Nondestructive Differentiation of Panax Species Using Visible and Shortwave Near-Infrared Spectroscopy

ARTICLE *in* FOOD AND BIOPROCESS TECHNOLOGY · JULY 2009

Impact Factor: 3.13 · DOI: 10.1007/s11947-009-0199-6

4 AUTHORS, INCLUDING:

Xiaojing Chen
Wenzhou University
22 PUBLICATIONS **175** CITATIONS
SEE PROFILE

Di Wu
Zhejiang University
96 PUBLICATIONS **967** CITATIONS
SEE PROFILE

Food Bioprocess Technol (2011) 4:753–761
DOI 10.1007/s11947-009-0199-6

ORIGINAL PAPER

Nondestructive Differentiation of Panax Species Using Visible and Shortwave Near-Infrared Spectroscopy

Xiaojing Chen · Di Wu · Yong He · Shou Liu

Received: 15 October 2008 / Accepted: 3 March 2009 / Published online: 17 March 2009

Abstract Visible and short-wave near-infrared (Vis-SWNIR) spectroscopy was investigated to differentiate species of panax, including American *Panax quinquefoliumI*, Chinese *Panax quinquefolium*, and Chinese *Panax ginseng*. Principal component analysis (PCA) was applied before least-square support vector machine (LS-SVM) modeling, and the vast points of the spectral data (376–1,025 nm) were effectively reduced. PCA-LS-SVM differentiated species with 100% correct classification rate for the tested samples. In addition, effective wavelengths were selected according to modeling power, discrimination power, regression coefficients, loading weights, and genetic algorithms, respectively. The optimal and simplified LS-SVM model with 100% correct classification rate was achieved using the effective wavelengths selected by genetic algorithms. The results showed that Vis-SWNIR spectroscopy technique can be applied as a high accuracy and fast way for the qualitative discrimination of herb species.

Keywords Visible and short-wave near-infrared spectroscopy (Vis-SWNIR) · Least-square support vector machine (LS-SVM) · Principal component analysis (PCA) · Genetic algorithms · Effective wavelengths · Panax and ginseng

Xiaojing Chen and Di Wu contribute equally to this paper.

X. Chen · D. Wu · Y. He (✉)
College of Biosystems Engineering and Food Science,
Zhejiang University, 268 Kaixuan Road,
Hangzhou 310029 Zhejiang,
People's Republic of China
e-mail: yhe@zju.edu.cn

X. Chen · S. Liu
Department of Physics, Xiamen University,
Xiamen 361005, China

Introduction

Asian ginseng (Radix et Rhizoma Ginseng, the root and rhizome of *Panax ginseng* (PG) C.A. Meyer, Araliaceae), mainly cultivated in East Asia (especially in China and Korea), has been widely adapted in traditional Chinese medicine (Nocerino et al. 2000; Bent and Ko 2004). *Panax quinquefolium* (PQ). L (Araliaceae), known as American ginseng, has the medical functions to reduce stress and high blood sugar and adjust immunity (Meng and Li 2003). In China, there are two kinds of PQ, American PQ and Chinese PQ. American PQ is planted in America while Chinese PQ is a kind of PQ transplant in China. Both PQ and PG extracts exhibit free radical scavenging activities depending on their ginsenoside contents, which are varied in ginseng species and specific plant parts (Hu and Kitts 2001). There are some differences in the contents of ginsenoside R_f and 24(*R*)-pseudoginsenoside F_{11} found in these two species (Li et al. 2000). The contents of crude saponin of PG and PQ were 4.8–5.2% and 7.0–7.3%, respectively (Levis 2003).

The sale price of PQ is 5–10 times higher than PG in Chinese herb market because PQ has higher medical value. American PQ has better medical value than Chinese PQ, so the price of American PQ is higher. American PQ, Chinese PQ, and PG are not easily distinguished by morphological features for normal consumers, especially when they are cut into fragments. PG is illegally sold as PQ, and Chinese PQ as American PQ as well. The discrimination between PG and PQ is mainly through morphological and microscope features by experienced experts (Lu et al. 2008) and analytical chemical methods. However, the former method is too subjective, and the latter methods are usually time consuming, destructive, and costly and require professional lab operations, such as UV characteristics (Fuzzati 2004), fluorescence of extract

(Dasgupta and Veras 2006), thin-layer chromatography (Vanhaelen-Fastré et al. 2000), and high-performance liquid chromatography with mass spectrometry (Li et al. 2000).

