



Prediction of crude protein and oil content of soybeans using Raman spectroscopy[☆]

Hoonsoo Lee^a, Byoung-Kwan Cho^{a,*}, Moon S. Kim^b, Wang-Hee Lee^a, Jagdish Tewari^c, Hanhong Bae^d, Soo-In Sohn^e, Hee-Youn Chi^f

^a Department of Biosystems Machinery Engineering, Chungnam National University, Daejeon, South Korea

^b Environmental Microbial and Food Safety Laboratory, USDA-ARS, Beltsville, MD 20705, USA

^c Fiber Science, Cornell University, Ithaca, NY, USA

^d School of Biotechnology, Yeungnam University, Gyeongsan, South Korea

^e Biosafety Division, National Academy of Agricultural Science, RDA, Suwon, South Korea

^f SBLab Co., Suwon, South Korea

ARTICLE INFO

Article history:

Received 9 January 2013

Received in revised form 22 April 2013

Accepted 23 April 2013

Available online 15 May 2013

Keywords:

Raman spectroscopy

Oil content

Protein content

Soybean

Partial least squares analysis

ABSTRACT

While conventional chemical analysis methods for food nutrients require time-consuming, labor-intensive, and invasive pretreatment procedures, Raman spectroscopy can be used to measure a variety of food components rapidly and non-destructively without supervision from experts once the instrument has been calibrated. The purpose of this study was to develop an optimal prediction model for determining the protein and oil contents of soybeans using a dispersive Raman spectroscopy method. In general, the crude oil content of soybeans is chemically determined using the Soxhlet extraction method, while the semimicro-Kjeldahl method and an auto protein analyzer have been used to assess crude protein content. In the present study, Raman spectra were measured in the 200–1800 cm⁻¹ wavenumber range and partial least squares (PLS) analysis methods were used to develop optimal models for predicting the crude protein and oil contents of soybeans. The resultant PLS models that used the effective wavenumber regions determined by intermediate PLS (iPLS) method were better than those models developed using the entire wavenumber range under investigation. The R_p^2 and SEP of the optimal PLS model for crude protein content were 0.916 and 0.636%, respectively. Likewise, the R_p^2 and SEP for crude oil content were 0.872 and 0.759%, respectively. The result suggests that the conventional Raman techniques investigated in this study can be applied to the prediction of soybean crude protein and oil content.

© 2013 The Authors. Published by Elsevier B.V. All rights reserved.

1. Introduction

In 2011, approximately 261 million tons of soybeans were produced worldwide. This represents approximately 10.1% of the total cereal production in 2010 [1]. Soybeans are an important food source for humans and animals and contain considerably more protein and oil than many other crops [2]. The protein and oil content of soybeans are considered to be important quality indicators and are announced immediately before packaging and distribution. Thus, the classification of soybeans could lead consumers to selectively take nutritionally high valued diets. Kjeldahl and Soxhlet extraction methods are widely used as standard methods of

measuring the crude protein and oil content of crops. However, these methods have several disadvantages, as they are time-consuming, labor-intensive, and destructive procedures that hamper the fast and economical quality evaluation of mass-produced soybeans. Therefore, the demand for rapid, robust, and nondestructive measurement methods for the main nutrient components of crops, such as the protein, oil, carbohydrate, water, and ash content of soybeans has recently increased.

Since the 1970s, near infrared reflectance spectroscopy has been widely used to measure the quality of agriculture and food materials. This noninvasive spectroscopic technique can rapidly provide physical and chemical information about specimens. Near infrared diffuse reflectance spectroscopy and near infrared transmittance spectroscopy have been used to predict the fatty acid composition of soybeans [3,4]. Baye et al. used near-infrared spectroscopy to predict maize seed composition [5] and Choung et al. employed near infrared reflectance spectroscopy to determine the protein and oil content of soybeans [6]. However, combination and overtone absorption bands in the near infrared region that are caused

[☆] This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author. Tel.: +82 42 821 6715; fax: +82 42 823 6246.

E-mail addresses: chobk@cnu.ac.kr, bx195@gmail.com (B.-K. Cho).

by O–H, C–H, and N–H functional groups result in broad overlapping spectra and the sensitivity of these techniques is limited to approximately 0.1% for trace components. Hence, a novel technique is necessary to improve accuracy and precision. As an alternative, Raman spectroscopy technique has emerged as a novel technology that could compensate for the disadvantage of NIR spectroscopy such as low sensitivity for minor components, providing high accuracy and precision [7]. Unlike near infrared spectroscopy, Raman spectroscopy provides detailed information about molecular vibrations. Extremely high sensitivity has been demonstrated for Raman spectroscopy in various fields [8]. This method involves the analysis of Raman scattering, which is observed in samples when they are illuminated by a strong light source. There are two types of Raman spectroscopy techniques that are of interest for the evaluation of agricultural commodities. One is dispersive Raman spectroscopy and the other is Fourier-transform (FT) Raman spectroscopy. Dispersive Raman spectroscopy typically uses various visible lasers, such as 473 nm, 532 nm and 633 nm lasers, and 785 nm near infrared (NIR) laser. In contrast, FT-Raman spectroscopy has used 1024 nm laser generally to minimize strong fluorescence signal from agro-product [9], preventing it from directly measuring the nutrient content of agricultural products. For this reason, many researchers have turned to interferometer-based NIR FT-Raman systems that cause less fluorescence for agriculture products than the dispersive Raman systems. The NIR FT-Raman spectroscopy has been used to predict the protein and apparent amylose contents of rice, determine milk fat content, monitor and quantify ethyl esters in soybean oil, and distinguish between oil and fat [10–13]. However, the large size of FT-Raman system due to interferometer is not suitable for recent demands on compact Raman system in the field of agriculture.

