# 基于红外光谱对野生冬虫夏草不同部位的识别

陈 逃[1]，郭 慧[1]，袁 满[1]，谭福元[3*]，李益洲[2*]，李梦龙[1]

1. 四川大学化学学院，四川 成都 610064
2. 四川大学网络空间安全学院，四川 成都 610064
3. 成都图径生物科技有限公司，四川 成都 610093

**摘 要** 冬虫夏草作为著名的传统中药材，由于其良好的药用价值而备受青睐。目前多数工作集中研究其活性成分含量以及药理药效。而对其不同部位的识别研究较为匮乏。基于红外光谱数据，结合化学计量学对多维度复杂体系的解析优势对冬虫夏草不同部位进行分类识别。首先对野生冬虫夏草五个不同部位包括子座头、子座中、头部、虫体中段、虫体尾段总共808个光谱数据使用标准正态变换(SNV)、多元散射矫正(MSC)进行数据预处理。而后用竞争自适应再权重取样(CARS)、变量组合种群分析(VCPA)挑选具有代表意义的特征变量。最后使用偏最小二乘判别分析(PLS-DA)、线性判别分析(LDA)进行建模预测分析。模型对训练集使用十倍交叉验证，以准确率(Acc)作为评价指标。结果表明，在该数据上PLS-DA模型在10倍交叉验证和独立测试集上的预测准确率分别是90.1%和92.0%，而使用LDA模型时，预测准确率分别降低到86.7%和85.8%。采用CARS和VCPA特征挑选方法可有效将特征从3 601维分别降到699和420维，同时保持预测准确率与全部特征的预测准确率相当。而挑选的特征波数630，625，1 024，1 028，1 084和1 089 $cm^{-1}$与虫草的甘露醇相关，879和874 $cm^{-1}$与虫草的多糖相关。通过对挑选的波数进行Wilcoxon rank-sum检验进一步表明虫草五个部位之间存在显著差异。研究表明化学计量学方法结合红外光谱能够有效识别冬虫夏草不同部位，有助于在分子层面上加深对冬虫夏草形成的认识，为针对虫草不同部位高效利用提供参考。

**关键词** 冬虫夏草；红外光谱；化学计量学；分类；特征选择

**中图分类号**：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)12-3727-06

## 引 言

冬虫夏草是菌丝体毛虫和真菌性基质芽的寄生复合物[1]，因为其出色的保护和免疫调节作用，成为备受推崇的传统中药材。冬虫夏草具有多种有效成分，包括多糖、虫草、腺苷、甘露醇、固醇、甘露聚糖和核苷[1]等。各种分析方法已经被应用到冬虫夏草活性成分的研究[2]。Li等[3]采用毛细管电泳测定冬虫夏草三种主要核苷的含量来探究与药理作用相关的成分。Yang等[4]优化毛细管电泳质谱法(CE-MS)同时测定天然虫草和人工虫草中的核苷和核苷碱基。Zhao等[5]结合亲水相互作用色谱(HILIC)和电喷雾电离质谱(ESI-MS)来表征和定量天然虫草。Hu等[6]使用高效液相色谱-串联质谱法(HPLC-MS/MS)在冬虫夏草中检测到有效的化学标记。凭借指纹分析功能，近红外光谱技术(NIR)也以其快速，低成本和无损检测等优势，广泛用于食品和药物的定性和定量分析[7]。Xie等[8]使用傅里叶变换近红外光谱(FT-NIR)定量测定冬虫夏草菌丝体中精氨酸的含量，并通过特征选择算法获得了预测精氨酸含量的最佳波数。而红外光谱在野生冬虫夏草不同部位差异性研究鲜有报道。

红外光谱信号通常会受到干扰，因此需要进行预处理提高光谱数据质量以便后续研究。标准正态变化(standard normal variation，SNV)[9]和多元散射校正(multiplicative scatter correction，MSC)[10]已广泛用于光谱数据的预处理。此外，通过变量选择消除无关冗余信息，降低模型复杂度并提高模型稳定性。

