DOI:10.1021/jf900474a

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY
ARTICLE

Determination of Protein Content of *Auricularia auricula* Using Near Infrared Spectroscopy Combined with Linear and Nonlinear Calibrations

Fei Liu, Yong He,* and Guangming Sun

College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China

Near infrared (NIR) spectroscopy was investigated to determine the protein content of *Auricularia auricula* (commonly called black woody ear or tree ear) using partial least-squares (PLS), multiple linear regression (MLR), and least-squares-support vector machine (LS-SVM). The performances of different preprocessing were compared including Savitzky−Golay (SG) smoothing, standard normal variate, multiplicative scatter correction (MSC), first derivative, second derivative, and direct orthogonal signal correction. A successive projections algorithm (SPA) was applied for relevant effective wavelengths selection. The combinations of various pretreatment and calibration methods were compared based on the prediction performance. The optimal full-spectrum PLS model was achieved by raw spectra, whereas the optimal SPA-MLR, SPA-PLS, and SPA-LS-SVM models were achieved by MSC spectra. The best prediction performance was achieved by the SPA-LS-SVM model, with correlation coefficients (r) = 0.9839 and a root mean squares error of prediction (RMSEP) = 0.16. The results indicated that NIR spectroscopy combined with SPA-LS-SVM was the most successful to determine the protein content of *A. auricula*.

KEYWORDS: Near infrared spectroscopy; direct orthogonal signal correction; successive projections algorithm; least-squares-support vector machine; *Auricularia auricula*; protein content

INTRODUCTION

Auricularia auricula (commonly called black woody ear or tree ear) is a macrofungus. It is traditionally used as a food and drug because of its rich nutrients, which include carbohydrates, proteins, fats, fibers, vitamins, and minerals. In recent years, some researchers have been focusing on its medicinal functions, such as its antioxidant activity and anticoagulant activity, as a functional food (*1*−*4*). The protein content is one of the most important quality indices in *A. auricula*, which has many medicinal functions. The protein in *A. auricula* is rich in amino acids, with an especially high content of lysine and leucine. These parameters are high nutritional factors in *A. auricula*. The determination of protein is also quite helpful for further studies of the fast detection of amino acids in *A. auricula*. Hence, the fast and accurate determination of protein contents are important for food quality and authenticity control, and it is quite helpful to keep a fair and reasonable competitive market. The traditional method for protein determination was the Kjeldahl method, the standard method of the Association of Official Analytical Chemists (AOAC). A recently developed method was the Dumas combustion method performed by Rapid N Cube (Elementar Analysensysteme, Hanau, Germany). As compared with the Kjeldahl method, the analytical time of the Dumas combustion method was reduced to some extent, whereas the consumables were quite costly. Many studies showed that there was little practical impact on the reliability of N values between Kjeldahl and Dumas methods or there was a simple equation to transfer these two kinds of N values (*5*). These two methods were laborious, costly, and not convenient enough for fast and nondestructive determination of the protein content. In this study, we were going to investigate the feasibility of determining the protein content of *A. auricula* using near-infrared (NIR) spectroscopy combined with linear and nonlinear calibrations. A NIR spectroscopic technique has wide applications due to its characteristics of high speed, low cost, and nondestructive analysis in the agriculture, pharmaceuticals, food, textiles, cosmetics, and polymer production industries (*6*−*11*). In the application aspect of *A. auricula*, some researchers studied the constitutions of some edible tropical species of mushrooms (*12*), the extraction of polysaccharide and hypoglycemia activity (*13*), and the determination of nickel using adsorbed resin phase spectrophotometry (*14*). Others studied the chemical constituents of *A. auricula* using Fourier transform infrared spectroscopy (*15*), the discrimination of three mushrooms (*A. auricula*, *Boletus aereus*, and *Tremella fuciformis*) by Fourier transform infrared spectroscopy (*16*), and the discrimination of producing areas of *A. auricula* using both visible and NIR spectra (*17*, *18*). However, to our knowledge, there were few reports about the determination of the protein content of *A. auricula* using NIR spectroscopy (1100−2500 nm) based on linear and nonlinear calibrations together with successive projection algorithms (SPA) for variable selection. The fast and accurate detection method of protein content using spectroscopic techniques is helpful for further studies in *A. auricula*, such as the

*To whom correspondence should be addressed. Tel: +86-571-86971143. Fax: +86-571-86971143. E-mail: yhe@zju.edu.cn.

 Published on Web 05/15/2009

detection of nutritional parameters (polysaccharides, amino acids, fat, and fiber) and the development of detection sensors and portable instruments for quality detection.

The objective of this study was to investigate the potential feasibility of using NIR spectroscopy combined with linear and nonlinear calibrations to determine the protein content of *A. auricula*. The performance of different preprocessing methods was compared including Savitzky−Golay (SG) smoothing, standard normal variate (SNV), multiplicative scatter correction (MSC), first derivative (1-Der), second derivative (2-Der), and direct orthogonal signal correction (DOSC). SPA was applied for relevant variable selection to develop and compare the prediction performance of multiple linear regression (MLR), partial least-squares (PLS), and least-squares-support vector machine (LS-SVM) models.

MATERIALS AND METHODS

Sample Preparation. Four major varieties (geographical origins) of commercial *A. auricula* were collected as representatives of different cultivated environments, including Qishan (Anhui, China), Heihe (Heilongjiang, China), Huangshan (Anhui, China), and Qingyuan (Zhejiang, China). These samples were cultivated in different environments and collected at different times. Therefore, these four varieties of *A. auricula* were more representative for further calibrations. These samples were first dried in an oven at 60 °C for 2 days and then ground in an electric mill (Universal High-Speed Smashing Machines, model FW100, Tianjin City Taisite Instrument Co., Ltd., Tianjin, China). The ground samples were screened through a 60 mesh sieve. These preparations were implemented to keep the samples in the same experimental condition and also to reduce the influence of other physical properties. Then, the performance of different calibration methods could be compared without external influences. Finally, 60 samples for each variety and a total of 240 samples were obtained and randomly separated into calibration, validation, and prediction sets. The calibration set was composed of 120 samples (30 for each variety), the validation set was 60 samples (15 for each variety), and the remaining 60 samples were used as the prediction set. The calibration and validation sets were applied to develop a stable calibration model, in which the validation set was used to validate the calibration stage and avoid overfitting problems instead of a full cross-validation procedure. The prediction set was applied as an independent set to evaluate and assess the prediction performance of the developed calibration models.