Visible and short-wave near-infrared (Vis-SWNIR) spectroscopy is widely applied as a fast and nondestructive analytical method (Sinija and Mishra 2009; Shao and He 2009). It is sample preparation free and is ideally suitable for online process monitoring and quality control (Chen et al. 2004; Woodcock et al. 2008). The visible bands contain the color information, and the short-wave near-infrared bands are mainly corresponding to C–H, O–H, and N–H vibrations. Vis-SWNIR spectroscopy has been applied for pharmaceutical materials, agricultural products, and foods (Nicolaï et al. 2007; Roggo et al. 2007; Luypaert et al. 2007; Li and He 2008). Lu et al. (2008) studied the mid-infrared spectroscopy of American PQ and PG. Variation in peak intensity were observed at about 1,640, 1,416, 1,372, and 1,048 cm^{-1} among these species. However, few contributions can be found on the use of Vis-NIR for differentiation of panax species. As the spectra in Vis-SWNIR region are originated from the fundamental bands in mid-infrared region, their results showed the possibility to use Vis-SWNIR spectroscopy technique for the discrimination. Vis-SWNIR spectroscopy technique has many advantages compared to mid-infrared spectroscopy, e.g., high transmittance ability to construct excellent detectors, deep energy penetration with less heating effect, reliable exploitation of single-beam data (Bittner et al. 1995), less effect of water vibration (Reeves 1994), and inexpensive fiber optics with inexpensive light sources (Mayes and Callis 1989). In addition, Vis-SWNIR spectroscopy can be used to design a nondestructive handheld instrument, while the sample is destructed in mid-infrared spectroscopy analysis.

This study investigated the potential application of Vis-SWNIR spectroscopy to differentiate species of American PQ, Chinese PQ, and PG. Different chemometric methods were combined to Vis-SWNIR spectroscopy technique to develop classification models. Least squares support vector machine (LS-SVM) models were developed based on different inputs, and their performances were compared using their correct classification rate (CCR, %).

Materials and Methods

Sample Preparation and Spectra Measurement

Panax samples were provided by School of Pharmacy of Wenzhou Medical College, China. They were PG, produced in Jilin province, China, American PQ in Canada, and Chinese PQ in Jilin province, China. The growing season is 2006. The year of production/batch is 2007. Colors of different varieties are substantial, so it is hard to differentiate panax varieties using naked eyes. The panax samples were dried, so there was less water influence for spectral analysis. Sliced panax fragments of each sample (approximate ellipse shape with macro axis of 2 cm, micro axis of 1 cm, and thickness of 0.5 mm) were evenly spread into the area of 65 mm in diameter for spectra measurement. Each variety included 26 samples for analysis.

Spectra measurement was taken using a handheld spectroradiometer, FieldSpec Pro FR (325–1,075 nm)/A110070 (Analytical Spectral Devices Inc., Boulder, CO, USA). Vis-SWNIR reflectance spectra in the 325–1,075 nm region were recorded. The spectroradiometer uses a 512-element photo-diode array with a resolution of 3.5 nm. The integration time is 128 ms. To exploit the 10° field-of-view of the probe, the spectroradiometer was placed at a distance of approximately 150 mm and 45° angle away from the measurement area. As the only illumination, a halogen light source of Lowell pro-lam 14.5 V Bulb/128690 tungsten (Ushio Lighting Inc. Osaka, Japan) was used. The halogen lamp can offer a constant light source and covers whole visible and near-infrared region. The lamp was placed about 300 mm away from the measurement area and 45° of horizon plane. Figure 1 shows the schematic diagram of the experiment apparatus. Vis-SWNIR reflectance was computed with spectral measurements from both samples and a white standard panel with approximately 100% reflectance from Vis-SWNIR region. Panel reflectance spectrum was taken before the spectral scanning experiment. The spectrum of each sample was measured 30 times. Then, these 30 values were averaged and stored as the absorbance value of this sample. To avoid low signal-to-noise ratio, only the

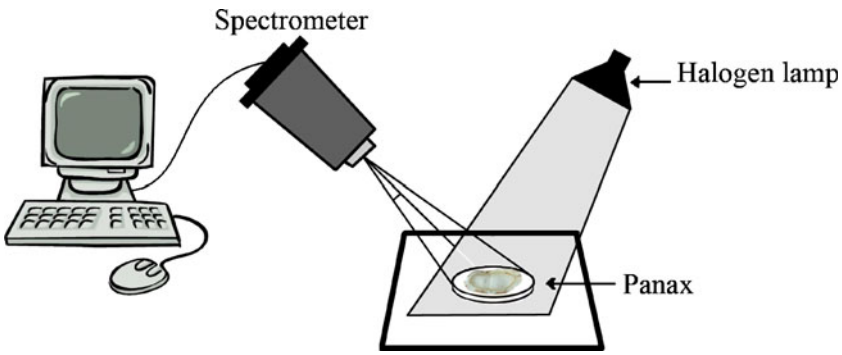

Fig. 1 Schematic diagram of the experiment apparatus

region of wavelengths (375–1024 nm) were employed for the calculations. Absorbance data were stored as log (1/*R*) (*R* = reflectance) at 1-nm intervals (650 spectra data points). Sgolay smooth with span of five was operated as the spectral pretreatment. All spectral data were stored in a computer and processed using RS3 spectral analysis software for Windows (Analytical Spectral Devices, Boulder, CO, USA) designed with a graphical user interface.