基于野生冬虫夏草不同部位的红外数据探讨了采用不同预处理SNV和MSC、特征挑选竞争自适应再权重取样

**收稿日期**：2020-11-17，**修订日期**：2021-03-08

**基金项目**：国家自然科学基金项目(21775107，21675114)资助

**作者简介**：陈 逃，1992年生，四川大学化学学院硕士研究生 e-mail：740270369@qq.com

*通讯作者 e-mail：liyizhou@scu.edu.cn；tanfuyuan@verygrass.com

(competitive adaptive reweighted sampling, CARS)[11]和变量组合种群分析(variable combination population analysis, VCPA)[12]、预测模型偏最小二乘判别分析(partial least squares discriminant analysis, PLS-DA)[13]和线性判别分析(linear discriminant analysis, LDA)[14]分别构建虫草部位的识别模型，并比较各方法的效果和以及对筛选的特征波长进行分析，有助于在分子层面上加深对野生冬虫夏草形成的认识，可为后期药物开发高效利用野生虫草提供参考。

## 1 实验部分

### 1.1 数据源

用于实验的冬虫夏草包括子座头、子座中、头部、虫体中段、虫体尾段总共 808 个样本，均由成都图径生物科技有限公司提供，样本详细信息如表 1 所示。所有样本采用美国 PerkinElmer 公司生产的 Spectrum 100 型傅里叶变换红外光谱仪，扫描范围为 400～4 000 $cm^{-1}$。训练集和测试集随机按 4∶1 生成，训练集使用十倍交叉验证，准确率(accuracy, Acc)作为评价指标。

**表 1 样本信息**

**Table 1 General information of samples**

| Organs | HS | MS | HD | ML | EL |
|---|---|---|---|---|---|
| Number of sample | 161 | 162 | 162 | 162 | 161 |

注：HS：子座头；MS：子座中；HD：头部；ML：虫体中段；EL：虫体尾段，下同

### 1.2 数据预处理

红外光谱在测量时，会受到背景噪声和散射因素影响，因此对光谱进行预处理，可以提高后续光谱数据分析的可靠性。本研究使用标准正态变换(SNV)消除基线变化所引起的潜在影响、使用多元散射校正(MSC)消除散射效应，增强红外吸收光谱信息。

### 1.3 变量挑选

CARS[11]首先采用蒙特卡洛(Monte Carlo)策略将样本数据集用于构建 PLS 模型，基于模型的系数来估计波长贡献。然后采用指数递减函数(exponentially decreasing function, EDF)除去系数绝对值小的波数。最后保留具有较大绝对值系数的波数作为特征选择结果。

VCPA[12]也常用于光谱数据变量选择。首先，使用二进制矩阵采样(binary matrix sampling, BMS)方法生成具有多样性变量组合子集。其次采用模型总体分析(model population analysis, MPA)和训练集交互验证均方根误差(root-mean squared error of cross-validation, RMSECV)评估子模型。然后根据指数递减函数(EDF)去除 PLS 模型系数绝对值较小波长。最后，具有最低 RMSECV 值的子集将作为最终变量选择结果。

### 1.4 建模预测

LDA 基本思想是在一定训练样本上设法将样本特征投影到子空间，使得同类样本投影点互相聚集，不同类样本投影点互相远离，这样相同类别之间距离最小，对于新样本进行分类时，投影到同一子空间，根据投影位置和距离确定新样本类别。

偏最小二乘判别分析(PLS-DA)是一种监督分类方法，根据偏最小二乘回归(PLSR)算法开发而来。PLS-DA 算法集主成分分析、多元线性回归和相关性分析等优点于一身，可以将特征变量和目标通过映射变换最终建立类别与光谱矩阵的判别关系。

## 2 结果与讨论

### 2.1 不同部位的红外光谱比较

冬虫夏草不同部位平均红外光谱图如图 1(a)所示，可看出部位间存在较大差异，但通过肉眼无法区分。冬虫夏草不同部位间皮尔森相关系数计算如图 1(b)所示，可看出不同部位之间有很强的相关性，但不完全相同，因此借助化学计量学方法进行识别。