Reference Method for Protein Content. The reference method for protein content detection was the Dumas combustion method using Rapid N Cube (Elementar Analysensysteme). The samples were weighed by a four decimal balance, Sartorius BS224S (Sartorius AG, Goettingen, Germany). After complete combustion, reduction, purification, and detection, the nitrogen content of *A. auricula* was obtained through the Rapid N Software V 3.4.0 (Elementar Analysensysteme). The protein content of *A. auricula* was calculated as the value of total $N \times 6.25$.

Spectral Acquisition. The milled samples stored in a refrigerator were taken out to reach room temperature at 20−23 °C. Three reflectance spectra of each sample were obtained by the Foss NIRSystems 5000 (Foss NIRSystems, Denmark). The scanning intervals were 2 nm within the region of 1100−2500 nm. The spectral collection software was WinISI II V1.5. Three replicate spectra of one sample were averaged into one spectrum. A total of 240 averaged spectra corresponding to the 240 samples were collected.

Spectral Preprocessing. To achieve the optimal prediction performance, different spectral preprocessing methods were applied for comparison. First, the spectral data were transformed into ASCII format, and then, the reflectance spectra were transformed into absorbance spectra by $\log(1/R)$ (R = reflectance). The applied data preprocessing methods included SG smoothing, SNV, MSC, 1-Der, 2-Der, and DOSC. The SG smoothing could be applied for denosing (*19*). SNV and MSC were applied for light scatter correction and reduction of the changes of light path length (*20*, *21*). The 1-Der and 2-Der were used to eliminate the baseline shift (*22*). These pretreatments were implemented by "The Unscrambler 9.8" (CAMO AS, Oslo, Norway). DOSC was applied to correct the major variance sources such as temperature effects, time influences, and instrumental differences in spectral data (*23*). The procedure of DOSC was implemented by MATLAB 7.0 (The Math Works, Natick, United States).

Variable Selection by SPA. SPA was recently developed as a relevant variable selection method. It is a forward variable selection algorithm, which applies vector projection operations in a vector space to select the most relevant variables with the least collinearity and redundancies for the development of multivariate calibration (*24*, *25*). In SPA, The matrix $\boldsymbol{X}$ was composed of the instrumental response data (spectral data). The dimensions of matrix $\boldsymbol{X}_{(N\times K)}$ are that the k-th variable x_k corresponds to the k-th column vector $x_k \in \mathcal{R}^N$. Let $\boldsymbol{M} = \min(N-1, K)$ be the maximum number of selected variables for later calibration models. First, the projections are carried on $\boldsymbol{X}$, which generate k chains of $\boldsymbol{M}$ variables each time. Each element in a chain is selected in order to display the least collinearity with previous ones. The construction of each chain starts from one of variables x_k, $k = 1, ..., K$ and follows a comparison step of projections until the needed relevant variable is selected. Then, the selected variables are thought as effective wavelengths (EWs). Herein, SPA was implemented to the spectra preprocessed by the aforementioned different pretreatment methods. Thus, the EWs selected by each preprocessing were used as the inputs of MLR, PLS, and LS-SVM models.

Modeling by MLR, PLS, and LS-SVM. MLR is a simple and easily interpreted calibration method but is interrupted by the collinearity between the variables (*26*). For the development of MLR models, the number of input variables should be less than the sample number and larger than the response chemical variable number. Herein, the input variables were the selected EWs by SPA with different preprocessing. EWs selected by SPA removed the most collinearity and redundancies in the preprocessed spectra. The response chemical variable was the protein content. Because all of the selected EWs were applied in the SPA-MLR model, the prediction performance could directly demonstrate the power of SPA and the prediction capability of SPA-MLR model.

PLS is another widely applied multivariate calibration method in the application of spectroscopic technique (*27*). The PLS model could develop a linear relationship between the inputs (spectral data) and the response chemical variable (protein content). During the calibration stage, PLS employs latent variables (LVs) instead of real variables (spectral data). To develop a parsimonious model, the selected EWs by SPA with different preprocessing were also applied as input data of PLS to develop SPA-PLS models. For comparison, full-spectrum PLS models were developed with aforementioned different preprocessing methods. In MLR and PLS, the samples in the validation set were used to validate the calibration model. The samples in the prediction set were applied to assess the prediction performance of developed MLR and PLS models.

Moreover, it is worth noting that MLR and PLS methods only deal the linear problems to develop a linear relationship between the spectral variables and the target chemical response (protein content of *A. auricula*). Considering latent nonlinear information existed in the spectral data, LS-SVM was employed to compare the prediction performance with MLR and PLS models.

LS-SVM has a good theoretical foundation in statistical learning methods and handles both linear and nonlinear multivariate problems in a relatively fast way (*28*−*31*). It employs a set of linear equations using support vectors instead of quadratic programming problems to reduce the complexity of optimization processes. The LS-SVM model can be expressed as follows:

$$y(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \quad (1)$$

where α_i is a Lagrange multiplier, $K(x,x_i)$ is a kernel function, and b is the bias value.

During the calibration stage of LS-SVM, input data are first settled using the selected EWs by SPA. The commonly used kernel is the radial basis function (RBF) kernel. The RBF kernel function can be expressed as follows:

$$K(x, x_i) = \exp(-||x - x_i||^2/\sigma^2) \quad (2)$$

where x_i is input data (selected EWs). σ is the RBF kernel parameter, and σ^2 is the bandwidth parameter. Two important parameters in

