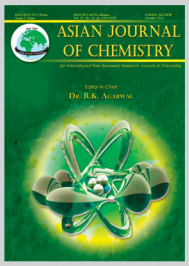
Asian Journal of Chemistry; Vol. 24, No. 3 (2012), 1019-1022

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in



Determination of *Rhizoma curcumaes* Using Visible and Near-Infrared Spectroscopy

XIN-XIANG LEI[1,2,*], XIAO-JING CHEN[3], LU LIU[2], AN-JIANG ZHANG[1] and LI-SHENG DING[1]

[1]Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, P.R. China
[2]College of Chemistry and Materials Engineering, Wenzhou University, Wenzhou 325035, P.R. China
[3]College of Physical and Electronic Informatiom, Wenzhou University, Wenzhou 325035, P.R. China

*Corresponding author: Fax: +86 577 86689179; Tel: +86 577 86689197; E-mail: xinxianglei@gmail.com

(*Received*: 24 January 2011; *Accepted*: 1 November 2011) AJC-10584

This study investigated the capacity of visible and near infrared (VIS-NIR) spectroscopy with chemometrics for the fast species discrimination of *Rhizoma curcumaes*. The powder samples of three *Curcuma* species *i.e.*, *C. phaeocaulis*, *C. kwangsiensis* and *C. wenyujin*, from five regions in China were used for the analysis. Least squares support vector machine (LS-SVM) was used to establish the discrimination model. Multiplicative scatter correction (MSC) was chosen as the best spectral pre-processing algorithm. Successive projections algorithm (SPA) was operated for the wavelength variable selection from thousands of original full-spectrum (FS) variables. The best correct classification rate of 99.11 % was obtained by MSC-SPA-LS-SVM model with only eight variables. The overall results demonstrate that VIS-NIR spectroscopy has the ability for the fast discrimination of different species of *Curcuma* species.

Key Words: Visible and near infrared spectroscopy, *Rhizoma curcumaes*, Successive projections algorithm.

INTRODUCTION

Rhizoma Curcumae (rhizomes of *Curcuma* species; Ezhu) is a common traditional Chinese medicine that has been used for more than a thousand years[1]. Genus *Curcuma* belongs to the Zingiberaceae family. There are about 20 species distributed in China and some of which are being used as herbal medicine. Generally, the rhizomes of *C. wenyujin*, *C. kwangsiensis* and *C. phaeocaulis* are used as Ezhu or Jianghuang according to Chinese Pharmacopoeia (2005 edition)[2]. However, it is difficult to distinguish their origins of raw materials in clinic because of their similar morphological characters, though their pharmacological activities and chemical components are obviously different[3-5]. Traditionally, the identification of the botanical drugs mainly depends on the difference of the appearance of the plants, which could only be figured out by the experienced herbalist physicians. Some modern analytical techniques such as gas chromatography (GC), gas chromatography-mass spectrometry (GC-MS), high performance liquid chromatography-mass spectrometry (HPLC-MS) and thin layer chromatography (TLC), have also been adopted for this purpose. Because of time-consuming and laborious pre-preparation of sample made these methods are unpopular. Therefore, a rapid, simple and accurate analytical method is essentially required for the qualitative analysis of the herbal medicines. Nowadays, visible-near infrared (VIS-NIR) spectroscopy is an attractive technique due to its non-destructive, simple and fast characteristics, which make this technology ideally suited for quality control. It has been increasingly adopted as an analytical tool for numerous applications in several fields, including medical, pharmaceutical and the most traditional food analysis[6-11]. Near infrared (NIR) spectral data of a sample can be treated as a signature, allowing samples to be grouped on the basis of their spectral similarities. Therefore, one of the most common applications of NIR technique is for classification combined with pattern recognition methods[12].

Because the modern spectroscopy instrumentations usually have a high resolution, their obtained spectral data sets often contain hundreds or thousands of variables. With these so many variables and hundreds of samples, the calibration process is time-consuming and not convenient to fulfill the high speed feature of spectroscopy. The elimination of irrelevant variables might predigest calibration modeling and improve the results in terms of accuracy and robustness. Better calibration model may be obtained by selecting characteristic information such as sample-specific or component-specific variables instead of the full spectra.

This study investigated the potential application of VIS-NIR spectroscopy to differentiate species of *Curcuma* species. The specific objectives of this research were (1) to establish the relationships between the VIS-NIR spectra and species differentiation of *Rhizoma curcumaes*, (2) to obtain effective

wavelength variables based on using successive projections algorithm (SPA) and (3) to establish least squares support vector machine (LS-SVM) model for the discrimination. Models' performances were compared using their correct classification rate (CCR, %).