Statistical Analysis

Principal Component Analysis

As one of the disadvantages of Vis-NIRS technique, the useful information is usually not prominent in the spectra signals. There were 650 spectra data points for each sample. These large number variables would affect the speed and accuracy of the calculation. PCA is a well-known projection method for re-expression multivariate data. It can reorient the data so that the first few dimensions can explain as much of the useful information as possible. By forming the relatively small number of dimensions named PCs, which are linear composites of the original variables, it is possible to account for most of the information in the original dataset.

In this study, PCA was executed on Vis-SWNIR spectra. Based on principal components (PCs), the principal information of spectrum can be reserved (Cozzolino et al. 2003) and the analysis of spectrum datum would be easier and manageable (Otsuka 2004). PCA was used to derive the first six PCs from the spectral data to ensure that all variability was considered by the analysis. PCA was carried out in Unscrambler ver. 9.7 (CAMO PROCESS AS, Oslo, Norway) computer program.

Model-to-Model Distance

Model-to-model distance was used to show how different two models are from each other (Xiccato et al. 2004). It is calculated by summing the residual standard deviations for all variables within one model when fitted to the other model and dividing this by the sum of the residuals for all variables when fitted to their own model. The distance from a model to itself is one. A model distance of greater than three indicates significantly different models. Model-to-model distance was calculated by Unscrambler ver. 9.7.

Modeling Power and Discrimination Power

Modeling power and discrimination power were used to obtain the effective wavelengths. The modeling power is a measure of the influence of a variable over a given model. It could be used to study the relevance of a variable and indicated how much of the variable's variance was used to describe the class (model). Modeling power can be computed as: modeling power = 1- square root of (variable residual variance/variable total variance). Modeling power was always between 0 and 1. The closer a model is to one, the more important that variable was in the class model. The value close to zero indicates a low contribution (Wold and Sjostrom 1977).

The discrimination power indicates the ability of each variable to contribute to classification. The measure of one shows no discrimination power at all. A larger value indicates a higher contribution to classification between the two corresponding classes (Wold and Sjostrom 1977). Modeling power and discrimination power were calculated by Unscrambler ver. 9.7.

Partial Least Squares Discriminant Analysis

Partial least squares discriminant analysis (PLS-DA) is a variant of partial least squares regression. In the application of PLS-DA, each variety in the calibration set is assigned a dummy variable as a reference value. In the development of PLS-DA models, full cross-validation was used to evaluate the quality and to prevent over-fitting of calibration model. The optimal number of latent variables (LVs) was determined by the lowest value of predicted residual error sum of squares.

PLS-DA was used in order to classify panax according to the spectra (Giunchi et al. 2008). Loading weights and regression coefficients of PLS-DA models were used to propose the effective wavelengths of the spectra (Wu et al. 2008a). The loading weights were specific to PLS-DA and expressed by the information at each wavelength (X-variable) related to the variation of different variety numbers. The loading weights were normalized so that their lengths could be interpreted as well as their directions. Wavelengths (X-variables) with large absolute loading weight values were important for the discrimination of panax. The regression coefficients were calculated from the raw data table, and it calculated the response value Y-variables from the X-variables. The size of the coefficients gave an indication of which variables had the important impact on the response variables (Y). Its task was to find which variables were important for predicting Y-variable. Large absolute values indicated the importance and the significance of the effect on the prediction of Y-variable preference. PLS-DA was executed by Unscrambler ver. 9.7.

Genetic Algorithms

Genetic algorithms optimize search results for problems with large data sets. Effective wavelengths were determined which can distinguish different varieties of panax from each other. A fitness function, which acts as selective pressure on all of the data points, is required in genetic algorithms. It determines which data points get passed on to or removed from each

subsequent generation. In this study, a linear combination of (1) the a posteriori probability and (2) the empirical error rate of a linear classifier (classify) was used to evaluate how well the data got grouped. The numbers of groups and generations were adjusted for the best results. Selection, crossover, and mutation events generated a new population in every generation. Gaussian mutation was used to create the mutated children. The values for scale or shrink were specified as 0.5 and 0.75, respectively. Scale controls what fraction of the gene's range is searched. A value of 0 results in no change; a scale of 1 results in a distribution whose standard deviation is equal to the range of this gene. Shrink controls how fast the scale is reduced as generations go by. A shrink value of 0 results in no shrinkage, yielding a constant search size. A value of 1 results in scale shrinking linearly to 0 as genetic algorithm progresses to the number of generations specified by the options structure. Finally, the effective wavelengths were proposed after genetic algorithms. Genetic algorithms were executed by MATLAB 7.1 (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. (2002). The LS-SVM has the capability for linear and nonlinear multivariate calibration and solves the multivariate calibration problems in a relatively fast way (Suykens et al. 2002). It uses a linear set of equations instead of a quadratic programming problem to obtain the support vectors (Li et al. 2007). As a nonlinear function and a more compacted supported kernel, RBF 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 (Lukas et al. 2004; Hou et al. 2008). Thus, RBF kernel was used. Grid-search technique was applied to find out the optimal parameter values which include regularization parameter *gam* (γ) and the RBF 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 (Wu et al. 2008b). Furthermore, the effective wavelengths were proposed through the regression coefficients of LS-SVM (Wu et al. 2008a). LS-SVM was executed by MATLAB 7.1 (The Math Works, Natick, USA). LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) was applied with MATLAB to derive all of the LS-SVM models.

