Plant Biotechnology 26, 451–457 (2009)

Original Paper

Special Issue

DrEFTIR: The data mining software for fourier transform near-infrared reflectance spectroscopy focused on food metabolic finger printing

Tatsuhiko Ikeda[1], Md. Altaf-Ul-Amin[2], Hiroki Takahashi[2], Eiichiro Fukusaki[1,*]

[1] Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan;
[2] Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan
* E-mail: fukusaki@bio.eng.osaka-u.ac.jp Tel & Fax: +81-6-6879-7424

Received September 9, 2009; accepted October 16, 2009 (Edited by T. Ikemura)

Abstract The near-infrared spectroscopy is used to analyze various foods because of the facts that the measurement time is very short and the coverage area is very wide. This analysis technique is recognized as an important technique to finger printing in addition to the element measurement. This research aimed at the development of the data analysis software for metabolic finger printing of food using the near-infrared spectroscopy. This software was made by using JAVA language that has advantages in developing graphical user interface. This software can perform feature extraction by using multi derivatives and Spearman's correlation with an arbitrary dependent variable after the preprocessing of the spectrum between 1000–2500 nm Wavelength by Standard Normal Variate. In addition, this software can visualize the tendency of the data by the PCA method, and determine the regression model by the PLS method. We demonstrated the usability of this software using Japanese green tea samples. A set of ranked green tea samples from a Japanese commercial tea contest was analyzed by Fourier transform near-infrared (FT-NIR) reflectance spectroscopy. This FT-NIR data was analyzed by our software, and quality prediction model was made. This prediction model had enough high accuracy.

Key words: Feature extraction, food science, fourier transform near-infrared reflectance spectroscopy, metabolomics profiling.

Near-Infrared (NIR) is part of the infrared region near the range of visible light frequency under the wavelength of 2500 nm. NIR has high permeability and reflects the feature of various compounds. NIR spectroscopy has important advantages: it allows the nondestructive analysis of solid samples, requires very little or no sample preparation, and enables extremely fast analysis (Blanco et al. 1998). Furthermore, NIR spectroscopy has already demonstrated its capacity to determine different types or species of herbal drugs to screen for their geographical origins or to quantify marker substances. It has been also utilized for authentication of foods of different geographical origins through measurement of various components (Downey and Boussion 1996; Downey et al. 1997; Sirieix and Downey 1993; Osborne et al. 1993; Evans et al. 1993; Laporte et al. 1998; O'Callaghana et al. 2000; Giardina et al. 2003; Guthrie and Walsk 1997; Buchanan et al. 1988; Dong et al. 1997; Navratil et al. 2004). Furthermore, it can be applied for examining food adulteration (Twomey et al. 1995). Therefore, NIR analysis is recognized as powerful tool for metabolomics, which is one of the post-genomics, especially in the field of food science. On the other hand, it is difficult to use Fourier Transform Near-Infrared Reflectance Spectroscopy (FT-NIR) for qualitative analysis without any chemometrics approaches, because the FT-NIR spectra of organic molecules were dominated by overtone and combination bands of fundamental vibrations involving high nonharmonic X–H (mainly C–H, N–H, and O–H) stretching modes (Williams and Norris 1990; Osborne et al. 1993). For that reasons, chemometrics approaches are needed in feature extraction and interpretation of NIR-spectra. Indeed supervised and unsupervised approaches are applied for interpretation of NIR-spectra such as data processing including SNV normalization (Blanco et al. 1998; Laasonen et al. 2002; Barnes et al. 1989), differentiation coefficient (Blanco et al. 1998; Laasonen et al. 2002), unsupervised multivariate analysis such as principal component analysis (PCA) (Blanco et al. 1998; Downey and Boussion 1996; Downey et al. 1997; Navratil et al. 2004; Laasonen et al. 2002), and supervised multivariate analysis such as multi-regression analysis and partial least square analysis (PLS) (Blanco et al. 1998; Downey et al. 1997; Laporte et al. 1998; Buchanan et al. 1988; Navratil et al. 2004; Laasonen et al. 2002). In previous