EXPERIMENTAL

Sample preparation: Samples in the training sets were designed as much as possible to include all sources of sample variability. All fresh plants were obtained from China: *C. phaeocaulis* from Shuangliu of Sichuan province (P); *C. kwangsiensis* from Qinzhou of Guangxi province (K); *C. wenyujin* from Yueqing (W1), Ruian (W2) and Yongjia (W3) of Zhejiang province. Sample origins and the numbers of samples are seen from Table-1. All samples were powdered using a cyclone mill fitted with 1 mm screen. The particle size of powder was below 20 meshes.

TABLE-1
DETAILED INFORMATION OF THE MATERIALS

Sample No.	Species	Origins	Varieties	Harvesting time
1-46	*C. phaeocaulis*	Shuangliu of Sichuan	1	2008.10
47-90	*C. kwangsiensis*	Qinzhou of Guangxi	2	2008.10
91-132	*C. wenyujin*	Yueqing of Zhejiang	3	2008.11
133-175	*C. wenyujin*	Yongjia of Zhejiang	4	2008.11
176-224	*C. wenyujin*	Ruian of Zhejiang	5	2008.11

Spectral measurement: Spectral measurement was taken using USB4000 Miniature Fiber Optic Spectrometer (Ocean Optics, USA). Fiber optic probe was set above from *Curcuma* tuber powder about 3 mm. Before the sample measurement, the dark spectra and white standard spectra were measured to standardize the spectrometer. A total of 30 scans were investigated for each sample using the sample transport module over the entire range of 346.02-1050.49 nm. Averaged transmittance spectra were used for further calculation. There were 46 samples obtained for *C. phaeocaulis*, 44 samples for *C. kwangsiensis*, 134 samples for *C. wenyujin* from three different geographical origins (42 samples from Yueqing; 43 samples from Yongjia and 49 samples from Ruian). The obtained samples were divided into a calibration set and a prediction set. In order to obtain a 1/1 division of calibration/prediction spectra, half samples of each species were selected into the calibration set. Finally the calibration set contains 112 samples and other 112 samples constitute the prediction set.

Spectral pre-processing: It is necessary to do spectral pre-processing for an optimal performance before the calibration stage. To avoid low signal-to-noise ratio, the region of wavelengths (474.01-940.93 nm, 2401 spectral variables) was employed for the calculations of *Curcuma* tuber fragments and the region of wavelengths (410.35-1048.86 nm, 3339 spectral variables) was employed for the calculations of *Curcuma* tuber. Absorbance data were stored as log (1/R) (R = reflectance). Several spectral pre-processing algorithms, including Savitzky-Golay smoothing (SGS)[13], multiplicative scatter correction (MSC)[14] and standard normal variate (SNV)[15], were operated in MATLAB 7.6 (The Math Works, Natick, USA). SGS is an averaging algorithm that fits a polynomial to the data points. MSC is a transformation method used to compensate for additive and/or multiplicative effects in spectral data. Standard normal variate is a row-oriented transformation which centers and scales individual spectra. The performances of these pre-processing algorithms were compared based on LS-SVM calibration.

Successive projections algorithm (SPA): In SPA process[16] the instrumental response data are disposed in a matrix X of dimensions (N × K) such that the k^{th} variable x_k is corresponding to the k^{th} column vector $x_k \in \Re^n$. Let M = min (N-1, K) be the maximum number of selected variables. First step consists of projections carried on the X matrix, which generate k chains of M variables. Each element in a chain is selected in order to display the least collinearity with the previous ones. The second step of SPA consists of evaluating candidate subsets of variables selected in the first step. The candidate subset of m variables starting from x_k is defined by the index set $\{SEl(1, k), SEL(2, k), \ldots, SEL(m, k)\}$. A total of M × K subsets of variables are tested and the best variable subset is selected. For this purpose root mean square error (RMSE) were adopted. The process of SPA was operated in MATLAB 7.6 (The Math Works, Natick, USA).