Binary Variety Number Encoding

In this study, each sample in the calibration set of PLS-DA, genetic algorithms, and LS-SVM was assigned a dummy variable as a reference value, which was an arbitrary number whether the sample belongs to a particular variety or not. PG was set as reference variety one, American PQ was set as two, and Chinese PQ was set as three. 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 (Allwein et al. 2000). In this study, three varieties were encoded in a codebook (using minimum output coding Suykens and Vandewalle 1999). Each variety was encoded into two numbers (−1 or 1) in two dimensions, respectively. Variety one was encoded into "−1, −1" for dimension one and two, respectively. Variety two was encoded into "−1, +1," and Variety three was encoded into "+1, +1."

Modeling

The whole Vis-SWNIR spectra, scores of obtained PCs or the spectra of selected effective wavelengths were used as the input matrix of the LS-SVM models, respectively. The outputs of the LS-SVM were encoded with binary matrix of classes corresponding to varieties in each model. Four LS-SVM models were established, where half randomly selected samples were used to training the model, and the remaining half were used for prediction.

Model 1 Species discrimination of Chinese PQ and PG which were from the same growing region (26 samples for each species).

Model 2 Species discrimination of PQ and PG (26 samples for PG and 52 samples for PQ, American PQ and Chinese PQ were set as the same specie).

Model 3 Species discrimination of American PQ and Chinese PQ which were from different growing region (26 samples for each variety).

Model 4 Species discrimination of American PQ, Chinese PQ and PG for all panax (total 78 samples, 26 samples for each variety).

Soft independent modeling of class analogy (SIMCA) method was also applied to evaluate the LS-SVM method. SIMCA is an elaborate method based on PCA (Sáiz-Abajo et al. 2004) and commonly used as a pattern recognition method (Woo et al. 2005; Chen et al. 2005; Sáiz-Abajo et al. 2004). SIMCA was executed by Unscrambler ver. 9.7.

Results and Discussion

Features of Vis-SWNIR Spectra

The average absorbance spectra and standard deviation curves (variance) described as log (1/R) of three varieties of

panax samples are shown in Fig. 2. As panax samples were from the same genus, there was no large spectral difference among them. American PQ and Chinese PQ were from the same species of *Panax quinquefolium*. L and had similar constituents while they were planted in different regions. Average spectral curves of two PQs were similar in shape, where the correlation coefficient was 0.9998. Spectral curves of PQ and PG average were less similar, where the correlation coefficients were 0.9842 and 0.9809 for PG average versus American PQ average and PG average versus Chinese PQ average, respectively. The modeling distances were 7.478, 40.271, and 181.259 for American PQ versus Chinese PQ, PG versus Chinese PQ, and PG versus American PQ, respectively. It could be seen that PG was closer to Chinese PQ than to American PQ, as PG and Chinese PQ grew in the same region.

PCA Results

PCA was executed before LS-SVM modeling to reduce original independent variables (650 wavelength points) into smaller new orthogonal variables with minor information loss. The accumulated contributions of PC1 to PC4 for all models exceed 99.90% except model 3. The score clustering plot of PC1, PC2, and PC3 of American PQ, Chinese PQ, and PG gave the information about variety clustering of panax samples (Fig. 3). The clustering result shows the borders of each clustering were not clear, which indicated only three PCs might not be enough for accurate discrimination.

PCA-LS-SVM Modeling

PCA-LS-SVM models built using different number of inputs (PCs) were evaluated using CCR. The CCRs of the LS-SVM prediction models with 1 to 6 PCs are shown in Fig. 4. The CCR increased quickly when the PC number increased from one to three PCs except model 2. When the number of PCs was more than four, the increase of CCR became slow. CCRs were 100% for each model based on three, three, four, and six PCs, respectively. The results show that all models had good performance, either for discriminating the species from the same growing region—Chinese (model 1) or for discriminating the different growing regions—American, Chinese of the same species (model 3).

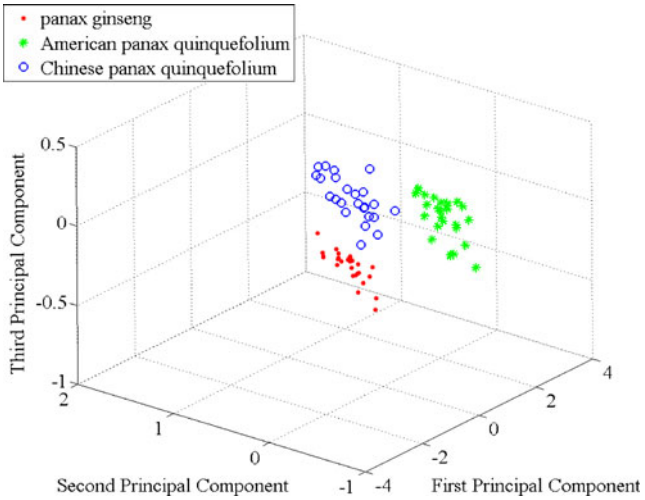

Fig. 3 Score clustering plot of PC1, PC2, and PC3 of American PQ, Chinese PQ, and PG

Selection of Effective Wavelengths

The method and procedure to select effective wavelengths were proposed and demonstrated taking model 4 as example. The effective wavelengths were selected based on a threshold as peaks or valleys of curves. Table 1 summarized the selection of effective wavelengths. The optimal number of loading weights was automatically determined as seven by Unscrambler from the minimum value of the PRESS by full cross validation.