To construct a compact Raman system which satisfies high sensitivity and low fluorescence emission, the strong fluorescence signal must be removed from conventional Raman data, particularly for dispersive Raman system. Recently, various mathematical methods, such as wavelet transformation, Fourier transformation, and polynomial fitting, have been developed to banish the strong fluorescence background signal caused by dispersive Raman spectroscopy [14]. Among the algorithms, the polynomial fitting method is popular because of its rapid and simple nature [15]. This method has been applied to dispersive Raman spectroscopy at 785 nm laser in order to detect lycopene changes in tomatoes [9].

As dispersive Raman data is available by reducing the fluorescence, the objective of this study is to develop an optimal prediction model for determining the protein and oil content of soybeans using data sets obtained by dispersive Raman spectroscopy. We aim at using PLS algorithm to use Raman spectra processed by polynomial fitting method in determination of the optimal model. Because high fluorescence limited the use of dispersive Raman system at the expense of high sensitivity, this study is to invent a customized mathematical method available for Raman spectra so that Raman spectroscopy can be applied to inspect and obtain accurate information for a specific agricultural product.

2. Materials and methods

2.1. Samples and reference analysis

A total of 45 types of soybean sample were purchased from retail stores and 50 types of soybeans were provided by the National Institute of Crop Science in Korea. Of all sample, the color of twenty-five samples were black and remain seventy samples were yellow. The average and standard deviation of the diameter of soybeans used in this experiment were 0.72 cm and 0.12 cm, respectively.

Table 1 illustrates indicate In addition, the range of moisture content in soybean was between 6.13% and 13.44% (**Table 1**). Each sample was milled using a grain grinder and passed through a 250- μ m-mesh sieve. Sieved samples were stored in plastic bottles and spectra were obtained using Raman spectroscopy. Crude protein was determined using the semimicro-Kjeldahl method and an auto protein analyzer (Kjeltec 2400 auto-analyzer, Hillerød, Denmark). The ground soybean sample was weighed 1 g accurately and nitrogen to protein conversion factor of 5.71 was used. In order to obtain crude oil content in soybean, weighted 1 g sample were put a dish and were dried during 2 h at 105 °C using drying oven. Then, crude oil content was determined by extraction using diethyl ether and a Soxhlet extractor (Soxtec System HT 1043 extraction unit, Höganäs, Sweden).

2.2. Raman spectroscopy

A 400-mW diode laser with a light source at 785 nm was used for all Raman measurements. The Raman system (Kaiser Optical Systems, Ann Arbor, MI) consisted of a charge-coupled device (CCD) detector and a holographic transmission grating. The spectral range was set to 200 cm^{-1} –1800 cm^{-1} in 0.3 cm^{-1} intervals and the laser point diameter was set to 3 mm. The same amount of each sample (2 g) was placed in a standard 96-hole wall plate and the plate moved automatically to align each sample with the Raman system prior to each measurement. For each sample, the spectrum measurement with 1-s exposure time was repeated 64 times for each sample and then the average spectrum was used for a representative spectrum for each sample.

2.3. Data analysis

The Raman scattering signal was generated by using a strong laser light source. Biological materials, such as agricultural products, may emit strong fluorescence signals that mask the characteristic Raman scattering signal. This problem has been considered to be a challenge for Raman spectroscopy [9]. In the present study, a widely used polynomial fitting method was employed to analyze Raman spectrum data and to correct for fluorescence because this method is efficient and simple [15]. Polynomial fitting involves determining the proper order polynomial for obtaining a baseline through iterative calculation. Various order of polynomial such as 4th, 5th, 8th, 12th, and 16th was test for fitting the spectrum data based on the previous studies that used 5th order polynomial for soybeans [7] and 8th order polynomial equation for lycopene in tomatoes [9]. Finally, a 16th order polynomial equation and the 100th iteration were employed to create the fluorescence correction baseline because of its best prediction than any other polynomials.

The corrected Raman spectra were further subjected to 8 preprocessing methods, including smoothing, mean normalization, maximum normalization, range normalization, multiplicative scatter correction (MSC), standard normal variate (SNV), Savitzky–Golay 1st derivative, and Savitzky–Golay 2nd derivative methods. The preprocessed spectra were used to develop an optimal partial least squares (PLS) model described below for prediction of the protein and oil content of soybeans.

2.4. Partial least squares (PLS) analysis

PLS was the main algorithm used to obtain the prediction models for the crude protein and oil content of soybeans. PLS is a multivariate analysis method that extracts new latent variables from raw spectra. It can compress the large amount of spectral data into a new structure known as latent variables or factors and these latent variables are able to describe the maximum covariance

Table 1
Reference values for soybean sample components.

Components	Data set	Mean (%)	Standard deviation (%)	Minimum (%)	Maximum (%)
Crude protein	Calibration	34.10	2.17	30.24	40.34
	Validation	34.19	2.44	30.30	41.21
Crude oil	Calibration	14.42	2.07	9.97	18.21
	Validation	14.45	2.08	9.83	18.62
Moisture	Entire data	9.19	2.03	6.13	13.44
Crude ash		5.24	0.50	4.26	6.27
Carbohydrate		37.54	1.89	33.21	44.92

Crude oil content in soybean consists of triacylglycerol (94.4%), phospholipids (3.7%), and unsaponifiable matter (1.3–1.6%) [5]. The triacylglycerol and phospholipids, major constitution in soybean, were considered as the oil in this study.

between reference oil value and the spectral data [16]. Equations of PLS method are as follows.

$$X = TP^T + E \quad (1)$$

$$Y = UC^T + F \quad (2)$$

$$U = TB + G \quad (B = (T^T T)^{-1} T^T U) \quad (3)$$

where X , and Y are spectra data, and protein and oil content in soybean, respectively. The T and U are score matrices projected on linear combinations and the P and C are loadings matrices. The matrices of E and F represent error matrices for X and Y data. After applying PCA at X and Y matrices like [Eqs. (1) and (2)], construct inner relation between spectra data and protein and oil content by using least squares [Eq. (3)] [17].