This article can be found at http://www.jspcmb.jp/

report, we indicated the effectiveness of derivative profiles and spearman's correlations in reducing the noise and amplifying the fundamental features (Ikeda et al. 2008). For evaluating the quality of foods by FT-NIR, we felt the need to develop an easy and useful software using multi derivatives and spearman's correlations. For that reason, we tried to make the software called "Dr.EFT-IR". In this report, we show the system of Dr.EFT-IR and disclose the software which is available from http://kanaya.naist.jp/DrEFT-IR/.

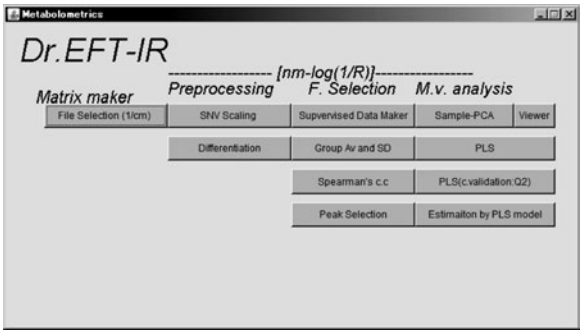

Figure 1. The main window of Dr.EFT-IR

Materials and methods

Sample Preparation for FT-NIR Analysis

The tea samples were obtained from the tea branches of the Nara Prefecture Agricultural Experiment Station. Professional tea tasters determined the ranking of teas based on the total scores (200 points full marks) of the sensory tests: leaf appearance (30 points), smell (70 points), color of the brew (30 points), and its taste (70 points). The ranks of all tea samples were determined from 1 (corresponding to the highest quality) to 67 (the lowest quality). Tea samples had test number independent of ranking number. The test numbers were given in order of registration to the sensory test. Of them, we selected samples having odd test number for constructing non-biased model equations to assess tea rank by FT-NIR spectra. As a result, 34 tea samples (ranking numbers: 1, 3, 6, 7, 8, 10, 12, 13, 14, 17, 18, 19, 20, 26, 27, 28, 29, 30, 35, 38, 39, 41, 42, 44, 45, 46, 47, 48, 49, 57, 58, 60, 62 and 64) were selected. These dried tea leaves (200 mg) in 2-ml Eppendorf tubes were ground with a Retsch ball mill (20 Hz, 10 min). The powder samples were transferred to individual 2-ml glass vials.

FT-NIR Measurements

Diffuse reflectance spectra of tea samples were measured using a NICOLET 6700 FT-IR (Thermo Electron K.K, Kanagawa, Japan) equipped with a Smart Near-IR UpDRIFT, a CaF_2 Beamsplitter and a cooled InGaAs detector. For each sample, a diffuse reflectance spectrum was measured three times. The FT-NIR spectra were recorded from 10000 to 4000 cm^{-1} at intervals of 3.857 cm^{-1}. The mirror velocity was 1.2659 $cm\,s^{-1}$, and the resolution was 8 cm^{-1}. The total number of data points was 1557 for each spectrum.

Flow of Data Processing in Dr.EFT-IR

Dr.EFT-IR has three folders and one batch file. 'Metabolometrics' Java folder has Java source files. The FT-NIR data files are kept in 'FTNIROriginal Data' folder. The FT-NIR data files must be put to 'FTNIROriginal Data' folder. The data files consist of two dimensions, wavenumber and reflectance. The result files of data analysis are saved in 'MetabolometricsOut' folder. MetabolometricsRun.bat file compile and perform the program. Figure 1 shows the main window of Dr.EFT-IR. As shown in the flow of data processing in Dr.EFT-IR (Figure 2), input and output files are managed by header names, for example, input files headed by 'Mv' can be analyzed by two preprocessing programs corresponding to 'SNV Scaling' and 'Differentiation' buttons of main window. Data processing in DrEFT-IR is divided into four major steps:

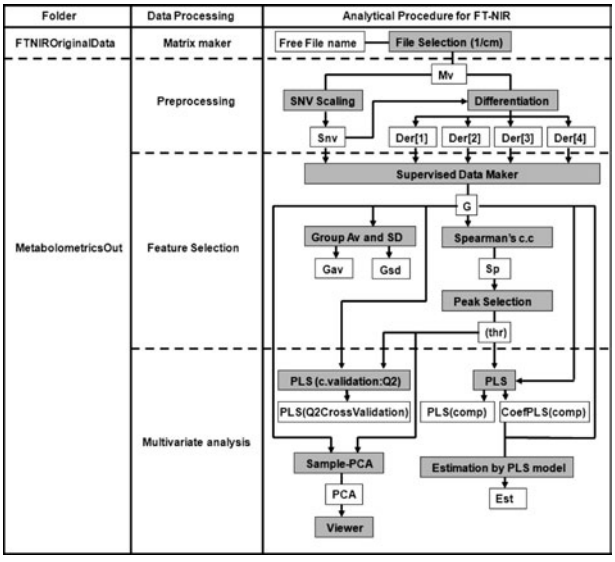

Figure 2. Flow Diagram of Data Processing in Dr.EFT-IR. Gray boxes correspond to individual processes and buttons to run each process, and white boxes correspond to prefix in input/output file names. Folder indicates data folders that are containing the files using in or made by each process.

Matrix maker, Preprocessing, Feature Selection, and Multivariate analysis. In Figure 2, white boxes correspond to prefix in input/output file names, and gray boxes correspond to individual processes and buttons to run each process.

Matrix maker: The button "File Selection (1/cm)" is clicked to open the file selection dialog. In the dialog, multiple files are selected and the data matrix is made. The FT-NIR data files have two types of information. One is wavenumber, and the other is reflectance. The file names of FT-NIR data files have no limitation. This is expressed as "Free file name" in Figure 2. The information of wavenumber is transformed to wavelength, and the reflectance (R) is transformed to $\log 1/R$. The Output file contains the matrix of wavelength and $\log 1/R$ of each sample and for the output file name "Mv" is added as header to the original file name.

Preprocessing: For normalizing, standard normal variate (SNV) is performed by "SNV Scaling". The button "SNV Scaling" is clicked to open the file selection dialog. In the dialog, only the files having "Mv" as header in file name are shown and can be selected. The mathematical transformation of the log $1/R$ spectra by the calculation of the standard normal

variation 5 at each wavelength removes the slope variation in an individual sample basis and is represented by Eq. (1) (Barnes et al. 1989):

$$x_i(s_j) = \frac{x_i'(s_j) - \bar{x}_i'}{\sqrt{\sum_{j=1}^{n} \{x_i(s_j) - \bar{x}_i'\}^2 / (n-1)}} \quad (1)$$

Here, $x_i'(s_j)$ represents the absorbance log $1/R$ of s_ith wavelength ($j = 1, 2, \ldots, n$) for jth measurements and n is the total number of channels of wavelength. $\bar{x}_i'$ is the average of $x_i'(s_j)$. "Snv" is added to output file name following "Mv".

For amplifying features, multi-differentiations are calculated by "Differentiation". Also in this process, file selection dialog is opened by clicking the button and only the files having "Mv" as header in file name are shown and can be selected. The first to fourth differentiations are calculated from the log $1/R$ spectra by using the Savitzky–Golay method (Savitzky and Golay 1964). By this processing, four output files are made for four successive differentiations. For the names of these four files we add to their original names "Der(1)", "Der(2)", "Der(3)" and "Der(4)" in case of 1st, 2nd, 3rd and 4th differentiation respectively following "MvSnv" or "Mv".