Methodology of LS-SVM: LS-SVM is an optimized algorithm based on the standard support vector machine by Suykens *et al.*[17]. The LS-SVM has the capability for linear and non-linear multivariate calibration and solves the multivariate calibration problems in a relatively fast way. It uses a linear set of equations instead of a quadratic programming problem to obtain the support vectors[18]. As a non-linear function and a more compacted supported kernel, radial basis function kernel was used in this study to reduce the computational complexity of the training procedure compared to other kernels while giving good performance under general smoothness assumptions[19,20]. Thus, radial basis function (RBF) kernel was used. Grid-search technique was applied to find out the optimal parameter values which include regularization parameter gama (γ) and the radial basis function kernel function parameter sig^2 (σ^2). For each combination of γ and σ^2 parameters, the root mean square error of cross-validation (RMSECV) was calculated and the optimum parameters were selected when produced smaller RMSECV. The details of LS-SVM algorithm could be found in the literature[21]. LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) was applied with MATLAB 7.6 (The Math Works, Natick, USA).

Binary variety number encoding: In this study, each sample in the calibration set of LS-SVM was assigned a dummy variable as a reference value (Five varieties, K, P, W1, W2 and W3), which was an arbitrary number if the sample belongs to a particular variety or if it does not. In order to solve the effect of different defined number for each class when the number of classes is more than two, a multiclass task with M classes needs to encode into a set of L binary classifiers[22]. In this study, five varieties were encoded in a codebook (using

minimum output coding)[23]. Each variety was encoded into two numbers (-1 or 1) in three dimensions, respectively. Varieties 1 was encoded into "-1, -1, -1" for dimension one, two and three, respectively. Varieties 2 was encoded into "-1, -1, +1", Varieties 3 was encoded into "-1, +1, -1", Varieties 4 was encoded into "-1, +1, +1" and Varieties 5 was encoded into "+1, -1, -1".

RESULTS AND DISCUSSION

Features of VIS-NIR spectra: The average absorbance spectra of *Curcuma* tuber samples are shown in Fig. 1. Although *Curcuma* tuber samples were from the same genus, the spectra of P and K are different from W1, W2 and W3. The average spectra of W1 are also different from W2 and W3, but the differences are not large. According to the spectra of samples from same species, when the spectra of more samples, not the average spectra, were considered, it is hard to distinguish samples of different species (Fig. 2), for example, to tell W1 from P and K. The average spectra of W2 and W3 are very close. It is similar in Fig. 2. Therefore, it is difficult to distinguish samples of W2 and W3 by directly observe their spectra.

Full-spectrum (FS) calculation: In order to get an overview of VIS-NIR spectra for the species discrimination of *Curcuma* tuber samples, LS-SVM models were established based on original full-spectrum and pre-processed spectra. The results are shown in Table-2. Both the pre-processing of MSC and SNV did the best, while the results of SGS and original spectra were not good. Overall, MSC and SNV were two effective spectral pretreatment methods and the pretreated spectra by MSC were used for the further analysis.

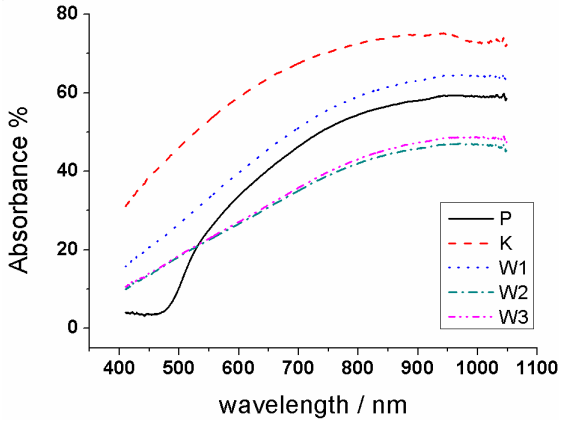


Fig. 1. Average absorbance spectra of *Curcuma* species samples

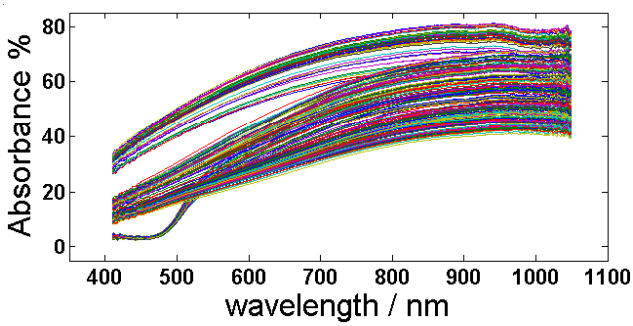


Fig. 2. Absorbance spectra of all *Curcuma* species samples

Successive projections algorithm calculation based on the full-spectrum: From the results of full-spectrum analysis, it can be concluded that VIS-NIR spectroscopy can be used for the non-invasive discrimination of *Curcuma* tuber. However, over thousands of variables caused the model establishment time-consuming. Therefore, variable selection should be executed to simplify and optimize the model.