The calibration and validation data sets were created by sorting the samples in the order of protein and oil content so that the data sets had a similar distribution of protein and oil content. For instance, first reference values of entire protein and oil content were extracted as validation data sets, while the second and third reference values were then selected as calibration data set. With this way, the calibration and validation data sets were extracted for entire reference data. As a results, the soybean flour samples were separated into a calibration data set ($n = 65$) and a validation data set ($n = 30$).

Cross validation was used to select the optimum number of regression factors and to avoid overfitting [18]. Detail theory of full cross validation is described in elsewhere [19]. The minimum root mean square error of validation (RMSEV) is given by the expression.

$$RMSEV = \sqrt{\frac{\sum_{i=1}^n (y_v - y_{ref})^2}{n}} \quad (4)$$

where y_v and y_{ref} are the estimated value for developed PLS model and reference value by standard methods, respectively, n is the number of spectrum [20].

Previously researchers have been reported that selected spectral regions are able to improve the performance of developed model using entire spectra range. In this study, In order to find an effective wavenumber regions contributed to the performance improvement for calibration model, the intermediate PLS (iPLS) method described by Norgaard et al. [21] was used. Wavenumbers between 200 cm^{-1} and 1800 cm^{-1} were divided into 32 sections in 50 cm^{-1} intervals and PLS results for individual wavenumber regions were obtained. Regions with the RMSEV that were lower than the average of the entire subinterval were selected as effective wavenumber regions. The selected effective wavenumber regions were the main source for developing the optimal PLS model. Removal of the fluorescence baseline, pre-processing, iPLS, and PLS analyses were performed using PLS toolbox in MATLAB (version 2009a, Mathworks, Natick, MA, USA).

3. Results and discussion

3.1. Reference and spectral analysis

Table 1 shows the reference values used for developing the optimal model for determining the oil and protein content of soybean flour. The mean oil contents of calibration and validation data sets were 14.42% and 14.45%, respectively, similar to the general oil contents of Korean soybean [22]. The standard deviation, minimum, and maximum values for oil were similar between the calibration and validation data sets. Protein ranges of the calibration and validation data sets were 30.24–40.34% and 30.3–41.21%, respectively.

Fig. 1A shows the original Raman spectra from 95 soybean flour samples. There was a large intensity variation due to the fluorescence background. A 16th-order polynomial equation was used to remove the fluorescence signal from the original soybean flour Raman signals. Fig. 1B contains the same spectra after removal of the fluorescence signal, clearly displaying the main peaks unlike in Fig. 1A. These peaks represent the various constituents of soybeans. Previous studies indicate that glutamic acid and phenylalanine are major amino acids, accounting for approximately 18% and 5% of soybean crude protein content [23]. The Raman peaks at 1003 cm^{-1} , 1517 cm^{-1} , and 1605 cm^{-1} are consistent with the peaks for glutamic acid and phenylalanine reported by Zhu et al. [24]. In addition, peaks at 1078 cm^{-1} and 1124 cm^{-1} were affected by miscellaneous skeletal structures such as C–C, C–N, and C–O [8], and the peak at 1657 cm^{-1} was associated with amide I bonds [25]. The peaks at 1264 , 1303 , 1445 , and 1745 cm^{-1} were related to the oil content of the samples and agree with those reported in previous Raman spectra studies [11–13,24]. These peaks were associated with double bonds in fatty acid molecules [11].

3.2. PLS calibration models for soybean protein content

The advantages of removing the fluorescence baseline from the original Raman scatter signals include improvement of prediction performance by the PLS model and identification of unique characteristics of the Raman peaks. To confirm this, PLS results of original Raman spectra and Raman spectra modified by removal of the fluorescence signal are compared (Table 2). The R_p^2 and standard error of prediction (SEP) of modified Raman spectra were 0.870% and 0.770%, respectively, resulting that the modified Raman spectra yielded better PLS results than the original spectra. This result demonstrates that the fluorescence signal influences the performance of the PLS model, which proves the enhanced predictive performance with the modified Raman spectra. Therefore, the spectra modified using a 16th-order polynomial equation was used to develop a PLS model for determination of soybean protein.

When developing a PLS model, it is important to determine the effective spectral ranges and number of PLS factors [26]. In the present study, spectral ranges and effective wavenumber regions were selected using iPLS algorithms by comparing the RMSEV, and

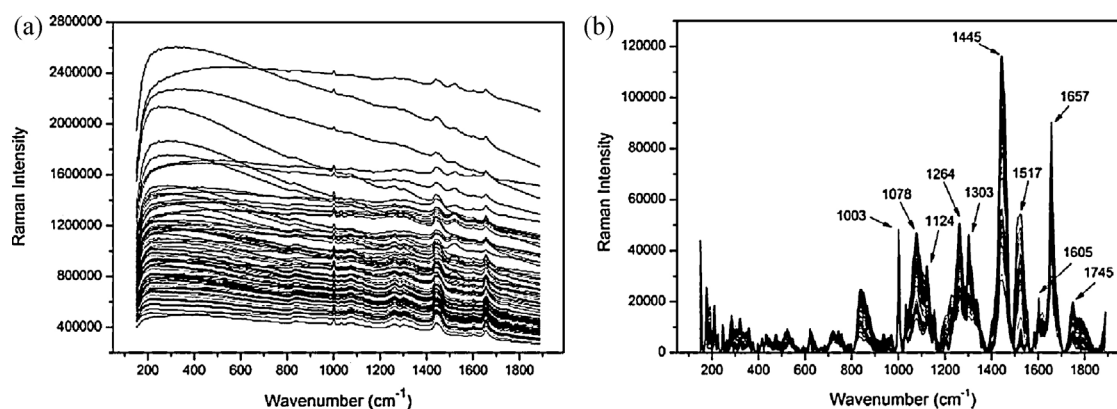


Fig. 1. Original Raman spectra (a) and Raman spectra modified by removal of the fluorescence signal using a polynomial equation (b) for 95 soybean flour samples.