Feature Selection: "Supervised Data Maker" makes the supervised data from the preprocessed data and dependent variables that are inputted by Supervised Data Maker dialog. The preprocessed data is in the selected file having "Mv" in file name. Dependent variables mean several quality parameters that are decided by the sensory test and so on. In the Supervised Data Maker dialog, original file names expressed as "Free file name" are shown and selected for grouping sample data. Simultaneously, the dependent variable is inputted in the dialog. Output file has "G" as new header in the file name and information about original file names, dependent variables and the numbers of samples included in each group.

After that, for checking the bias of supervised data, the average reflectance and the standard deviation of each wavelength and samples included in each group are calculated by "Group Av and SD". Input file of "Group Av and SD" has to have "G" as a header in the file name. "Group Av and SD" make two output files. One is Average data file of each group in input file and "Gav" is added to the end of the file name header, another is standard deviation data file and "Gsd" is added to the file name header.

"Spearman's c.c" performs the calculation of Spearman's rank correlation coefficient. The Spearman's rank correlation coefficient does not require the assumption that the relationship between the variables is linear: it also does not require the variables to be measured at interval scales. It can be used for the variables measured at ordinal levels. The Spearman's rank correlation coefficient is represented by Eq. (2) (Spearman 1904):

$$r(s_j) = 1 - \frac{6\sum_{i=1}^{N} (xrank_i^{(k)}(s_j) - y_i)^2}{N^3 - N} \quad (2)$$

Here, s_j is the wavelength, $r(s_j)$ is the Spearman's rank correlation coefficient at s_j, N is the sample number, k is the number of derivation, and $xrank_i^{(k)}(s_j)$ is the order of the k^{th} derivative value at s_j of the i^{th} sample. y_i is dependent variable. Because dependent variable is necessary, input file has to have "G" in head of file name. "Sp" is added to the head of the output file name of "Spearman's c.c" and this file contains the data of Spearman's rank correlation coefficient of each wavelength. Using the Spearman's rank correlation coefficient, feature extraction is performed by "Peak Selection". The button "Peak Selection" is clicked to open the file selection dialog. In this dialog, we have to choose two files and input threshold value. One of the two files is correlation file having "Sp" in head of file name, and another is supervised data file having "G" in head of file name. "Peak Selection" searches the wavelengths that have larger correlation coefficients than the threshold from the correlation file. After that, the peaks that have same wavelength searched in previous step are selected from the supervised data file. "(thr)" is added to the end of the output file name header of "Peak Selection". "thr" means the value of threshold. The output file has the data of dependent variables and peak intensity of selected wavelength.

Multivariate analysis: Principal component analysis (PCA) is performed as an exploratory data analysis for formulating the predictive models by "Sample-PCA". PCA is a technique used to reduce multidimensional data sets to lower dimensions for analysis purposes. The files having "Mv" or "G" in head of file name can be applied as input files. PCA is defined by following model represented by Eq. (3).

$$T = XL \quad (3)$$

Here, T is Score matrix of PCA, L is loading matrix and X is original data matrix of the input data file. "PCA" is added to the end of the output file name header of "Sample-PCA" and this file contains the data of T and L. This output file can be inputted to "Viewer". "Viewer" displays the score plot of PCA. Because PCA can analyze the tendency and dispersion of multivariable data, PCA is utilized for verifying success of feature extraction. If the verifying is over or not necessary, the partial least-squares (PLS) is performed.

For formulating the prediction models, PLS technique is employed. PLS is a method for linearly relating two data matrices, X (M×N) and y (M×1). The model equation is represented by Eq. (4).

$$\mathbf{X} = \sum_{k=1}^{l} \mathbf{t}_k \mathbf{p}_k^T + \mathbf{E} \quad (4a)$$

$$\mathbf{y} = \sum_{k=1}^{l} \mathbf{t}_k \mathbf{q}_k + \mathbf{e} \quad (4b)$$