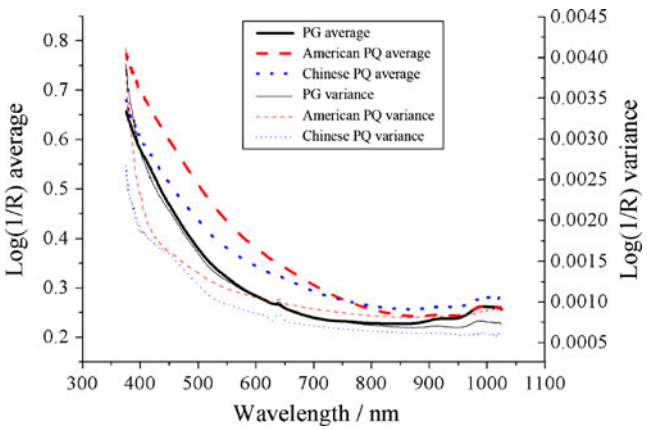

Fig. 2 Average spectra and variance curve related to panax samples at 650 points in the region of 376–1,025 nm

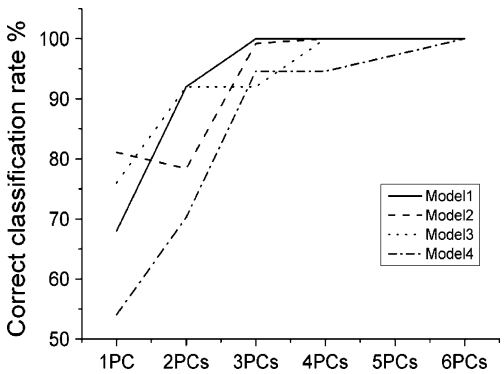

Fig. 4 Correct classification rate of prediction models of LS-SVM using different PCs

Table 1 Selection methods of effective wavelengths using different methods for model 4

Methods	Selection threshold	Effective wavelengths	Purpose
Modeling power	Higher than 0.96	407 and 649 nm	For PG discrimination model
		398, 492, and 758 nm	For American PQ discrimination model
		400, 510, and 783 nm	For Chinese PQ discrimination model
Discrimination power	Higher than 9.0	407, 626, and 909 nm	For PG versus Chinese PQ discrimination model
	Higher than 3.5	387, 733, and 955 nm	For American PQ versus Chinese PQ discrimination model
	Higher than 17.0	402, 523, 674, and 918 nm	For PQ versus American PQ discrimination model
Regression coefficients of PLS-DA	Absolute value larger than 6.0	397, 668, and 984 nm	For the whole discrimination model
Regression coefficients of LS-SVM	Absolute value larger than 2.0	397, 668, and 984 nm	For the whole discrimination model
Loading weights	With the largest positive and negative values of each weight	385, 515, 585, 670, and 984 nm	For the whole discrimination model
Genetic algorithms	/	547 and 1,001 nm	For dimension one (Table 1)
	/	715 and 956 nm	For dimension two (Table 1)

In addition, the overall discrimination power value of PG versus American PQ was the highest and that of American PQ versus Chinese PQ was the lowest. It indicated that American PQ and Chinese PQ were close. The difference between species with the same growing region (PG versus Chinese PQ) was more obvious than that between different growing regions (American PQ versus Chinese PQ).

Effective wavelengths of other three models were proposed by the same ways as used in model 4 (Table 2). For the same model, some effective wavelengths were both selected by different selection methods, or some effective wavelengths were quite close to each other. Take model 1 as an example, 400 nm was selected by both modeling power and regression power, while it was quite close to 407 nm which was

Table 2 Effective wavelengths of models proposed by different methods

Model	Method	Effective wavelength [nm]
1	Modeling power	400, 407, 510, 649, 783, and 896
	Discrimination power	407, 626, 660, and 910
	Regression coefficients	376, 400, 983, and 1025
	Loading weights	385, 585, 667, and 982
	Genetic algorithms	508, 660, and 751
2	Modeling power	403, 523, and 912
	Discrimination power	400, 402, 495, 512, 760, and 785
	Regression coefficients	405, 428, 528, 638, and 666
	Loading weights	386, 447, 1015, and 1023
	Genetic algorithms	449, 467, 541, 621, and 629
3	Modeling power	401, 409, 509, 651, 787, 898, and 983
	Discrimination power	403, 525, and 912
	Regression coefficients	397, 667, and984
	Loading weights	385, 503, and 985
	Genetic algorithms	512, 718, and 761
4	Modeling power	398, 400, 407, 492, 510, 649, and 783
	Discrimination power	387, 402, 407, 523, 626, 674, 733, 909, 918, and 955
	Regression coefficients	397, 668, and 984
	Loading weights	385, 515, 585, 670, and 984
	Genetic algorithms	547, 715, 956, and 1001