Table 2

Comparison of PLS models for estimation of protein constructed using original Raman spectra and Raman spectra modified by removal of the fluorescence signal.

Spectra	R^2_c	SEC (%)	Factors	R^2_p	SEP(%)	Bias
Original	0.909	0.662	10	0.740	1.408	−0.021
Modified	0.915	0.684	6	0.870	0.770	−0.094

the minimum RMSEV value was used to determine the number of effective factors for PLS model. The Raman spectra in this study were divided into 32 equal intervals, each approximately 50 cm^{-1} intervals.

Fig. 2 shows the results of selection of effective wavenumber region using the RMSEV values for protein. Effective wavenumber regions were those with RMSEV values that were lower than the average RMSEV value for the entire range. The average RMSEV value was 1.57% for protein and there were 8 effective wavenumber regions. The R^2 values for the selected effective wavenumber regions were higher than those of other regions. The selected effective wavenumber regions were: 450 cm^{-1} –500 cm^{-1} , 600 cm^{-1} –650 cm^{-1} , 950 cm^{-1} –1050 cm^{-1} , 1150 cm^{-1} –1200 cm^{-1} , 1350 cm^{-1} –1450 cm^{-1} , and 1650 cm^{-1} –1700 cm^{-1} . These regions include 1003, 1124, 1303, and 1657 cm^{-1} and are in good accord with regions of interest in Fig. 1B.

Table 3 shows the results of PLS models for prediction of protein. Use of the entire wavenumber range containing 8 PLS factors resulted in a maximum of R^2_p value of 0.918 and a minimum SEP value of 0.633% for the PLS model preprocessed using SNV. Then, we constructed PLS models using effective wavenumber regions

selected by the iPLS algorithm, and they were comparable to those of PLS models developed using the entire wavenumber range. The highest R^2_p (0.916) and the lowest SEP (0.636%) value were obtained for this PLS models which is also preprocessed using SNV. In addition, the number of PLS factors for model applied SNV method was the same either for models constructed using effective wavenumber regions or those built by using the entire wavenumber range. These results suggest that optimal PLS models for detection of protein in soybeans can be constructed using effective wavenumber regions based on RMSEV.

The optimum number of PLS factors was determined by selecting the model with the minimum RMSEV value. Fig. 3 shows the RMSEV values for model developed using the effective wavenumber regions that contain 1–20 PLS factors. The minimum RMSEV value results from the 8-factor PLS model, and increases with additional PLS factors, demonstrating that the optimal number of PLS factors was 8. The protein content predicted by the optimal PLS model and the measured protein values for the soybean samples are shown in Fig. 4. Both calibration and prediction data sets were strongly correlated with measured values, indicating predictability of the developed 8-factor PLS model for the protein contents.

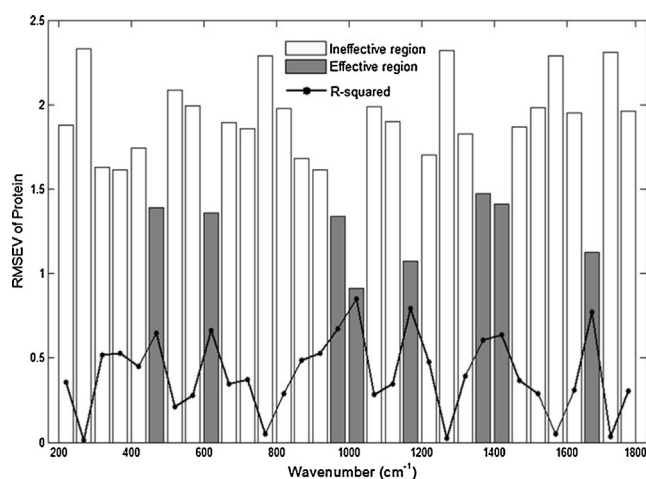


Fig. 2. Effective wavenumber region selection using RMSEV for protein.

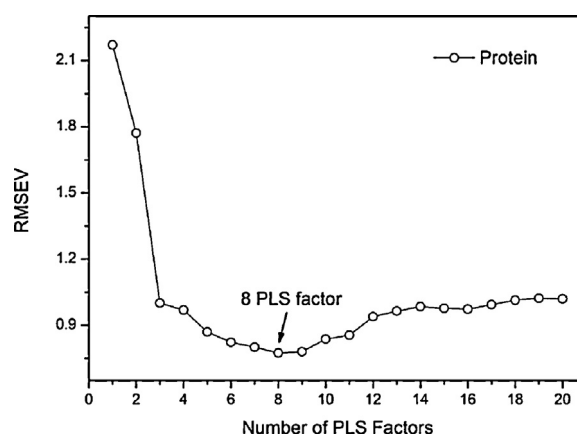


Fig. 3. The root mean square error of validation for each number of partial least squares factors added to a model for prediction of soybean protein content.

Table 3

Results of PLS models constructed to estimate the protein content of soybeans, presented by wavenumber range of Raman spectra.