Here, $\mathbf{p}_k$ and $\mathbf{q}_k$ are called the loading vector of $\mathbf{X}$, and the coefficient of y for k^{th} component, respectively. X is original data matrix and y is dependent variable vector of the input data file. l is the number of components and $\mathbf{t}_k$ is a score vector for k^{th} component. E and e represent the residual matrix and vector. The number of PLS components, l, is determined to maximize Q^2 by leave-one-out cross-validation for each component. This leave-one-out cross-validation is performed by "PLS (c. validation: Q2)". Q^2 is defined by Eq. (5).

$$Q^2 = 1 - \frac{\sum (y_{obs} - y_{pred})^2}{\sum (y_{obs}^2)} \quad (5)$$

Here, y_{obs} and y_{pred} are an experimental and a predicted y values, respectively. Input file of "PLS (c. validation: Q2)" has to have "G" in head of file name. "PLS (Q2 Cross Validation)" is added to the head of the output file name of "PLS (c. validation: Q2)" and contains the data of y_{obs}, y_{pred} and Q^2 value in each components number. After that, using this optimum components number having largest Q^2 value PLS is performed by "PLS". The condition of input file of "PLS" is same to that of "PLS (c. validation: Q2)". PLS is calculated from the spectra data and dependent value of input file by using Eq. (4a) and (4b) given above. The PLS equations (Eq. (4a) and (4b)) can also be transformed into a linear form represented by Eq. (6).

$$\mathbf{y} = \mathbf{Xb} + \mathbf{f} \quad (6)$$

Here, **b** is a regression coefficient vector and its elements are represented by b_j ($j = 1, 2, \ldots, N$). By "PLS", two output files are made. One is the file of regression coefficient vector. This file has only the data of regression coefficient vector and "CoefPLS(comp)" in the head of file name. Another is the result file of PLS. This file has the data of regression coefficient vector, y_{obs}, y_{pred} and the predicted residual sum of squares (R^2 value), and "PLS(comp)" in the head of file name. In these two files, "comp" means the value of components number. If users want to know the dependent value of unknown samples, "Estimation by PLS model" estimates the dependent value using the regression coefficient vector file having "CoefPLS(comp)" in the head of file name. "Estimation by PLS model" require two input files. One is the regression coefficient vector file; another is the supervised data file of unknown samples. This supervised data file of unknown samples is made by matrix maker, preprocessing and feature selection from new original FT-NIR data files.

Results and discussion

We indicate the features of Dr.EFT-IR with an example of practical use. The data for practical use is the FT-NIR data of Japanese green tea. This green tea samples were ranked by professional tea tasters. Using Dr.EFT-IR, the prediction of ranking number of Japanese green tea was performed.

By clicking the "File Selection (1/cm)" button, we can open the File Selector dialog window which can be used to browse to the FT-NIR data files (Figure 3). In the File Selector dialog, we chose the CSV files and click the "> Add >" button. After that, we can get the data matrix by clicking the "Start Merge" button. The name of the data matrix can be decided by writing in the text field labeled "INPUT Output File". If we need preprocessing, we can do that by clicking the "SNV Scaling" or "Differentiation" button. In practical case, both of the "SNV Scaling" and "Differentiation" button were clicked.

The input of dependent variables is performed in the Supervised Data Maker dialog window (Figure 4). In the left panel of the Supervised Data Maker dialog window, we chose one sample or some samples that have same dependent variables and click the "> Add >" button. Subsequently, we input the dependent variables (ranking number of tea) in the text field labeled "input quantity of targeted variable" and click the "Decide" button. When all dependent variables have been decided, we click the "Start Grouping" button and the supervised data is made. The supervised data is used in the next step. When "Group Av and SD" or "Spearman's c.c" button is pressed, file chooser dialog is opened. Average and standard deviation or spearman's rank correlation coefficient is calculated by using the selected supervised data (Figure 5).

