 90 representative ginseng samples including root, fiber and powder were collected. NIR spectra of the samples were obtained directly with wrapped polyethylene packing film. MSC and first derivative were performed after the elimination of notable packing film absorbance in raw spectra. Then the informative wave bands were chosen by moving window partial least-squares regression method. PLS-DA, PCA-DA and SVM discrimination models were founded and their results were compared. SVM was proven to be the most effective method with 100% accurate identification rate for validation set. It indicates that the method founded is precise and convenient and can be practically used in practice for quality control and fast screening of raw herb materials.

Keywords Near infrared spectroscopy; *Panax ginseng*; *Panax quinquefolium*; Moving window partial least squares; Pattern recognition

* Corresponding author

(Received Dec. 9, 2009; accepted Mar. 12, 2010)