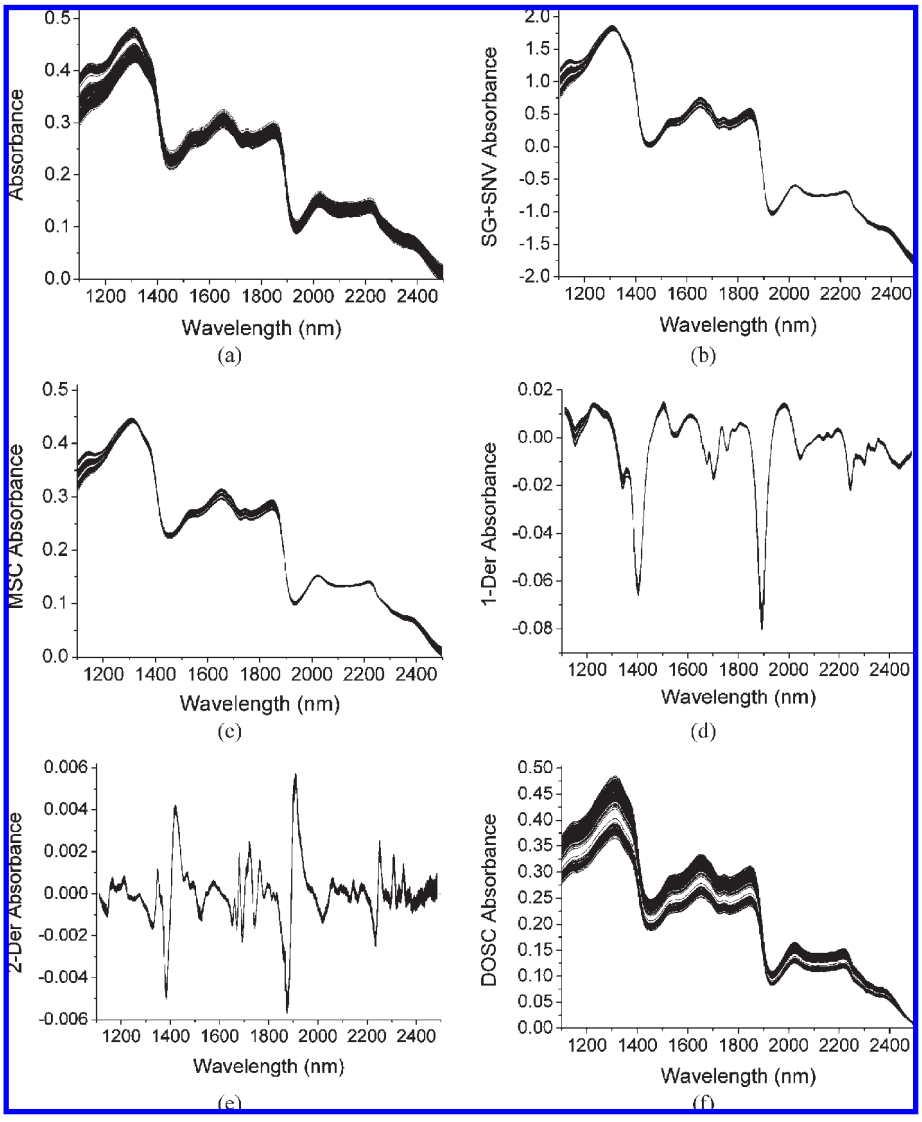

Figure 1. Raw absorbance spectra (**a**) and preprocessed spectra by SG + SNV (**b**), MSC (**c**), 1-Der (**d**), 2-Der (**e**), and DOSC (**f**) of *A. auricula*.

LS-SVM with RBF kernel are the regularization parameter γ and the width parameter σ^2. The regularization parameter γ determined the trade-off between minimizing the training error and minimizing the model complexity. The width parameter σ^2 implicitly defined the nonlinear mapping from input space to some high dimensional feature space. In this paper, the optimal combination of (γ, σ^2) was achieved by a two-step grid search technique. In this procedure, leave-one-out cross-validation was used to avoid overfitting problems. The ranges of γ and σ^2 within (10^{-3}−10^5) were based on experience and previous researches (*10*, *30*). All of the calculations for modeling were performed using MATLAB 7.0 (The Math Works). The free LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) was applied with MATLAB to develop the calibration models. The prediction performance was evaluated and assessed by the following indices (*24*, *25*): correlation coefficients (r), root mean squares error (RMSE) of calibration set (RMSEC), validation set (RMSEV), and prediction set (RMSEP), and residual predictive deviation (RPD). Generally, a good model should have a higher r value and lower RMSEC, RMSEV, and RMSEP values. An acceptable model should have a RPD value of more than three. RMSE is calculated as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \qquad (3)$$

where n is the number of samples and y_i and $\hat{y}_i$ are the reference and predicted values of the i-th sample, respectively.

One more thing should be mentioned about the software for spectral preprocessing and calibration. The data were collected by WinISI software, and the preprocessing and calibration were implemented by Unscrambler and MATLAB software. WinISI software could not perform DOSC, SPA, MLR, and LS-SVM. These procedures were helpful for valuable information exploring. Although exporting data into Unscrambler and MATLAB required a little additional time, the models developed using SPA and LS-SVM were more parsimonious and performed a better prediction performance. Hence, Unscrambler and MATLAB software were employed for spectral preprocessing and calibration.

RESULTS AND DISCUSSION

Spectral Features of *A. auricula*. The raw spectra of *A. auricula* are shown in **Figure 1a**. The preprocessed spectra by SG + SNV, MSC, 1-Der, 2-Der, and DOSC are shown in **Figure 1b**−**f**, respectively. The trends of raw spectra were quite similar, but at the region of 1100−1350 nm, the absorbance values were separated, especially in **Figure 1b**−**d**. The SG + SNV and MSC spectra kept all spectral features in raw spectra such as the peaks and valleys. The DOSC spectra were quite different from raw spectra with two separated absorbance bands along with the wavelength. The visual difference might be caused by the orthogonal projection procedure in DOSC pretreatment.

The statistical values of protein content of *A. auricula* in calibration, validation, and prediction sets are shown in **Table 1**.