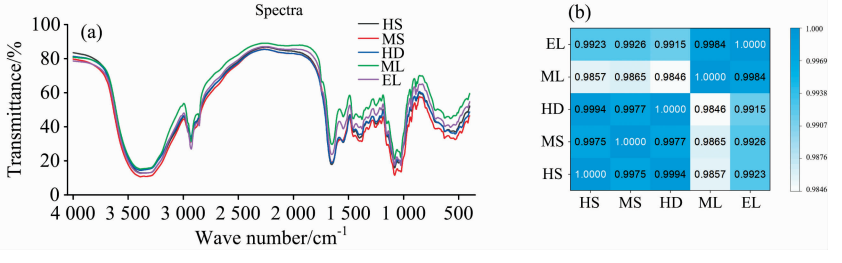


**图 1 冬虫夏草不同部位均值红外光谱(a)与相似性(b)**

**Fig. 1 The averaged Fourier-transform infrared spectra for different parts of Cordyceps (a); The similarities between each two parts of Cordyceps (b)**

### 2.2 不同部位的分类结果

表 2 可以看出，PLS-DA 经过 CARS 和 VCPA 特征挑选之后，特征维数大幅下降，分别从 3 601 降到 669 和 420，而且准确率 90.1%，91.4%与全部特征预测准确率 92.0%相当。而 LDA 结果相对较差，模型最高预测准确率为 85.8%，经特征挑选后准确率分别为 80.9%和 82.1%。结果表明 PLS-DA 预测效果优于 LDA，特征挑选有利于降低模型复杂程度。

针对不同特征挑选、建模方法所得独立测试集预测结果进一步用混淆矩阵分析如图 2 所示。结果表明，大多数错误

**表 2　不同部位的分类结果**

**Table 2　The model performance on discriminating different cordyceps parts**

| Pre-treatment methods | Methods | 10fold-CV/% | Independent test set/% | Variable number |
|---|---|---|---|---|
| SNV+MSC | PLS-DA | 90.1 | 92.0 | 3 601 |
| SNV+MSC | LDA | 86.7 | 85.8 | 3 601 |
| SNV+MSC | CARS-PLS-DA | 90.0 | 90.1 | 669 |
| SNV+MSC | CARS-LDA | 76.3 | 80.9 | 669 |
| SNV+MSC | VCPA-PLS-DA | 89.6 | 91.4 | 420 |
| SNV+MSC | VCPA-LDA | 83.0 | 82.1 | 420 |

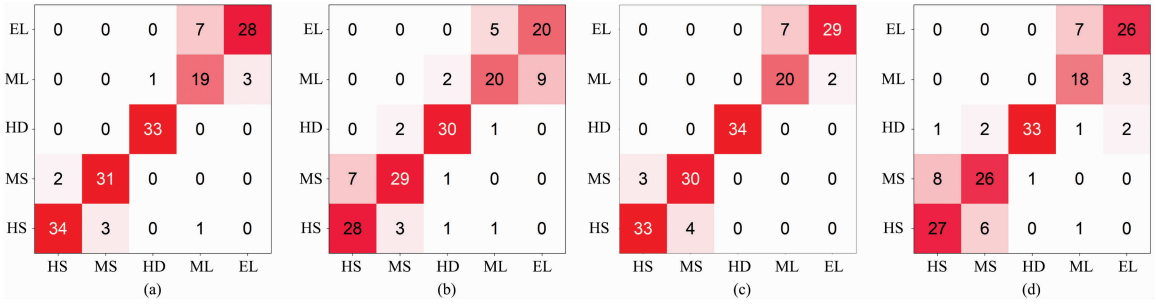


**图 2　不同方法独立测试集的混淆矩阵**

(a)：CARS-PLS-DA；(b)：CARS-LDA；(c)：VCPA-PLS-DA；(d)：VCPA-LDA

**Fig. 2　The confusion matrix of independent data set by different methods**

(a)：CARS-PLS-DA；(b)：CARS-LDA；(c)：VCPA-PLS-DA；(d)：VCPA-LDA

预测情况都出现在靠近对角线附近区域，表明该样本被预测为临近部位。

### 2.3　变量分析

对 CARS 和 VCPA 挑选的变量分析发现有 85 个共享特征，对于特征挑选结果差异性应该来源于算法本身的差异。特征波数选择结果对应光谱图中位置如图 3(a,b)所示。