selected by both modeling power and discrimination power. Six hundred sixty nanometers was selected by both discrimination power and genetic algorithms, while it was quite close to 667 nm selected by loading weights. Some other effective wavelengths were quite close, such as 508 nm (genetic algorithms) and 510 nm (Modeling power) and 983 nm (regression coefficients) and 982 nm (loading weights). Although effective wavelengths selected according to different methods might not be completely the same, they have close discrimination accuracy. It might be because the selected wavelengths from one method were sufficient but not necessary condition for the discrimination. So, although some effective wavelengths, selected according to different selection methods, were different, they can obtain the similar results.

LS-SVM Models Using Selected Effective Wavelengths

The performances of PCA-LS-SVM, LS-SVM (effective wavelengths), LS-SVM (whole spectra), and SIMCA models for prediction are compared in Fig. 5. Model 2 for the species discrimination between PG and PQ was better than other models, which responded to the larger difference between different species than that between the same species from different growing regions. Overall, the results of model 3 were worse than model 1. The difference between PG and Chinese PQ both from the same growing region was larger than that between the same species. The region discrimination was harder than species discrimination. The discrimination between American PQ and Chinese PQ was more difficult than others, and the CCRs of model 4 were worse than model 2.

PCA-LS-SVM and LS-SVM with genetic algorithms obtained the best results with 100% of CCRs for all four models. SIMCA showed the worst results. For the methods using the whole spectral data, the prediction results of PCA-LS-SVM models were better than those of LS-SVM (whole spectra) and SIMCA models. Because of reduced inputs, the calculation of PCA-LS-SVM model was simpler and faster than LS-SVM (whole spectra).

Compared to PCA-LS-SVM method using the whole spectral data, LS-SVM (effective wavelengths) had reduced input data matrices using the spectra of effective wavelengths. The training time could be reduced by LS-SVM (effective wavelengths) because the training time increased with the square of the number of training samples and linearly with the number of variables (Chauchard et al. 2004). The performances of effective wavelengths selected by modeling power and discrimination power were similar. LS-SVM with regression coefficients did a second best prediction results than other three selection methods. Only one sample was wrongly discriminated for models 3 and 4, respectively. The best prediction performance was achieved based on genetic algorithms. Moreover, all the effective wavelength selection methods were executed based on artificial setting thresholds except genetic algorithms. This kind of artificial setting is subjective. The manual selection, especially the identification whether the wavelength was at certain peak, was time consuming. Thus, the artificial setting to select effective wavelength have many disadvantages, and genetic algorithms shows its objective to select effective wavelengths compared with other methods.

However, the results of this study are based on sliced panax fragments. The differentiation of whole panax should be further researched. Moreover, the range of samples with different properties is not enough. More robust models and improved comparison results would be obtained using wider range of samples with different properties.

Conclusion

The results of this study indicated that Vis-SWNIR spectroscopy was a feasible way to differentiate panax species. The

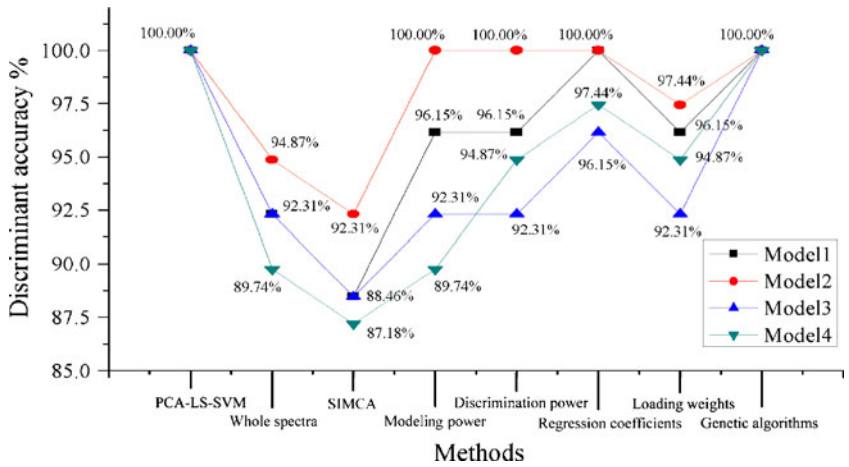

Fig. 5 Comparison of discrimination accuracy of PCA-LS-SVM, LS-SVM (whole spectra), SIMCA, and LS-SVM (effective wavelengths): LS-SVM (modeling power), LS-SVM (discrimination power), LS-SVM (regression coefficients), LS-SVM (loading weights), and LS-SVM (genetic algorithms)

distinguishing results of PCA-LS-SVM to differentiate species of panax samples can reach 100% CCR. The results showed that the vast spectra data can be effectively reduced and can remain useful information using PCA. The best LS-SVM (effective wavelength) model with 100% CCR was achieved through genetic algorithms. The selection process based on genetic algorithms is objective and has few manual works compared with other methods. In further study, a large number of panax samples with more different species, different manufactures, different growing regions and years, and different batches of one same species should be taken into consideration before industry online application.