Table 1. Statistical Values of Protein Content of *A. auricula*

data set	sample no.	range (%)	mean (%)	standard deviation
calibration	120	8.83−11.55	10.65	0.88
validation	60	9.12−11.44	10.65	0.87
prediction	60	8.97−11.50	10.65	0.88
all	240	8.83−11.55	10.65	0.87

Table 2. Prediction Results of Protein Content by Full-Spectrum PLS Models

		calibration		validation		prediction	
preprocessing	LVs	*r*	RMSEC	*r*	RMSEV	*r*	RMSEP
raw	8	0.9792	0.18	0.9610	0.26	0.9632	0.23
SG + SNV	8	0.9787	0.18	0.9607	0.26	0.9606	0.24
MSC	6	0.9770	0.19	0.9606	0.25	0.9624	0.24
1-Der	8	0.9873	0.14	0.9560	0.27	0.9555	0.26
2-Der	6	0.9882	0.13	0.9485	0.28	0.9471	0.32
DOSC	1	0.9879	0.14	0.9902	0.12	0.9575	0.25

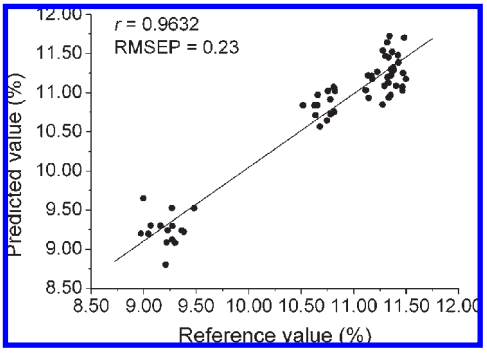

Figure 2. Reference vs predicted values of protein content by an optimal full-spectrum PLS model using raw spectra. The prediction samples were denoted by black circles, which distributed along the regression line.

The ranges of protein content in calibration set are 8.83−11.55%, which covered a larger range than validation and prediction sets. This would be helpful for the development of a stable and general calibration model. It was a coincidence that the mean values were identical with 10.65% in all three sets.

Full-Spectrum PLS Models. Full-spectrum PLS models were developed using raw and preprocessed spectra with aforementioned preprocessing methods (SG + SNV, MSC, 1-Der, 2-Der, and DOSC). The validation set was applied to validate the achievement of a stable calibration model. Different LVs were used according to different preprocessing. The prediction performance was assessed by prediction set. The main evaluation standards were the *r* values and RMSEP of the prediction set. The optimal prediction performance was achieved by raw spectra with $r = 0.9632$ and RMSEP = 0.23. Then, the following model was MSC spectra with $r = 0.9624$ and RMSEP = 0.24 (**Table 2**). The reference vs predicted values of protein content by full-spectrum PLS (raw) model are shown in **Figure 2**. As indicated in **Figure 2**, there were three obvious levels of sample scatters corresponding to four varieties of *A. auricula*. The reason was that the mean values of protein content in prediction set were 9.22, 10.72, 11.26, and 11.40% for Qishan, Heiheng, Huangshan, and Qingyuan, respectively. The mean values of Huangshan (11.26%) and Qingyuan (11.40%) were quite close, in that their scatters were mixed together in the top level in **Figure 2**. Although all developed full-spectrum PLS models obtained acceptable results, the input variables were 700 wavelengths, which included much collinearity and redundant information. Hence, a variable selection method should be introduced to remove the collinearity and redundancies of the spectral data. Herein, SPA was recommended to implement such a procedure, as shown as an effective method (*24*, *25*, *32*, *33*).

Table 3. Selected EW by SPA with Different Preprocessing

preprocessing	no.	selected wavelengths (nm)
raw	7	1318, 1100, 1396, 2214, 1834, 1520, 2498
SG + SNV	9	1304, 2306, 2012, 1520, 2390, 2176, 1918, 2052, 2268
MSC	6	1316, 2498, 1894, 2074, 2186, 1908
1-Der	5	1718, 2216, 2118, 2226, 2108
2-Der	2	1690, 1314
DOSC	1	2320

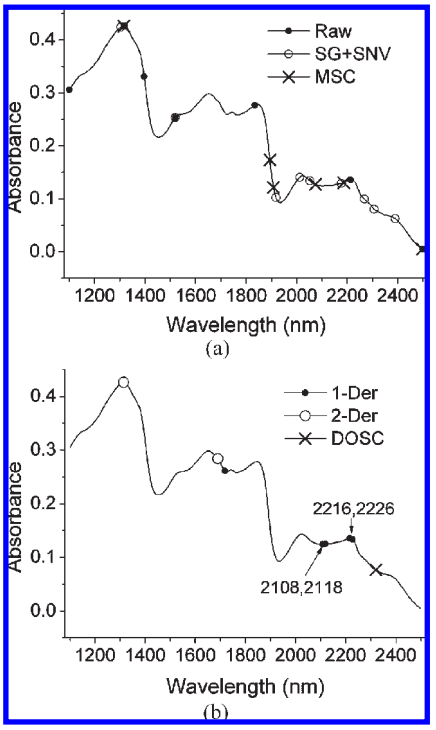

Figure 3. Selected EWs by SPA with preprocessing of raw, SG + SNV, and MSC (**a**) and 1-Der, 2-Der, and DOSC (**b**).

EWs Selected by SPA. The SPA procedure was implemented to the spectra, which were pretreated by different preprocessing methods. It was worth noting that the validation set was applied for the guidance of selection of candidate subsets of variables. The prediction set was utilized in the final performance evaluation of the resulting models. The maximum number of selected EWs was set as 30. **Table 3** shows the selected EWs by SPA according to each preprocessing. The locations of EWs in the spectra are shown in **Figure 3**. The number of selected EWs was less than 10, which would be helpful to develop more parsimonious models. The EWs selected by each preprocessing in **Table 3** were listed according to the significance, with the most important listed first. The most important wavelength was that selected first by SPA for each preprocessing. Take MSC spectra for instance, 1316 nm was thought to be the most relevant one. To show the locations of selected EWs, the selected EWs by SPA with different preprocessing are shown in **Figure 3**.