The peak selection is performed after calculation of the spearman's rank correlation coefficient. File chooser dialog is opened by clicking the "Peak Selection" button. The dialog selects two files. One is supervised data file and the other is the result file of "Spearman's c.c". Using these two files, peak selection is performed. By "Peak Selection" the wavelengths with a correlation value larger than 0.50 were extracted. No derivative data had 93 data points with correlation coefficients larger than 0.50, 1st derivative data had 286 data points, 2nd derivative data had 185 data points, 3rd derivative data had 54 data points and 4th derivative data had

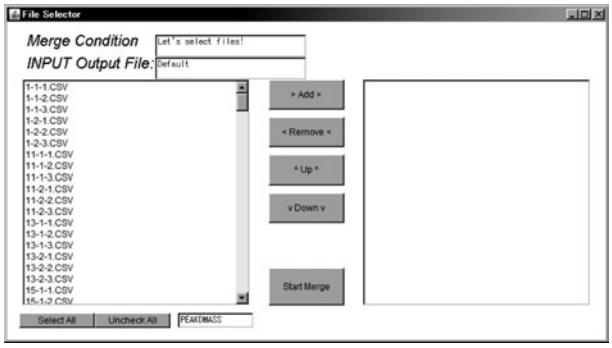

Figure 3. The File Selector dialog window.

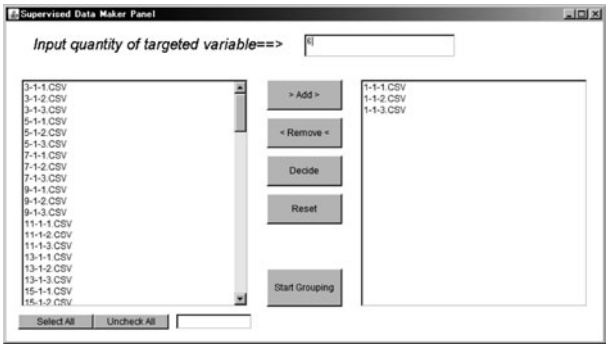

Figure 4. The Supervised Data Maker dialog window.

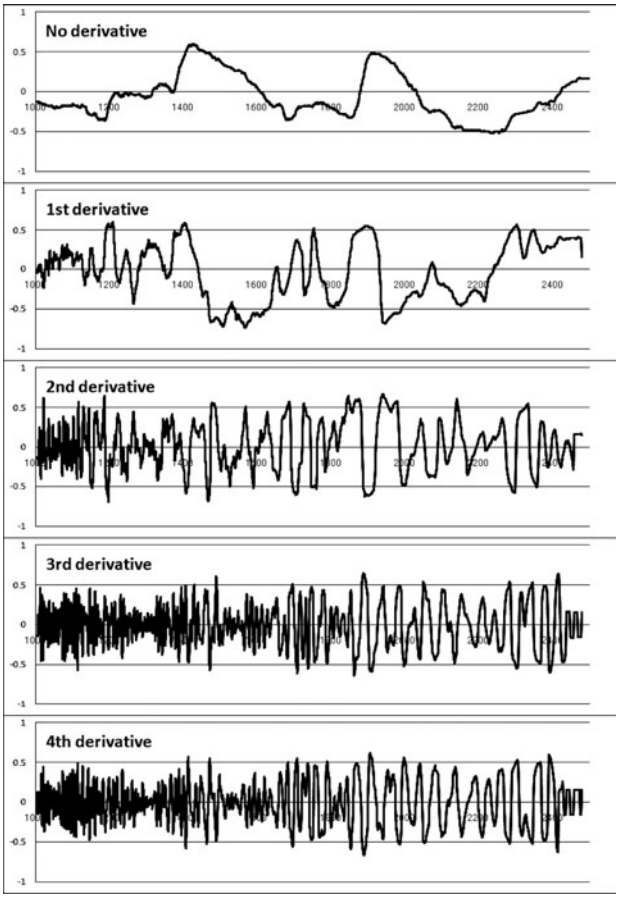

Figure 5. Spearman's rank correlation coefficients. X and Y axes indicate wavelength and correlation coefficients.