Wavenumber (cm ⁻¹)	Preprocessing	R _c ²	SEC (%)	Factors	R _p ²	SEP (%)	Bias
200–1800	Smoothing	0.949	0.528	8	0.892	0.704	−0.074
	Mean normalization	0.984	0.292	10	0.899	0.697	−0.064
	Maximum normalization	0.952	0.512	7	0.859	0.803	−0.005
	Range normalization	0.952	0.512	7	0.859	0.803	−0.005
	MSC	0.979	0.341	9	0.880	0.742	0.001
	SNV	0.973	0.389	8	0.918	0.633	0.022
	Savitzky_Golay_1st	0.999	0.091	8	0.904	0.704	−0.085
	Savitzky_Golay_2nd	0.996	0.148	6	0.760	1.174	−0.040
	Raw	0.950	0.525	8	0.892	0.704	−0.074
Effective wavenumber regions based on RMSEV	Smoothing	0.990	0.241	13	0.842	0.855	−0.018
	Mean normalization	0.988	0.252	12	0.895	0.697	0.071
	Maximum normalization	0.948	0.536	7	0.857	0.819	0.048
	Range normalization	0.953	0.507	8	0.845	0.845	0.024
	MSC	0.977	0.360	10	0.842	0.860	0.851
	SNV	0.959	0.474	8	0.916	0.636	0.047
	Savitzky_Golay_1st	0.975	0.373	5	0.880	0.742	−0.203
	Savitzky_Golay_2nd	0.937	0.591	4	0.676	1.216	−0.441
	Raw	0.990	0.235	13	0.841	0.859	−0.024

Table 4

Comparison PLS models constructed to estimate the oil content of soybeans using original Raman spectra and Raman spectra modified by removal of the fluorescence signal.

Spectra	R _c ²	SEC (%)	Factors	R _p ²	SEP (%)	Bias
Original	0.794	0.945	10	0.698	1.202	0.007
Modified	0.816	0.895	5	0.820	0.903	−0.041

3.3. PLS calibration models for soybean oil content

The results of the PLS model for prediction of crude oil content using original Raman spectra and Raman spectra modified by removal of the fluorescence signal are shown in Table 4. The R_p² of the PLS model constructed using original Raman spectra was lower, and the SEP value was higher than those of the PLS model constructed using modified Raman spectra. Therefore, Raman spectra modified using a 16th-order polynomial equation were used to develop an optimum PLS model for determining the crude oil content of soybeans. The same iPLS algorithm and approach in previous section were used to select the effective wavenumber regions for crude oil model development.

Fig. 5 illustrates the RMSEV value for each wavenumber region. The average RMSEV value was 1.61% and 17 regions with RMSEV values were below the average with relatively high R² values. The lowest RMSEV value was observed at the effective region between

1300 cm⁻¹ and 1350 cm⁻¹. Of the 17 selected effective regions, 1400 cm⁻¹–1450 cm⁻¹ and 1250 cm⁻¹–1350 cm⁻¹ are consistent with the region that indicates the presence of double bonds in fatty acid molecules (Fig. 1B).

The resultant PLS models that were constructed to predict the oil content of soybeans are presented in Table 5. When the entire wavenumber range was used, the best models resulted in an R_p² of 0.835 and SEP of 0.865%, and were constructed using maximum normalization and range normalization preprocessing methods. Only with the effective wavenumber ranges, the best model developed using MSC preprocessing exhibited an R_p² of 0.872 and SEP of 0.759%. The better results were obtained with the use of effective wavenumber ranges that the entire range when developing the model. Therefore, the optimal PLS models were achievable with effective wavenumber regions selected based on RMSEV values to determine soybean oil content.

The RMSEV value greatly decreased with the addition of the 5th PLS factor, but the RMSEV values again increased with the addition of the 6th PLS factor (Fig. 6). Because the minimum RMSEV

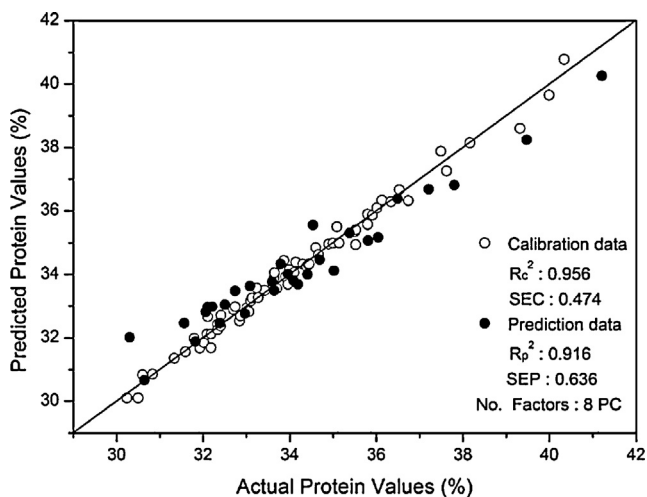


Fig. 4. Scatter plot of soybean protein values predicted using an optimal partial least squares model of Raman spectra and measured soybean protein values.