The wavelength bands selected between 1254 and 1348 nm (1304, 1314, 1316, and 1318 nm) could be attributed to the combination of the first overtone of N−H stretch with fundamental N−H in plane bend and C−N stretch with N−H in-plane

Table 4. Prediction Results of Protein Content by SPA-MLR, SPA-PLS, SPA-LS-SVM, and Linear Function Models

		calibration		validation		prediction		
preprocessing	EWs/LVs/(γ, σ^2)	r	RMSEC	r	RMSEV	r	RMSEP	RPD
				SPA-MLR				
raw	7/−/−	0.9699	0.21	0.9560	0.26	0.9590	0.25	3.5
SG + SNV	9/−/−	0.9771	0.19	0.9589	0.25	0.9429	0.29	3.0
MSC	6/−/−	0.9705	0.21	0.9599	0.25	0.9599	0.25	3.5
1-Der	5/−/−	0.9354	0.31	0.9392	0.30	0.9176	0.38	2.3
2-Der	2/−/−	0.9398	0.30	0.9206	0.34	0.9122	0.39	2.3
DOSC	1/−/−	0.9873	0.14	0.9905	0.12	0.9577	0.25	3.5
				SPA-PLS				
raw	7/6/−	0.9685	0.22	0.9506	0.29	0.9549	0.26	3.4
SG + SNV	9/7/−	0.9760	0.19	0.9613	0.24	0.9476	0.28	3.1
MSC	6/3/−	0.9645	0.23	0.9561	0.26	0.9635	0.24	3.7
1-Der	5/2/−	0.9315	0.32	0.9418	0.29	0.9148	0.38	2.3
2-Der	2/1/−	0.9398	0.30	0.9214	0.34	0.9124	0.39	2.3
DOSC	1/1/−	0.9783	0.14	0.9905	0.12	0.9577	0.25	3.5
				SPA-LS-SVM				
raw	7/−/(546.4, 27.8)	0.9864	0.14	0.9844	0.15	0.9737	0.20	4.4
SG + SNV	9/−/(31.4, 5.0)	0.9975	0.06	0.9980	0.05	0.9757	0.19	4.6
MSC	6/−/(32.8, 4.9)	0.9945	0.09	0.9970	0.07	0.9839	0.16	5.5
1-Der	5/−/(51.4, 14.0)	0.9803	0.17	0.9818	0.16	0.9586	0.25	3.5
2-Der	2/−/(9.5, 1.1)	0.9829	0.16	0.9910	0.12	0.9752	0.16	5.5
DOSC	1/−/(1.7×10^3, 0.2)	0.9920	0.11	0.9970	0.07	0.9607	0.24	3.7
				linear function				
DOSC	1/−/−	0.9869	0.14	0.9905	0.12	0.9576	0.25	3.5

bend. This region was thought to be associated with the amide vibrations characteristic for protein (*34*). EWs around 1908 and 1918 nm could confirm highly hydrophilic properties of protein since these wavelengths could be attributed to the combination of the O−H stretch and the O−H deformation (*35*). The selected EWs around 2012, 2052, 2108, and 2118 nm might be correlated with N−H stretch and N−H in-plane motion (*34*). Wavelengths around 2176, 2186, 2214, 2216, 2226, and 2268 nm might also be attributed to the combination of N−H, C−N, and C═O stretch (*36*, *37*). Wavelengths around 2390 and 2498 nm could be associated with the stretching and bending vibrations of the CH_2 groups of the side chains of different amino acids (*38*). The above analysis indicated the EWs selected by SPA had a close relationship with the response of protein. Different preprocessing could select similar EWs (**Table 3**), such as 1520 nm by both raw and SG + SNV spectra, 2498 nm by both raw and MSC spectra, 2214 (raw) and 2216 nm (1-Der), 1314, 1316, and 1318 nm by 2-Der, MSC, and raw spectra, respectively. These results indicated the relevance of the selected similar EWs.

SPA-MLR, SPA-PLS, and SPA-LS-SVM Models. Using the selected EWs by SPA, the protein content of *A. auricula* was determined by SPA-MLR, SPA-PLS, and SPA-LS-SVM models. Herein, SPA-MLR and SPA-PLS were linear calibrations, whereas SPA-LS-SVM belonged to nonlinear calibration. The EWs selected using different preprocessed spectra were compared for prediction performance. **Table 4** shows the overall results of each model based on a different preprocessing combined SPA selection and modeling method.

SPA-MLR models were developed directly using the EWs, and the prediction performance could directly indicate the effectiveness of the EWs. The prediction results are shown in **Table 4**. As can be seen, the optimal prediction results were obtained by MSC spectra with $r = 0.9599$ and RMSEP = 0.25. The performance was not as good as a full-spectrum PLS (raw) model with $r = 0.9632$ and RMSEP = 0.23. However, the prediction performance was not severely impaired after reducing the number of wavelengths from 700 to 6 and was acceptable for applications ($r > 0.95$).

SPA-PLS models were developed based on the selected EWs, and different LVs were applied in the calibration models. The optimal prediction performance was also achieved by MSC spectra with $r = 0.9635$ and RMSEP = 0.24. Three LVs were used in this SPA-PLS (MSC) model. The result was slightly better than the full-spectrum PLS models except the RMSEP value in the full-spectrum PLS (raw) model. The RMSEP = 0.24 by SPA-PLS (MSC) model was slightly larger than RMSEP = 0.23 by the full-spectrum PLS (raw) model. However, these two values (RMSEP = 0.24 and RMSEP = 0.23) were quite close to each other. Considering the variables used in these two models (SPA-PLS with six wavelengths and full-spectrum PLS with 700 wavelengths), the SPA-PLS model was more simple and parsimonious to understand. The selected EWs in MSC spectra would be helpful for further practical applications like commercial portable instrument development for protein detection of *A. auricula*. In this point of view, the SPA-PLS (MSC) model was better than full-spectrum PLS models.

Considering that the latent nonlinear information existed in the spectral data, LS-SVM was recommended to develop the SPA-LS-SVM model to determine the protein content of *A. auricula*. Using the selected EWs could reduce the computational time to develop LS-SVM models because the training time using LS-SVM increased with the square of the number of training samples and linearly with the number of variables (dimension of spectra) (*39*). The kernel function was the aforementioned RBF kernel. The model parameters (γ, σ^2) were determined by a two-step grid search technique as stated above. The optimal combinations of (γ, σ^2) were determined according to preprocessing methods. The validation set was applied for validation of