Acknowledgements This study was supported by the National Science and Technology Support Program of China (2006BAD10A09, 2006BAD10A0711), 863 National High-Tech Research and Development Plan (2007AA10Z210), Natural Science Foundation of China (Project No: 30671213), Science and Technology Department of Zhejiang Province (Project No. 2005C12029). National Special Public Sector Research of Agriculture (200803037) and Science and Technology Department of Ningbo (Project No. 2007C10034).

References

Allwein, E. L., Schapire, R. E., & Singer, J. Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, *1*, 113–141. doi:10.1162/15324430152733133.

Bent, S., & Ko, R. (2004). Commonly used herbal medicines in the United States: A review. *American Journal of Medicine*, *116*, 478–485. doi:10.1016/j.amjmed.2003.10.036.

Bittner, A., Marbach, R., & Heise, H. M. (1995). Multivariate calibration for protein, cholesterol and triglycerides in human plasma using short-wave near-infrared spectrometry. *Journal of Molecular Structure*, *349*, 341–344. doi:10.1016/0022-2860(95)08779-U.

Chauchard, F., Cogdill, R., Roussel, S., Roger, J. M., & Bellon-Maurel, V. (2004). Application of LS-SVM to nonlinear phenomena in NIR spectroscopy: Development of a robust and portable sensor for acidity prediction in grapes. *Chemometrics and Intelligent Laboratory Systems*, *71*, 141–150. doi:10.1016/j.chemolab.2004.01.003.

Chen, J., Arnold, M. A., & Small, G. W. (2004). Comparison of combination and first overtone spectral regions for near infrared calibration models for glucose in aqueous solutions. *Analytical Chemistry*, *76*, 5405–5413. doi:10.1021/ac0498056.

Chen, Q., Zhao, J., Zhang, H., Liu, M., & Fang, M. (2005). Qualitative identification of tea by near infrared spectroscopy based on soft independent modelling of class analogy pattern recognition. *Journal of Near Infrared Spectroscopy*, *13*, 327–332.

Cozzolino, D., Smyth, H. E., & Gishen, M. J. (2003). Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins. *Journal of Agricultural and Food Chemistry*, *51*, 7703–7708. doi:10.1021/jf034959s.

Dasgupta, A., & Veras, E. (2006). Effectiveness of activated charcoal and equilibrium dialysis in removing Asian, American, Siberian and Indian ginseng from human serum. *Clinica Chimica Acta*, *367*, 144–149. doi:10.1016/j.cca.2005.12.005.

Fuzzati, S. N. (2004). Analysis methods of ginsenosides. *Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences*, *812*, 119–133.

Giunchi, A., Berardinelli, A., Ragni, L., Fabbri, A., & Silaghi, F. A. (2008). Non-destructive freshness assessment of shell eggs using FT-NIR spectroscopy. *Journal of Food Engineering*, *89*(2), 142–148. doi:10.1016/j.jfoodeng.2008.04.013.

Hou, T., Zhang, W., Case, D. A., & Wang, W. (2008). Deciphering the binding patterns of the peptide ligands of amphiphysin-1 SH3 domain by analyzing the molecular interaction fields. *Journal of Molecular Biology*, *376*, 1201–1214. doi:10.1016/j.jmb.2007.12.054.

Hu, C., & Kitts, D. D. (2001). Free radical scavenging capacity as related to antioxidant activity and ginsenoside composition of asian and North American ginseng extracts. *Journal of the American Oil Chemists' Society*, *78*, 249–255. doi:10.1007/s11746-001-0253-8.

Levis, W. H. (2003). *Medical botany* (p. 608). New York: Wiley.

Li, X., & He, Y. (2008). Evaluation of least squares support vector machine regression and other multivariate calibrations in determination of internal attributes of tea beverages. *Food Bioprocess Technology*. doi:10.1007/s11947-008-0101-y.

Li, W., Gu, C., Zhang, H., Awang, D. V. C., Fitzloff, J. F., Fong, H. H. S., et al. (2000). Use of high performance liquid chromatography-tandem mass spectrometry to distinguish Panax ginseng C.A. Meyer (Asian ginseng) and Panax quinquefolius L. (North American ginseng). *Analytical Chemistry*, *72*, 5417–5422. doi:10.1021/ac0006501.

Li, J. Z., Liu, H. X., Yao, X. J., Liu, M. C., Hu, Z. D., & Fan, B. T. (2007). Structure-activity relationship study of Oxindole-based inhibitors of Cyclin-dependent kinases based on least-squares support vector machines. *Analytica Chimica Acta*, *581*, 333–342. doi:10.1016/j.aca.2006.08.031.