其中共同波数参照文献[15]报道见表 3 所示，如波数 630 与 625 $cm^{-1}$对应冬虫夏草活性成分甘露醇，说明特征挑选方法的特征波数具有一定化学意义，挑选特征具有可行性。

对 CARS 和 VCPA 挑选波数画出 box-plot 图和 Wilcoxon rank-sum 检验热图如图 4、图 5 所示。从图 4 当中可看出，子座中段 MS 与虫体中段 ML 的 $p$ 值最低，该数据说明冬虫夏草这两部位活性成分差异性最显著。如在图 4 波数 1 084 $cm^{-1}$(b)所示，结果显示在该波数下不同部位之间活性成分有显著性差异。类似情况在图 5 也可观察得到。结果表明，冬虫夏草不同部位之间活性成分有显著性差异。

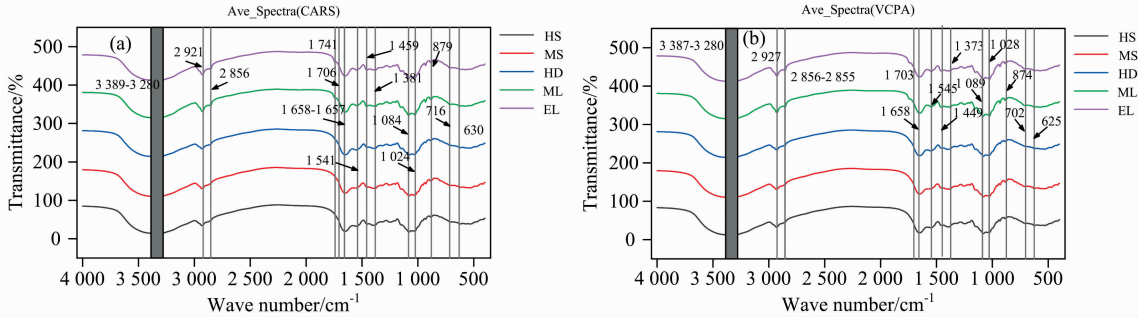


**图 3　CARS (a)和 VCPA (b)特征选择结果**

**Fig. 3　The result of feature selection method CARS (a) and VCPA (b)**

## 3　结　论

通过化学计量学模型结合红外光谱数据，实现对野生冬虫夏草不同部位有效识别。总体而言，PLS-DA 模型优于 LDA 模型，准确率在 90.0%以上，冬虫夏草不同部位在活性成分上确实存在较大的差异。特征挑选方法可以保证准确率的同时降低模型复杂程度，同时挑选的特征具有一定的化学可解释性，说明特征挑选的可行性。本研究可有助于在分子水平上加深对野生冬虫夏草形成的认识，并对后期药物开发高效利用虫草提供参考，为合理有效利用名贵中草药提供依据。

**表 3 虫草特征挑选与化学解释**

**Table 3 Holistic assignment of infrared spectroscopy spectra of Cordyceps**

| Base group and vibration mode | CARS /cm$^{-1}$ | VCPA /cm$^{-1}$ | Base group and vibration mode | CARS /cm$^{-1}$ | VCPA /cm$^{-1}$ |
|---|---|---|---|---|---|
| | 3 389 | | $\nu_{as}$(C—H) in $CH_3$, $CH_2$ | 2 921 | 2 927 |
| | 3 388 | | $\nu_s$(C—H) in $CH_3$, $CH_2$ | 2 853 | 2 856 |
| | 3 347 | 3 387 | | | 2 855 |
| | 3 345 | 3 372 | $\nu$(C═O) in ester | 1 741 | |
| | 3 343 | 3 360 | $\nu$(C═O) in Carboxy acid | 1 706 | 1 703 |
| | 3 338 | 3 358 | $\nu$(CO) in Amide Ⅰ | 1 658 | 1 658 |
| | 3 329 | 3 345 | | 1 657 | |
| $\nu$(O—H, N—H) in OH, NH | 3 328 | 3 313 | $\delta$(N—H) in Amide Ⅱ | 1 541 | 1 545 |
| | 3 325 | 3 306 | $\delta$(C—H) in $CH_3$ | 1 459 | 1 449 |
| | 3 321 | 3 304 | $\delta_s$(C—H) in $CH_3$ | 1 381 | 1 373 |
| | 3 316 | 3 302 | $\nu$(C—OH) in Mannitol | 1 084 | 1 089 |
| | 3 311 | 3 296 | | 1 024 | 1 028 |
| | 3 296 | 3 285 | $\delta$(C—O—C) in Saccharides, ring | 879 | 874 |
| | 3 284 | 3 280 | breathing | | |
| | 3 282 | | $\nu$(C—C) in —$(CH_2)_4$— | 716 | 702 |
| | 3 280 | | $\delta$(C—O—H) in Mannitol, wagging | 630 | 625 |