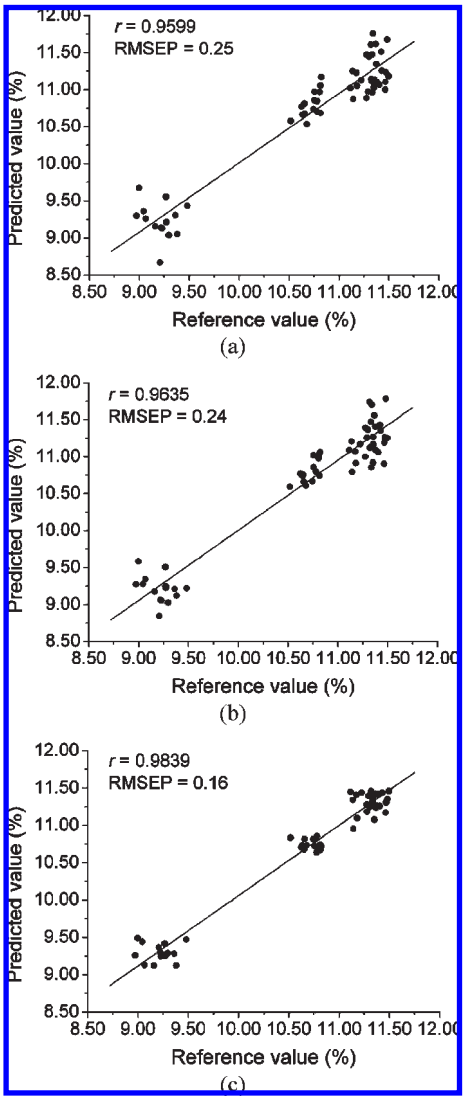

Figure 4. Reference vs predicted values of protein content by optimal SPA-MLR (**a**), SPA-PLS (**b**), and SPA-LS-SVM (**c**) models using MSC spectra.

calibration model, and prediction set was used to assess and evaluate the prediction performance of the developed model. The prediction results by SPA-LS-SVM models are shown in **Table 4**. As can be seen, the optimal prediction performance was achieved by LS-SVM (MSC) model with $r = 0.9839$ and RMSEP = 0.16, which was the best model as compared with the full-spectrum PLS, SPA-MLR, SPA-PLS, and SPA-LS-SVM models. The prediction performance of all developed SPA-LS-SVM models was better than that of linear PLS and MLR models except SPA-LS-SVM (1-Der and DOSC) models. The reason for a better performance was that LS-SVM took both linear and latent nonlinear relevant information of the selected EWs, and this nonlinear information improved the prediction performance. The similar results were also in agreement with previous studies (*10*, *30*).

The reference vs predicted values of protein content are shown in **Figure 4a**−**c** for SPA-MLR, SPA-PLS, and SPA-LS-SVM models using MSC spectra, respectively. The sample scatters were much closer to the regression line in **Figure 4c** than that in **Figures 2** and **Figure 4a**,**b**. This also indicated the better prediction performance of SPA-LS-SVM (MSC) model (**Table 4**). Comparing all developed SPA-MLR, SPA-PLS, and SPA-LS-SVM models, it was worth noting that the selected EWs in 2-Der spectra made a bad performance in linear SPA-MLR and SPA-PLS models, whereas they showed a good performance in the nonlinear SPA-LS-SVM model. Moreover, the optimal preprocessing was MSC for all SPA-based models. This might indicate that the different calibration methods would need similar preprocessing for better prediction performance when using a small number of input variables especially in such specific cases. Furthermore, the prediction performance was identical in SPA-MLR and SPA-PLS using DOSC wavelength (2320 m). The reason might be that only one wavelength 2320 nm was selected as EW, and MLR and PLS were both linear calibration methods. The prediction performance of SPA-LS-SVM (DOSC) was different and slightly better than SPA-MLR and SPA-PLS (DOSC) models. It could also be found that the DOSC-based SPA-MLR and SPA-PLS models achieved a good performance in calibration and validation sets, whereas acceptable prediction results ($r > 0.95$) in the prediction set. After comparison, it could be concluded that NIR spectroscopy combined with SPA-LS-SVM could be the most successful to determine the protein content of *A. auricula*.

Considering only one wavelength (2320 nm) was selected as EWs by SPA in DOSC spectra, a direct linear function and other kinds of functions like polynomial function were developed to determine the protein content. The developed functions are shown in **Table 5**. Herein, only the direct linear function was performed for prediction performance. These functions were developed using the samples in calibration and validation sets. The prediction performance was assessed by prediction set. The prediction results are shown in **Table 4**. The reference vs predicted values of protein content are shown in **Figure 5**. As can be seen, the prediction results were acceptable with $r = 0.9576$ and RMSEP = 0.25. The results indicated that the direct linear function could be applied for the determination of protein content of *A. auricula*. The samples distributed not as close as **Figure 4c** by SPA-LS-SVM (MSC) to the regression line. However, this linear function supplied a simple and effective way for further applications in the field of quality detection of *A. auricula*.

Table 5. Direct Linear Function and Other Functions for Protein Content Determination

function type	function $y(x)$ [a]	r
direct linear function	$y(x) = 129.4x + 0.247$	0.9879
polynomial function	$y(x) = -747.2x^2 + 245.1x - 4.196$	0.9884
logarithm function	$y(x) = 9.984 \ln(x) + 35.85$	0.9884
exponent function	$y(x) = 3.835 \exp(12.66x)$	0.9879
power function	$y(x) = 125.2x^{0.977}$	0.9884

[a] For $y(x)$, y is the protein content of *A. auricula*, and x is the value of DOSC spectra (wavelength at 2320 nm) of a certain sample.

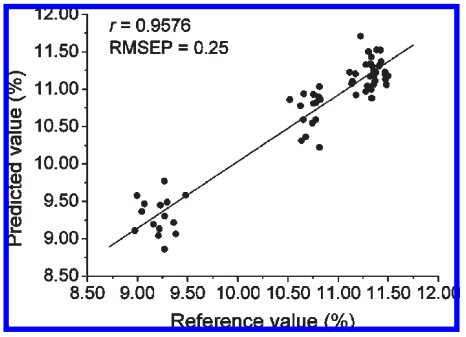

Figure 5. Reference vs predicted values of protein content by direct linear function using DOSC wavelength.