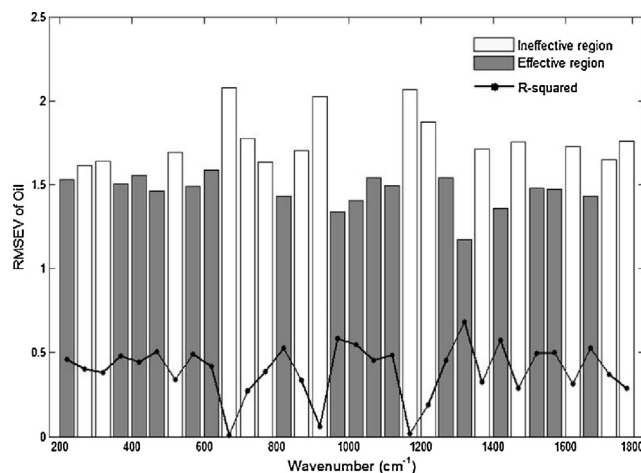
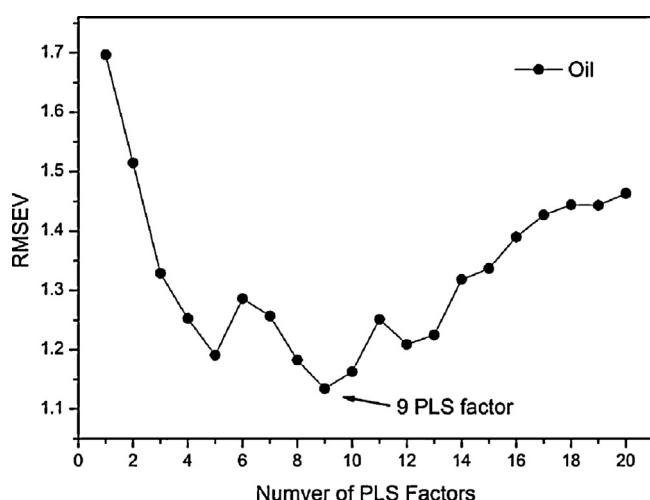
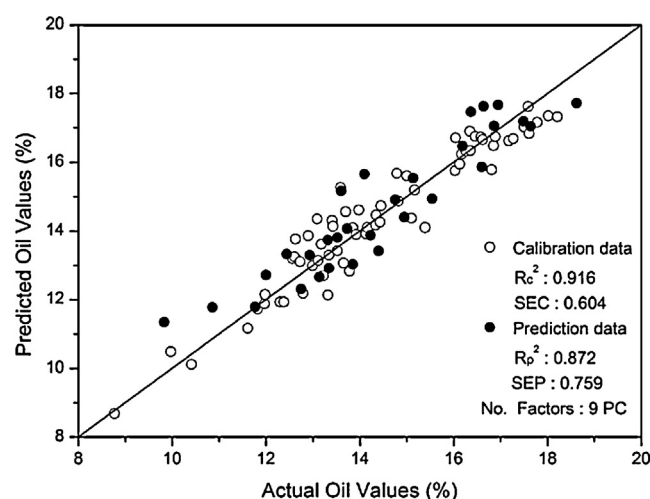


Fig. 5. Effective wavenumber region selection using RMSEV for oil.

Table 5

Results of PLS models constructed to estimate the oil content of soybeans, presented by wavenumber range of Raman spectra.

Wavenumber (cm ⁻¹)	Preprocessing	R^2_c	SEC (%)	Factors	R^2_p	SEP (%)	Bias
200–1800	Smoothing	0.795	0.943	4	0.799	0.965	−0.001
	Mean normalization	0.796	0.942	5	0.830	0.879	0.019
	Maximum normalization	0.801	0.928	6	0.835	0.865	0.024
	Range normalization	0.801	0.928	6	0.835	0.865	0.024
	MSC	0.788	0.959	5	0.828	0.879	−0.022
	SNV	0.764	1.013	4	0.827	0.899	0.048
	Savitzky_Golay_1st	0.962	0.404	5	0.810	0.954	0.034
	Savitzky_Golay_2nd	0.959	0.423	3	0.688	1.222	0.205
	Raw	0.795	0.942	4	0.798	0.966	−0.001
Effective wavenumber region based on RMSEV	Smoothing	0.775	0.989	3	0.767	1.041	−0.006
	Mean normalization	0.947	0.478	11	0.801	0.946	0.166
	Maximum normalization	0.936	0.526	11	0.863	0.787	0.184
	Range normalization	0.766	1.009	5	0.817	0.922	0.013
	MSC	0.916	0.604	9	0.872	0.759	0.174
	SNV	0.783	0.970	5	0.831	0.875	0.031
	Savitzky_Golay_1st	0.949	0.472	5	0.846	0.866	0.041
	Savitzky_Golay_2nd	0.782	0.972	2	0.682	1.225	0.258
	Raw	0.775	0.988	3	0.767	1.041	−0.006

**Fig. 6.** The RMSEV for each number of partial least squares factors added to a model for prediction of soybean oil content.**Fig. 7.** Scatter plot of soybean oil values predicted using an optimal partial least squares model of Raman spectra and measured soybean oil values.

value was associated with the 9th PLS factor, the optimal PLS model included 9 PLS factors for predicting soybean oil content. The RMSEV values increased with each additional PLS factor above the 9th PLS factor, suggesting the use of more than 9 PLS factors would over-fit the Raman data.

Fig. 7 illustrates the agreement of the measured oil content of soybean samples with the predicted oil content in the same samples using the optimal model. The distribution of the prediction data is similar to that of the calibration data. The variation of the oil content predicted by the model was slightly higher than that of the protein content prediction model.

4. Conclusions

In this study, a conventional dispersive Raman spectroscopy and suggested optimal PLS model have been successfully used to predict the protein and oil content of soybeans. Modification of Raman spectra by removing the fluorescence background signal produced better PLS models compared to the direct use of the raw Raman spectra. With modified Raman data, the optimal PLS model was developed with the effective wavenumber regions selected by iPLS algorithm, and reported better predictability than the model including the entire wavenumber range. The R^2_p and SEP were 0.916 and 0.636%, respectively, for the protein prediction model constructed using SNV-preprocessed data. The R^2_p and SEP were 0.872 and 0.759%, respectively, for the oil prediction model constructed using MSC-preprocessed data. The PLS result for the prediction of protein content was slightly better than that for the prediction of oil content. The Raman-based PLS results in this study were slightly lower than those previously reported for NIR by others. However, NIR techniques are highly susceptible to moisture contents and conventional Raman techniques can be used to predict both minor and major constituents of various crop grains with minimal water interference. Moreover, Raman technique has a potential to detect for trace of contamination on and classify origin for agro-products [27]. Future work will evaluate to determinate minor component in crops such as vitamin and inorganic component.