Successive projections algorithm was carried out for selecting effective wavelength variables from the full spectra. Fig. 3 shows the RMSE scree plot. The solid square shows the selected variable numbers. As can be seen, a sharp fall is shown in the starting part of the RMSE curve as the numbers of selected variables were from 1-3. Then the trends of RMSE curves become marginal with further increasing number of selected variables. The curve tends to level off after the determination of selected variables by F-test criterion[24] with α = 0.25. Finally eight (RMSE = 0.51231) variables were selected. Fig. 4 is the plot of eight wavelength variables selected by SPA. Black columns represent the selected wavelengths. The curve shows the original spectrum. Different three main regions around 520, 800-860 and 980-1040 nm were selected. The selected wavelength variables were set as the inputs of LS-SVM models (Table-2). The performance of MSC-SPA-LS-SVM model which only had eight input variables was better than MSC-FS-LS-SVM model of 3339 variables (99.11 % *versus* 96.43 % for prediction set). As there were only eight variables, the training time of LS-SVM procedure can be saved.

High discrimination result of 99.11 % CCR was obtained by MSC-SPA-LS-SVM model, which shows that VIS-NIR spectroscopy can be used for the non-invasive discrimination of *Curcuma* tuber. To analyze the discrimination results of specific species in Table-2, P and K can be well distinguished than other three varieties, While, W1, W2 and W3 were difficult

TABLE-2
CORRECT DISCRIMINATION RATES OF MODELS FOR THE DISCRIMINATION OF *Curcuma* TUBER SPECIES

Pre-processing	Variable selection	Variable number	Species						
			(%)	P	K	W1	W2	W3	All
No	No	3339	Calibration	100	100	100	100	100	100
			Prediction	100	100	100	79.17	100	91.07
SGS	No	3339	Calibration	100	100	100	100	100	100
			Prediction	100	100	100	75	100	90.18
SNV	No	3339	Calibration	100	100	100	100	100	100
			Prediction	100	100	100	87.5	100	96.43
MSC	No	3339	Calibration	100	100	100	100	100	100
			Prediction	100	100	100	87.5	100	96.43
MSC	SPA	8	Calibration	100	100	100	100	100	100
			Prediction	100	100	100	95.83	100	99.11

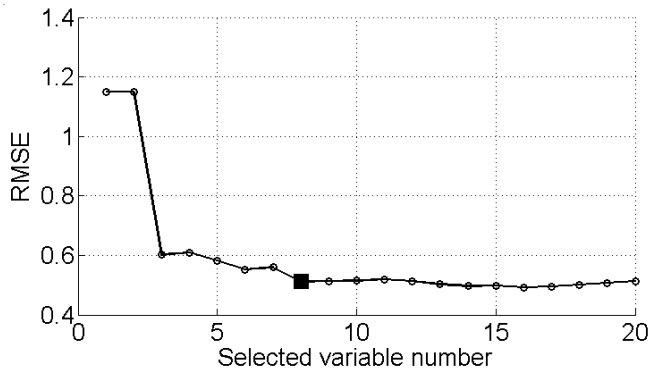


Fig. 3. RMSE scree plot obtained by SPA based on the whole spectra of *Curcuma* species samples

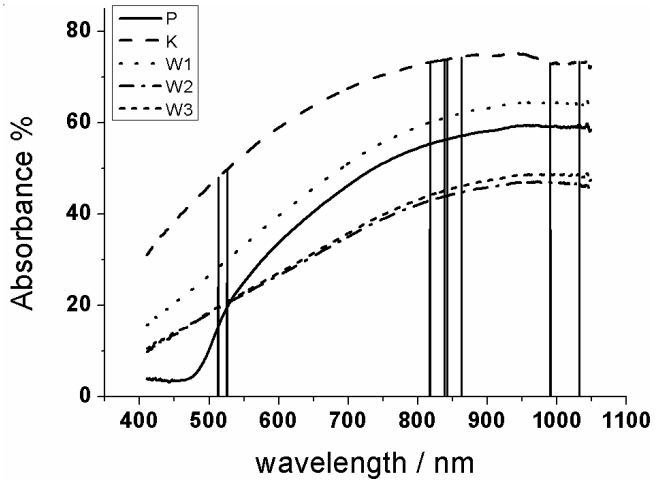


Fig. 4. Plot of eight selected wavelength variables by SPA. Black columns represent selected wavelength variables. The curve shows the original spectrum

to be distinguished from each other, especially for W2 and caused some errors. Because they are from the same species and the same Province (Yueqing,Yongjia and Ruian lie in the south of Zhejiang Province). To distinguish the same species of *Curcuma* tuber samples of different varieties from the same region should be further analyzed.