Discussion about the Used Chemometric Methods. The best model for protein content detection is SPA-LS-SVM model. This work is a systematic comparison and analysis of preprocessing methods, variable selection method, and calibration methods. A discussion about the advantages and disadvantages of chemometric methods is addressed as follows:

1. Systematic comparison of preprocessing methods of SG, SNV, 1-Der, 2-Der, MSC, and DOSC: The first five preprocessing (SG, SNV, 1-Der, 2-Der, and MSC) are commonly used methods, which only take the ***X***-variables (spectral data) into consideration and implement the process depending on spectral data and not considering the influence of the ***Y***-variable (chemical constituents). It indicated that the performance of these methods was quite influenced by the spectral data, which might cause three potential problems. First, the preprocessing methods would bring in new noise and irrelevant information to spectral data, like a derivative process (1-Der and 2-Der). Second, some useful and relevant information for the prediction of chemical constituents was removed or reduced by these preprocessing methods. Third, the combination influences the first and second points. However, the newly developed preprocessing method, DOSC, could take both the spectral data and the chemical constituents into consideration. The preprocessed spectra by DOSC could correct the major variance sources such as temperature effects, time influences, and instrumental differences. The preprocessed spectra are more relevant and related to the chemical constituents.
2. Application of variable selection method of SPA: Many studies developed models using full wavelength bands. There are many uninformative and irrelevant variables in the full wavelength region. This makes the model more complex and higher time cost. SPA could select relevant variables with the least collinearity and redundancies. It is a powerful way for variable selection and helpful to reduce the computational time and model complexity. Moreover, the selected relevant variables would be helpful for the development of detection sensors and portable instrument.
3. Comparison of linear (MLR and PLS) and nonlinear (LS-SVM) calibration methods: The spectral data contain both linear and nonlinear useful information for protein content determination. MLR could make full use of the input variables and directly demonstrate the effectiveness of the input data. One constraint of MLR is that the sample number must be larger than the variable number, and the variable number must be equal or more than the ***Y***-variable number. PLS analysis using the LVs instead of original input spectral data could be used to develop the model. LVs could reduce the computational time, but the PLS model could not directly demonstrate which variable was important and how the performance of the input variables is. Furthermore, both MLR and PLS methods only deal with the linear relationship between the spectral data and the chemical constituents. The latent nonlinear information in spectral data could not be applied to improve the predictive performance. However, LS-SVM could take advantage of both linear and nonlinear information in the spectral data. That is why the optimal model for protein content determination is the SPA-LS-SVM model. SPA selected the most relevant and informative variables, and LS-SVM made full use of both linear and nonlinear relations between the selected variables and the protein content to achieve a good prediction performance.
4. The DOSC preprocessed spectral followed by SPA process could select the most useful wavelength to predict the protein content in *A. auricula*. In most DOSC-SPA cases tried in this paper, only one wavelength was selected, and an acceptable prediction performance was achieved. These were new trials and discoveries to bring a bright future to develop detection sensors and portable instrument for quality control in *A. auricula*, foods, and other related fields.

In conclusion, the protein content of *A. auricula* was successfully determined using NIR spectroscopy combined with the SPA-LS-SVM model. The most suitable preprocessing was MSC in SPA-based models. Moreover, SPA was a powerful way for the most relevant variable selection, and the developed SPA-MLR, SPA-PLS, and SPA-LS-SVM models were more simple and parsimonious for further applications such as portable instrument development. The best prediction performance was achieved by the SPA-LS-SVM (MSC) model with $r = 0.9839$ and RMSEP = 0.16. Further studies would be focused on the variable selection and parsimonious function development with higher prediction precision and less number of effective variables.

ABBREVIATIONS USED

NIR, near-infrared; MLR, multiple linear regression; PLS, partial least-squares; LS-SVM, least-squares-support vector machine; SG, Savitzky−Golay; SNV, standard normal variate; MSC, multiplicative scatter correction; 1-Der, first derivative; 2-Der, second derivative; DOSC, direct orthogonal signal correction; SPA, successive projections algorithm; *r*, correlation coefficients; RMSE, root mean squares error; RMSEC, root mean squares error of calibration; RMSEV, root mean squares error of validation; RMSEP, root mean squares error of prediction; RPD, residual predictive deviation; AOAC, Association of Official Analytical Chemists; R, reflectance; EW, effective wavelength; LV, latent variable; RBF, radial basis function.

LITERATURE CITED

(1) Fan, L. S.; Zhang, S. H.; Yu, L.; Ma, L. Evaluation of antioxidant property and quality of breads containing *Auricularia auricula* polysaccharide flour. *Food Chem*. **2006**, *101*, 1158–1163.
(2) Acharya, K.; Samui, K.; Rai, M.; Dutta, B. B.; Acharya, R. Antioxidant and nitric oxide synthase activation properties of *Auricularia auricula*. *Indian J. Exp. Biol*. **2004**, *42*, 538–540.
(3) Takeuchi, H.; He, P. M.; Mooi, L. L. Reductive effect of hot-water extracts from woody ear (*Auricularia auricula-judae* quel.) on food intake and blood glucose concentration in genetically diabetic KK-AY mice. *J. Nutr. Sci. Vitaminol*. **2004**, *50*, 300–304.
(4) Yoon, S. J.; Yu, M. A.; Pyun, Y. R.; Hwang, J. K.; Chu, D. C. The nontoxic mushroom *Auricularia auricula* contains a polysaccharide with anticoagulant activity mediated by antithrombin. *Thromb. Res*. **2003**, *112*, 151–158.
(5) Simonne, A. H.; Simonne, E. H.; Eitenmiller, R. R.; Mills, H. A.; Cresman, C. P.III. Could the Dumas method replace the Kjeldahl digestion for nitrogen and crude protein determinations in foods? *J Sci. Food Agric*. **1997**, *73*, 39–45.
(6) Cen, H. Y.; He, Y. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci. Technol*. **2007**, *18*, 72–83.