Acknowledgements

Financial support for this work was provided by the Next-Generation Bio-Green-21 program (No. PJ008055), Rural Development Administration (RDA), Republic of Korea. Authors would like to appreciate Dr. Kangjin Lee, Dr. Sukwon Kang and

Mr. Changyeun Mo in RDA, Republic of Korea for helping us to use Raman spectroscopy.

References

- [1] FAOSTAT, Database on cereal production, 2011.
- [2] D.L. Pazdernik, A.S. Killam, J.H. Orf, Analysis of amino and fatty acid composition in soybean seed, using near infrared reflectance spectroscopy, *Agronomy Journal* 89 (1997) 679–685.
- [3] S.R. Delwiche, K.S. McKenzie, B.D. Webb, Quality characteristics in rice by near-infrared reflectance analysis of whole-grain milled samples, *Cereal Chemistry* 73 (1996) 257–263.
- [4] A.G. Patil, M.D. Oak, S.P. Taware, S.A. Tamhankar, V.S. Rao, Nondestructive estimation of fatty acid composition in soybean [*Glycine max* (L.) Merrill] seeds using near-infrared transmittance spectroscopy, *Food Chemistry* 120 (2010) 1210–1217.
- [5] T.M. Baye, T.C. Pearson, A.M. Settles, Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy, *Journal of Cereal Science* 43 (2006) 236–243.
- [6] M.G. Choung, I.Y. Baek, S.T. Kang, W.Y. Han, D.C. Shin, H.P. Moon, et al., Determination of protein and oil contents in soybean seed by near infrared reflectance spectroscopy, *Korean Journal of Crop Science* 46 (2001) 106–111.
- [7] M.V. Schulmerich, M.J. Walsh, M.K. Gelber, R. Kong, M.R. Kole, S.K. Harrison, et al., Protein and oil composition predictions of single soybeans by transmission Raman spectroscopy, *Journal of Agricultural and Food Chemistry* 60 (2012) 8097–8102.
- [8] N.K. Afseth, V.H. Segtnan, B.J. Marquardt, J.P. Wold, Raman and near-infrared spectroscopy for quantification of fat composition in a complex food model system, *Applied Spectroscopy* 59 (2005) 1324–1332.
- [9] J.W. Qin, K.L. Chao, M.S. Kim, Investigation of Raman chemical imaging for detection of lycopene changes in tomatoes during postharvest ripening, *Journal of Food Engineering* 107 (2011) 277–288.
- [10] D.S. Himmelsbach, F.E. Barton, A.M. McClung, E.T. Champagne, Protein and apparent amylose contents of milled rice by NIR-FT/Raman spectroscopy, *Cereal Chemistry* 78 (2001) 488–492.
- [11] R.M. El-Aabassy, P.J. Eravuchira, P. Donfack, B. von der Kammer, A. Materny, Fast determination of milk fat content using Raman spectroscopy, *Vibrational Spectroscopy* 56 (2011) 3–8.
- [12] G.F. Ghesti, J.L. de Macedo, V.S. Braga, A.T.C.P. de Souza, V.C.I. Parente, E.S. Figueredo, et al., Application of Raman spectroscopy to monitor and quantify ethyl esters in soybean oil transesterification, *Journal of the American Oil Chemists Society* 83 (2006) 597–601.
- [13] V. Baeten, P. Hourant, M.T. Morales, R. Aparicio, Oil and fat classification by FT-Raman spectroscopy, *Journal of Agricultural and Food Chemistry* 46 (1998) 2638–2646.
- [14] G. Schulze, A. Jirasek, M.M.L. Yu, A. Lim, R.F.B. Turner, M.W. Blades, Investigation of selected baseline removal techniques as candidates for automated implementation, *Applied Spectroscopy* 59 (2005) 545–574.
- [15] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, *Applied Spectroscopy* 57 (2003) 1363–1367.
- [16] J.V. Kresta, T.E. Marlin, J.F. Macgregor, Development of inferential process models using PLS, *Computers & Chemical Engineering* 18 (1994) 597–611.
- [17] J. Liu, K. Min, C.H. Han, K.S. Chang, Robust nonlinear PLS based on neural networks and application to composition estimator for high-purity distillation columns, *Korean Journal of Chemical Engineering* 17 (2000) 184–192.
- [18] G. ElMasry, D.W. Sun, P. Allen, Non-destructive determination of water-holding capacity in fresh beef by using NIR hyperspectral imaging, *Food Research International* 44 (2011) 2624–2633.
- [19] J. Shao, Linear-model selection by cross-validation, *Journal of the American Statistical Association* 88 (1993) 486–494.
- [20] N. Berardo, V. Pisacane, P. Battilani, A. Scandolara, A. Pietri, A. Marocco, Rapid detection of kernel rots and mycotoxins in maize by near-infrared reflectance spectroscopy, *Journal of Agricultural and Food Chemistry* 53 (2005) 8128–8134.
- [21] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, *Applied Spectroscopy* 54 (2000) 413–419.
- [22] H. Song, Oil Content and Fatty Acid Composition in Soybean Genetic Resources, Ph.D. Thesis, Department of Crop Science, Chungbuk National University, Jeonju, South Korea (2011).
- [23] I.V. Kovalenko, G.R. Rippke, C.R. Hurburgh, Determination of amino acid composition of soybeans (*Glycine max*) by near-infrared reflectance spectroscopy, *Journal of Agricultural and Food Chemistry* 54 (2006) 3485–3491.
- [24] G.Y. Zhu, X. Zhu, Q. Fan, X.L. Wan, Raman spectra of amino acids and their aqueous solutions, *Spectrochimica Acta A* 78 (2011) 1187–1195.
- [25] E.C.Y. LiChan, The applications of Raman spectroscopy in food science, *Trends in Food Science and Technology* 7 (1996) 361–370.
- [26] M.S. Ku, H. Chung, Comparison of near-infrared and Raman spectroscopy for the determination of chemical and physical properties of naphtha, *Applied Spectroscopy* 53 (1999) 557–564.
- [27] Y. Kim, S. Lee, H. Chung, H. Choi, K. Cha, Improving Raman spectroscopic differentiation of the geographical origin of rice by simultaneous illumination over a wide sample area, *Journal of Raman Spectroscopy* 40 (2009) 191–196.