Note: $\nu$: stretching or frame vibration; as: asymmetrical; s: symmetrical; $\delta$: bending vibration

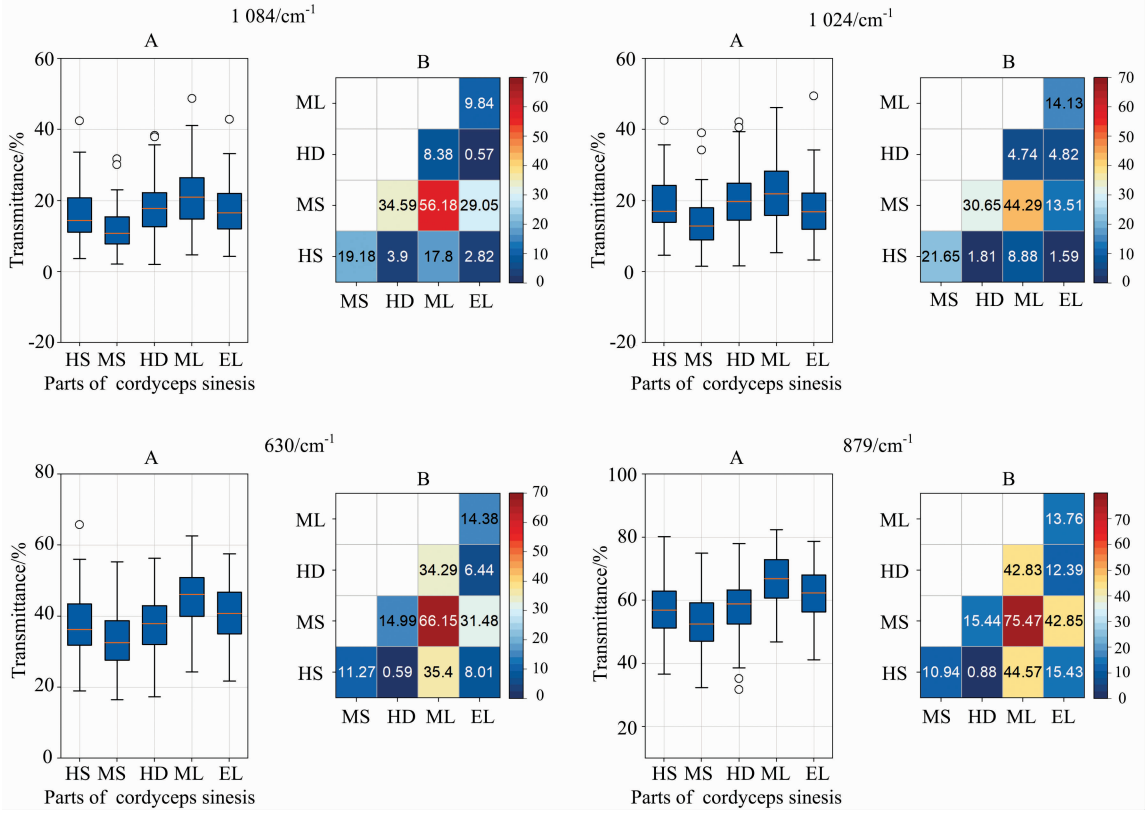


**图 4 CARS 选择波数对应箱线图 A(1 084, 1 024, 630, 879 cm$^{-1}$)和 Wilcoxon rank-sum 检验热图 B(1 084, 1 024, 630, 879 cm$^{-1}$)**

**Fig. 4 The box-plot A (1 084, 1 024, 630, 879 cm$^{-1}$) and heat-map for Wilcoxon rank-sum test of wavenumbers selectedby CARS B(1 084, 1 024, 630, 879 cm$^{-1}$)**