Lu, G., Zhou, Q., Sun, S., Leung, K. S., Zhang, H., & Zhao, Z. (2008). Differentiation of Asian ginseng, American ginseng and Notoginseng by Fourier transform infrared spectroscopy combined with two-dimensional correlation infrared spectroscopy. *Journal of Molecular Structure*, *883–884*, 91–98. doi:10.1016/j.molstruc.2007.12.008.

Lukas, L., Devos, A., Suykens, J. A. K., Vanhamme, L., Howe, F. A., Majós, C., et al. (2004). Brain tumor classification based on long echo proton MRS signals. *Artificial Intelligence in Medicine*, *31*, 73–89. doi:10.1016/j.artmed.2004.01.001.

Luypaert, J., Massart, D. L., & Van der Heyden, Y. (2007). Near-infrared spectroscopy applications in pharmaceutical analysis. *Talanta*, *72*, 865–883. doi:10.1016/j.talanta.2006.12.023.

Mayes, D. M., & Callis, J. B. (1989). A photodiode-array-based near-infrared spectrophotometer for the 600–1100 nm wavelength region. *Applied Spectroscopy*, *43*, 27–32. doi:10.1366/0003702894202067.

Meng, F. Z., & Li, B. (2003). *Panax quinquefolium L.* Beijing: Science and Technology Publishing House.

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., et al. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, *46*, 99–118. doi:10.1016/j.postharvbio.2007.06.024.

Nocerino, E., Amato, M., & Izzo, A. A. (2000). The aphrodisiac and adaptogenic properties of ginseng. *Fitoterapia*, *71*, S1–S5. doi:10.1016/S0367-326X(00)00170-2.

Otsuka, M. (2004). Comparative particle size determination of phenacetin bulk powder by using Kubelka-Munk theory and principal component regression analysis based on near-infrared spectroscopy. *Powder Technology*, *141*, 244–250. doi:10.1016/j.powtec.2004.01.025.

Reeves III., J. B. (1994). Effects of water on the spectra of model compounds. *Journal of Near Infrared Spectroscopy*, *2*, 199–212.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*, 683–700. doi:10.1016/j.jpba.2007.03.023.

Sáiz-Abajo, M. J., González-Sáiz, J. M., & Pizarro, C. J. (2004). Near infrared spectroscopy and pattern recognition methods applied to the classification of vinegar according to raw material and elaboration process. *Journal of Near Infrared Spectroscopy*, *12*, 207–219.

Shao, Y., & He, Y. (2009). Measurement of soluble solids content and pH of yogurt using visible/near infrared spectroscopy and chemometrics. *Food Bioprocess Technology.* doi:10.1007/s11947-008-0180-9.

Sinija, V. R., & Mishra, H. N. (2009). FTNIR spectroscopic method for determination of moisture content in green tea granules. *Food Bioprocess Technology.* doi:10.1007/s11947-008-0149-8.

Suykens, J. A. K., & Vandewalle, J. (1999). Multiclass least squares support vector machines. In Proc. the Int. Joint Conf. on Neural Networks (IJCNN'99), Washington, DC. 900–903.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific Publishing.

Vanhaelen-Fastré, R. J., Faes, M. L., & Vanhaelen, M. H. J. (2000). High-performance thin-layer chromatographic determination of six major ginsenosides in Panax ginseng. *Journal of Chromatography. A*, *868*, 269–276. doi:10.1016/S0021–9673(99)01253–4.

Wold, S., & Sjostrom, M. (1977). In B. R. Kowalski (Ed.), *Chemometrics: Theory and applications ACS symposium. Ser. No. 52* (p. 243). Washington DC: American Chemical Society.

Woo, Y. A., Kim, H. J., Ze, K. R., & Chung, H. (2005). Near-infrared (NIR) spectroscopy for the non-destructive and fast determination of geographical origin of Angelicae gigantis Radix. *Journal of Pharmaceutical and Biomedical Analysis*, *36*, 955–959. doi:10.1016/j.jpba.2004.08.037.

Woodcock, T., Fagan, C. C., O'Donnell, C. P., & Downey, G. (2008). Application of near and mid-Infrared spectroscopy to determine cheese quality and authenticity. *Food Bioprocess Technology*, *1* (2), 117–129. doi:10.1007/s11947-007-0033-y.

Wu, D., He, Y., & Feng, S. (2008a). Short-wave near-infrared spectroscopy analysis of major compounds in milk powder and wavelength assignment. *Analytica Chimica Acta*, *610*(2), 232–242. doi:10.1016/j.aca.2008.01.056.

Wu, D., He, Y., Feng, S., & Sun, D. W. (2008b). Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM. *Journal of Food Engineering*, *84*, 124–131. doi:10.1016/j.jfoodeng.2007.04.031.

Xiccato, G., Trocino, A., Tulli, F., & Tibaldi, E. (2004). Prediction of chemical composition and origin identification of european sea bass (Dicentrarchus labrax L.) by near infrared reflectance spectroscopy (NIRS). *Food Chemistry*, *86*(2), 275–281. doi:10.1016/j.foodchem.2003.09.026.