Biographies

Hoonsoo Lee received his B.S., and M.S. degrees in Department of Agriculture Machinery Engineering at the Chungnam National University, Daejeon, South Korea, in 2007 and 2009, respectively. In 2010, he joined the Environmental Microbial and Food Safety Laboratory, USDA-ARS, Beltsville, Maryland, as an intern of USDA/US Forest Service International Program. He is currently a full time Ph.D. student in the Non-destructive Bio-Sensing laboratory at the Chungnam National University, Daejeon, South Korea. His research interests are quality and safety measurements of agriculture products using hyperspectral imaging techniques, such as Vis/NIR, SWIR, and Raman imaging.

Byoung-Kwan Cho received his B.S., and M.S. degrees in Department of Agricultural Engineering at Seoul National University, Seoul, South Korea, in 1993 and 1998, respectively, and received Ph.D. in Department of Agricultural and Biological Engineering from the Pennsylvania State University, University Park, USA in 2003. He joined Department of Agricultural and Biological Engineering at Purdue University, West Lafayette, IN, USA in 2003 and ARS, USDA, Beltsville, MD, USA in 2005 as a post-doctoral researcher. Since 2006 he has been a professor in the Department of Biosystems Machinery Engineering at the Chungnam National University Daejeon, South Korea. His research areas are non-destructive biosensing for quality and safety evaluation of agricultural and food materials.

Moon S. Kim received his B.S., M.S., and Ph.D. degrees from the University of Maryland, College Park, MD in 1988, 1994, and 1999, respectively. He joined ARS, USDA in 1999 as a research physicist working on development of optical sensing technologies for food safety research projects. Prior to joining ARS-USDA, he held professional positions at NASA/GSFC, Greenbelt, MD for over 10 years. Since 2009, he has served as the chair of the Sensing for Agriculture and Food Quality and Safety Conference, SPIE. He leads a multidisciplinary team of researchers to develop innovative sensing methodologies and technologies to address food safety concerns for food production and to aid in reducing food safety risks in food processing.

Wang-Hee Lee received his B.S degree in Agricultural Machinery and Process Engineering at the Seoul National University, Seoul, South Korea, in 2004, and received M.S. and Ph.D. in Agricultural and Biological Engineering from the Purdue University, West Lafayette, in 2006 and 2011, respectively. In 2011, he joined the Laboratory for Biological Systems Analysis at the Georgia Institute of Technology, Atlanta, as a post-doctoral research fellow. Since 2012 he has been an assistant professor in the Department of Biosystems Machinery Engineering at the Chungnam National University, Daejeon, South Korea. His research interests are biosystem modeling and data analysis applicable for agricultural, biological and food systems including their products and process engineering.

Jagdish Tewari is Research Scientist (Lab Manager) at Cornell University, Ithaca, NY. Dr. Tewari is well known scientist in the field of vibrational spectroscopy. His major research interest includes Vibrational Spectroscopy and Mass spectrometry combined with chemometrics and artificial neural network.

Hanhong Bae received B.S. degree in Forestry from Seoul National University (Seoul, Korea) in 1987, M.S. degree in Forest Science from Oregon State University (Corvallis, OR, USA) in 1992, and Ph.D. degree in Genetics from Iowa State University (Ames, IA, USA) in 2001. In 2001, he joined USDA-ARS-Plant Sciences Institute (Beltsville, MD, USA) as a Molecular Biologist. Since 2009 he has been on the faculty of school of biotechnology in Yeungnam University (Gyeongsan, Korea). His main researches focus on molecular plant-microbe interactions and plant responses in response to soundwave.

Soo-In Sohn received her B.S. and M.S. degrees in Biology at the Catholic University of Korea, in 1993 and 1995, respectively. In 2000 she received Ph.D. of thesis of studying embryo stage-specific gene expression in *Pimpinellabrachycarpa* at Kangwon National University. In 2000, she participated in the Department of Pharmacy as a postdoctoral research fellow at same University. As an invited researcher, she also joined the Department of Biotechnology of Tokyo University of Agriculture and Technology in 2000. In 2001, she joined the Laboratory for Proteome Analysis System at Pohang University as a post-doctoral fellow. From 2002 to 2005, she joined the department of Crop Environment and Biotechnology Division in National Institute of Crop Science as a post-doctoral fellow. From 2006 to present, she participate in the subject of Environmental Risk Assessment of GM (genetically-modified) crops as a researcher in the Biosafety Division in National Academy of Agricultural Science.

Hee-Youn Chi received B.S., M.S., and Ph.D. degrees focused on applied Bioscience from the Agricultural at Konkuk University, Seoul, Korea, in 1997, 2000, and 2005, respectively. From 2005 to 2008, she undertook postdoctoral research at the Rural Development Administration in South Korea. She worked as Senior Researcher at R&D Center of Smateome company in Korea, was involved in project for developing diverse near-infrared (NIR) spectroscopic systems and Raman spectroscopy for non-destructive analysis from 2009 to 2012. Now, she is an part-time lecture of the Department of Applied Bioscience at Konkuk University, and the Department of Plant Life & Environment Science at Hankyong National University. Currently her research interest is about the analysis of natural products using HPLC, GC, and MS and also evaluation of their biological functions on the in vitro, in situ, or in vivo level.