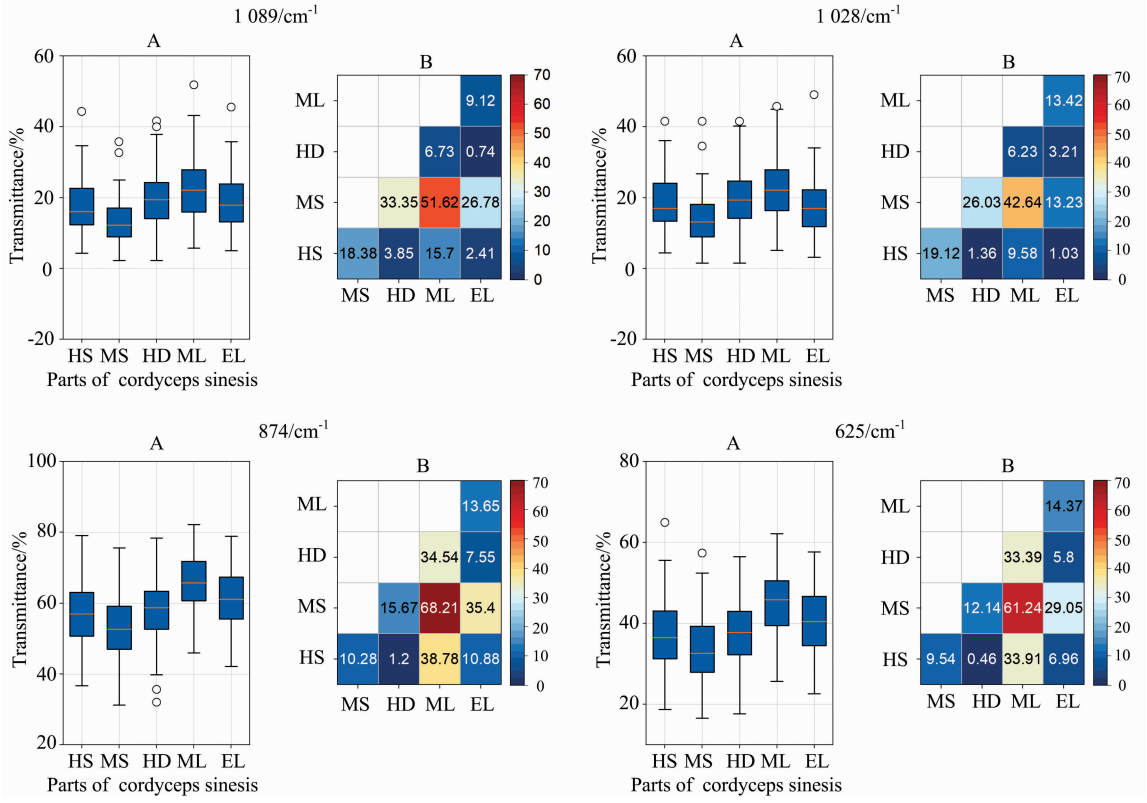


**图 5　VCPA 选择波数对应箱线图 A(1 089，1 028，874，625 $cm^{-1}$)和 Wilcoxon rank-sum 检验热图 B(1 089，1 028，874，625 $cm^{-1}$)**

**Fig. 5　The box-plot A(1 089，1 028，874，625 $cm^{-1}$) and heat-map for Wilcoxon rank-sum test of wavenumbers selected by VCPA B(1 089，1 028，874，625 $cm^{-1}$)**

## References


[1] Lo H C, Hsieh C, Lin F Y, et al. Journal of Traditional & Complementary Medicine, 2013, 3(1): 16.

[2] Li S P, Yang F Q, Tsim K W K. Journal of Pharmaceutical & Biomedical Analysis, 2006, 41 (5): 1571.

[3] Li S P, Li P, Dong T T X, et al. Electrophoresis, 2001, 22(1): 144.