(7) Liu, F.; He, Y.; Wang, L.; Pan, H. M. Feasibility of the use of visible and near infrared spectroscopy to assess soluble solids content and pH of rice wines. *J. Food Eng*. **2007**, *83*, 430–435.
(8) Liu, F.; Zhang, F.; Jin, Z. L.; He, Y.; Fang, H.; Ye, Q. F.; Zhou, W. J. Determination of acetolactate synthase activity and protein content of oilseed rape (*Brassica napus* L.) leaves using visible/near infrared spectroscopy. *Anal. Chim. Acta* **2008**, *629*, 56–65.
(9) Yan, Y. L.; Zhao, L. L.; Han, D. H.; Yang, S. M. *The Foundation and Application of Near-infrared Spectroscopy Analysis*; China Light Industry Press: Beijing, China, 2005; pp 1−3.
(10) Liu, F.; He, Y. Use of visible and near infrared spectroscopy and least squares-support vector machine to determine soluble solids content and pH of cola beverage. *J. Agric. Food Chem*. **2007**, *55*, 8883–8888.
(11) Cen, H. Y.; He, Y.; Huang, M. Measurement of soluble solids contents and pH in orange juice using chemometrics and Vis/NIRS. *J. Agric. Food Chem*. **2006**, *54*, 7437–7443.
(12) Aletor, V. A. Compositional studies on edible tropical species of mushrooms. *Food Chem*. **1995**, *54*, 265–268.
(13) Han, C. R.; Ma, Y. Q.; Tang, J. Extraction of polysaccharide from *Auricularia auricula* and its hypoglycemia activity. *J. Food Sci. Biotechnol*. **2006**, *25*, 111–114.
(14) Li, R.; Jiang, Z. T.; Mao, L. Y.; Shen, H. X. Adsorbed resin phase spectrophotometric determination of nickel. *Anal. Chim. Acta* **1998**, *363*, 295–299.
(15) Shi, Y. M.; Liu, G.; Kuy, J. H.; Song, D. H. Identification of *auricularia auricula* from different regions by Fourier transform infrared spectroscopy. *Acta Opt. Sin*. **2007**, *27*, 129–132.
(16) Guo, L. Y.; Liu, G.; Song, D. S.; Liu, J. H.; Zhou, Y. L.; Ou, J. M.; Sun, S. Z. FT-IR study of the mushrooms *Auricularia auricular*, boletus aereus and tremella fuciformis. *J. Yunnan Normal Univ*. **2005**, *25* (3), 48–50.
(17) Liu, F.; He, Y. Discrimination of producing areas of *Auricularia auricula* using visible/near infrared spectroscopy. *Food Bioprocess Technol*. 2008, in press.
(18) Liu, F.; Sun, G. M.; He, Y. Geographical origin discrimination of *Auricularia auricula* using variable selection method of modeling power. *Spectrosc. Spectr. Anal*. 2009, in press.
(19) Gorry, P. A. General least-squares smoothing and differentiation by the convolution (Savitzky−Golay) method. *Anal. Chem*. **1990**, *62*, 570–573.
(20) Dhanoa, M. S.; Lister, S. J.; Sandersona, R.; Barnesb, R. J. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *J. Near Infrared Spectrosc*. **1994**, *2*, 43–47.
(21) Chen, J. Y.; Iyo, C.; Terada, F.; Kawano, S. Effect of multiplicative scatter correction on wavelength selection for near infrared calibration to determine fat content in raw milk. *J. Near Infrared Spectrosc*. **2002**, *10*, 301–307.
(22) Chu, X. L.; Yuan, H. F.; Lu, W. Z. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. *Prog. Chem*. **2004**, *16*, 528–542.
(23) Westerhuis, J. A.; De Jong, S.; Smilde, A. K. Direct orthogonal signal correction. *Chemom. Intell. Lab. Syst*. **2001**, *56*, 13–25.
(24) Araújo, M. C. U.; Saldanha, T. C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst*. **2001**, *57*, 65–73.
(25) Galvão, R. K. H.; Araújo, M. C. U.; Fragoso, W. D.; Silva, E. C.; José, G. E.; Soares, S. F. C.; Paiva, H. M. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemom. Intell. Lab. Syst*. **2008**, *92*, 83–91.
(26) Næs, T.; Mevik, B. H. Understanding the collinearity problem in regression and discriminant analysis. *J. Chemom*. **2001**, *15*, 413–426.
(27) Geladi, P.; Kowalski, B. R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
(28) Suykens, J. A. K.; Vanderwalle, J. Least squares support vector machine classifiers. *Neural Process. Lett*. **1999**, *9*, 293–300.
(29) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.
(30) Liu, F.; He, Y.; Wang, L. Comparison of calibrations for the determination of soluble solids content and pH of rice vinegars using visible and short-wave near infrared spectroscopy. *Anal. Chim. Acta* **2008**, *610*, 196–204.
(31) Guo, H.; Liu, H. P.; Wang, L. Method for selecting parameters of least squares support vector machines and application (in Chinese). *J. Syst. Simul*. **2006**, *18*, 2033–2036.
(32) Liu, F.; Jiang, Y. H.; He, Y. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer. *Anal. Chim. Acta* **2009**, *635*, 45–52.
(33) Liu, F.; He, Y. Application of successive projections algorithm for variable selection to determine organic acids of plum vinegar. *Food Chem*. **2009**, *115*, 1430–1436.
(34) Daszykowski, M.; Wrobel, M. S.; Czarnik-Matusewicz, H.; Walczak, B. Near-infrared reflectance spectroscopy and multivariate calibration techniques applied to modeling the crude protein, fibre and fat content in rapeseed meal. *Analyst* **2008**, *133*, 1523–1531.
(35) Font, R.; Del Rio-Celestino, M.; Cartea, E.; De Haro-Bailón, A. Quantification of glucosinolates in leaves of leaf rape (*Brassica napus* ssp. *pabularia*) by near-infrared spectroscopy. *Phytochemistry* **2005**, *66*, 175–185.
(36) Cowe, I. A.; Koester, S.; Paul, C.; McNicol, J. W.; Cuthbertson, D. C. Principal component analysis of near infrared spectra of whole and ground oilseed rape (*Brassica napus* L.) samples. *Chemom. Intell. Lab. Syst*. **1998**, *3*, 233–242.
(37) Czarnik-Matusewicz, B. *Useful and Advanced Information in the Field of Near Infrared Spectroscopy*; Research Signpost: Trivandrum, India, 2003.
(38) Wang, J.; Sowa, M. G.; Ahmed, M. K.; Mantsch, H. H. Potoacoustic near-infrared investigation of homo-polypeptides. *J. Phys. Chem*. **1994**, *98*, 4748–4755.
(39) Chauchard, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V. Application of LS-SVM to non-linear phenomena in NIR spectroscopy: Development of a robust and portable sensor for acidity prediction in grapes. *Chemom. Intell. Lab. Syst*. **2004**, *71*, 141–150.

Received February 10, 2009. Revised manuscript received April 24, 2009. This study was supported by the National Science and Technology Support Program (2006BAD10A09), 863 National High-Tech Research and Development Plan (2007AA10Z210), Natural Science Foundation of China (Project 30671213), and Science and Technology Department of Zhejiang Province (Project 2005C12029).