Conclusion

The predictive capacity of VIS-NIR spectroscopy to the fast species discrimination of *Curcuma* tuber was demonstrated. Multiplicative scatter correction was chosen as the best spectral pre-processing algorithm. Successive projections algorithm was operated for the wavelength variable selection, and eight effective variables were selected from thousands of original FS variables. The best CCR of 99.11 % was obtained by MSC-SPA-LS-SVM model based on powder samples. Samples from different species can be correctly distinguished, while the discrimination of the same species from different growing region had some error. The overall results demonstrate the ability of VIS-NIR spectroscopy for the fast discrimination of different species of *Curcuma* species and successive projections algorithm is a useful algorithm for the spectral variable selection.

ACKNOWLEDGEMENTS

The work was supported in part by grants from NSFC 30900129, ZJNSFC Y2080331, ZJST 2009C34014 and WZST H20080052.

REFERENCES

1. W.J. Lang, A Thousand Formulae of Prepared Decoctions for Anticancer, Chinese Medicinal Technology Press: Beijing, China, p. 119 (1992).
2. X.Y. Zheng, Pharmacopoeia of the People's Republic of China, Chemistry Industry Press: Beijing, China, p. 230 (2000).
3. C. Selvam, S.M. Jachak, R. Thilagavathi and A.K . Chakraborti, *Bioorg. Med. Chem. Lett*., **15**, 1793 (2005).
4. J.J. Johnson and H. Mukhtar, *Cancer Lett*., **255**, 170 (2007).
5. N.Y. Qin, F.Q. Yang and Y.T. Wang, *J. Pharm. Biomed. Anal*., **43**, 486 (2007).
6. F.W. McClure, *J. Near Infrared Spectrosc*., **11**, 487 (2003).
7. M. Blanco and I. Villaroya, *Trends Anal. Chem*., **21**, 40 (2002).
8. Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond and N. Jent, *J. Pharm. Biomed. Anal*., **44**, 683 (2007).
9. H. Huang, H. Yu, H. Xu and Y. Ying, *J. Food Eng*., **87**, 303 (2008).
10. B.M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. Theron, and J. Lammertyn, *Postharvest Biol. Technol*., **46**, 99 (2007).
11. J. Luypaert, D.L. Massart and Y. Vander Heyden, *Talanta*, **72**, 865 (2007).
12. N. Smola and U. Urleb, *Anal. Chim. Acta*, **410**, 203 (2000).
13. A. Savitzky and M. Golay, *Anal. Chem*., **36**, 1627 (1964).
14. I.S. Helland, T. Naes and T. Isaksson, *Intell. Lab. Syst*., **29**, 233(1995).
15. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc*., **43**, 772 (1989).
16. R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares and H.M. Paiva, *Chemometrics Intell. Lab. Syst*., **92**, 83 (2008).
17. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, Least Squares Support Vector Machines, World Scientific Publishing, Singapore, p. 71 (2002).
18. J.Z. Li, H.X. Liu, X.J. Yao, M.C. Liu, Z. Hu and B.T. Fan, *Anal. Chim. Acta*, **581**, 333 (2007).
19. L. Lukas, A. Devos, J.A.K. Suykens, L. Vanhamme, F.A. Howe, C. Majos, A. Moreno-Torres, M. Van der Graaf, A.R. Tate, C. Arus and S. Van Huffel, *Artif. Intell. Med*., **31**, 73 (2004).
20. T. Hou, W. Zhang, D.A. Case and W. Wang, *J. Mol. Biol*., **376**, 1201 (2008).
21. D. Wu, Y. He, S. Feng and D.-W. Sun, *J. Food Eng*., **84**, 124 (2008).
22. E.L. Allwein, R.E. Schapire and Y. Singer, *J. Machine Learn Res*., **1**, 113 (2000).
23. J.A.K. Suykens and J. Vandewalle, In Proc. the Int. Joint Conf. on Neural Networks (IJCNN'99), Washington, DC; p. 900 (2008).
24. M.C. Breitkreitz, I.M. Raimundo, J.J.R. Rohwedder, C. Pasquini, H.A.D. Filho, G.E. José and M.C.U. Araújo, *Analyst*, **128**, 1204 (2003).