[4] Yang F Q, Ge L, Yong J W H, et al. Journal of Pharmaceutical and Biomedical Analysis, 2009, 50(3): 307.

[5] Zhao H Q, Wang X, Li H M, et al. Molecules, 2013, 18: 9788.

[6] Hu H, Xiao L, Zheng B, et al. Analytical and Bioanalytical Chemistry, 2015, 407(26): 8059.

[7] Wang P, Zhang H, Yang H, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 137: 1403.

[8] Xie C, Xu N, Shao Y, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 149: 971.

[9] Chen H, Lin Z, Tan C. Journal of Pharmaceutical and Biomedical Analysis, 2018, 161: 239.

[10] Yu Y, Yu H, Guo L, et al. Analytical Methods, 2018, 10(26): 3224.

[11] Xu S, Zhao Y, Wang M, et al. CATENA, 2017, 157: 12.

[12] Yun Y H, Wang W T, Deng B C, et al. Analytica Chimica Acta, 2015, 862: 14.

[13] Alladio E, Giacomelli L, Biosa G, et al. Forensic Science International, 2018, 282: 221.

[14] Szabó é, Gergely S, Salgó A. Journal of Chemometrics, 2018, 32(4): e3005. s

[15] Yang P, Song P, Sun S Q, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2009, 74(4): 983.

# Recognition of Different Parts of Wild Cordyceps Sinensis Based on Infrared Spectrum

CHEN Tao[1], GUO Hui[1], YUAN Man[1], TAN Fu-yuan[3*], LI Yi-zhou[2*], LI Meng-long[1]

1. College of Chemistry, Sichuan University, Chengdu 610064, China
2. School of Cyber Science and Engineering, Sichuan University, Chengdu 610064, China
3. Biological Process Science and Technology Co., Ltd., Chengdu 610093, China

**Abstract** Cordyceps Sinensis, a famous Chinese medicinal material, is favored due to its good medicinal value. Recently, investigations have focused on the study of its active ingredient content and pharmacological effects. However, scarce studies were reported on the identification of different parts of wild Cordyceps. This study is based on infrared spectroscopy data, combined with the analytical preponderance of chemometrics in multi-dimensional complex systems to classify and identify different parts of Cordyceps Sinensis. First, preprocessing methods, standard normal variation (SNV) and multiplicative scatter correction (MSC) were used on a total of 808 spectral data of five different parts of wild Cordyceps, including head of stroma (HS), middle of stroma (MS), head (HD), the middle larva body (ML) and the end larva body (EL). Then, competitive adaptive reweighted sampling (CARS) and variable combination population analysis (VCPA) were hired to select characteristic variables with representative significance. Ultimately, partial least squares discriminant analysis (PLS-DA) and linear discriminant analysis (LDA) were engaged for modeling and predictive analysis. Ten-fold cross-validation was used on the training set, and accuracy (Acc) was employedas the evaluation index. The results showed that the prediction accuracies of the PLS-DA model on the 10-fold cross-validation and independent test set on this data were 90.1% and 92.0%, respectively, while using the LDA model, the prediction accuracies reduced to 86.7% and 85.8%, respectively. In addition, the dimensions of the features can be effectively reduced from 3 601 to 669 and 420, respectively, when using CARS and VCPA feature selection methods, but keeping the prediction accuracies equivalent to that of all features. The selected wavenumbers 630, 625, 1 024, 1 028, 1 084, and 1 089 $cm^{-1}$ were related to mannitol in cordyceps, and 879 and 874 $cm^{-1}$ were related polysaccharides in cordyceps. The Wilcoxon rank-sum test on the selected wavenumbers further showed significant differences between the five parts of Cordyceps. This study showed that chemometric methods combined with infrared spectroscopy could effectively identify different parts of Cordyceps Sinensis, thereby deepening the understanding of the formation of Cordyceps at the molecular level and providing a reference for the efficient use of different parts of Cordyceps.

**Keywords** Cordyceps sinensis; Infrared spectroscopy; Chemometrics; Classification; Feature selection

(Received Nov. 17, 2020; accepted Mar. 8, 2021)

* Corresponding authors