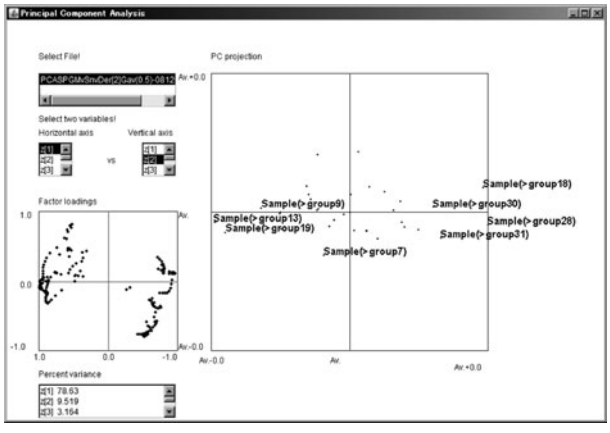

Figure 6. The Viewer of principal component analysis. In left top window, we can chose file for showing the result of PCA. In left second windows, we chose the principal component as the horizontal axis and vertical axis individually. Left third window shows loading plot of selected components. Left bottom window shows percents of variances of each component. Right large window shows score plot of PCA. Each spot indicates score of each sample. If we want to know the information of the spot, sample name is shown by clicking the spot. X and Y axis indicate scores of the principal components selected in left second windows.

56 data points. "Sample-PCA", "PLS" and "PLS (cross-validation)" can use the supervised data file or the result file of "Peak Selection". The result of "PCA" can be shown by using "Viewer" (Figure 6). In this viewer, Score plot and loading plot of selected result file of "PCA" is shown. Horizontal axis and vertical axis can select any factor. Percent variance of each factor is shown in left-down side of the viewer. Figure 6 shows PCA score plot of 2nd derivative data. Not only the result of 2nd derivative data but also that of other derivatives data showed the approximate separation according to the ranking number (data not shown).

For making PLS model, the optimum number of factors for calibration is decided by clicking of the "PLS (cross-validation)" button. As the result, we got the optimum numbers (Figure 7). The number of significant factors is 1 at no derivative, 2 at 1st derivative, 4 at 2nd derivative, 2 at 3rd derivative, and 2 at 4th derivative. Using this optimum number, making prediction model was performed by clicking the "PLS" button. The PLS regression model of 2nd derivative had sufficient accuracy, and other models did not have sufficient accuracy (Figure 8). If we need the estimation of new unknown sample, "Estimation by PLS model" is performed by using the result file of PLS and the supervised or peak selected data file.

In previous report, we use paste tea samples with glycerol (Ikeda et al. 2007). We could not make prediction model using powder samples, in that case, because FT-NIR data of powder samples had much noise. However, in this study, we succeeded making prediction model from FT-NIR data of powder samples. This result shows that our software can reduce the noise and has benefit for making prediction model from FT-NIR data of powder samples.

A software tool for data mining focused on food

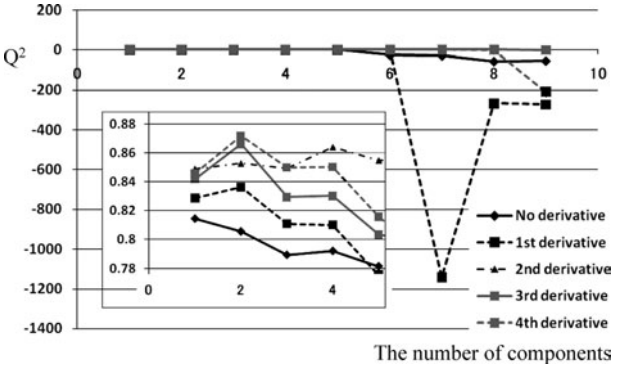

Figure 7. The cross validation of PLS. X axis indicates number of components used for making PLS prediction model. Y axis indicates Q^2 value in cross validation of PLS. The small figure shows magnified plots. X and Y axis of small figure also indicate number of components and Q^2 value. The number of significant factors is 1 at no derivative, 2 at 1st derivative, 4 at 2nd derivative, 2 at 3rd derivative, and 2 at 4th derivative.

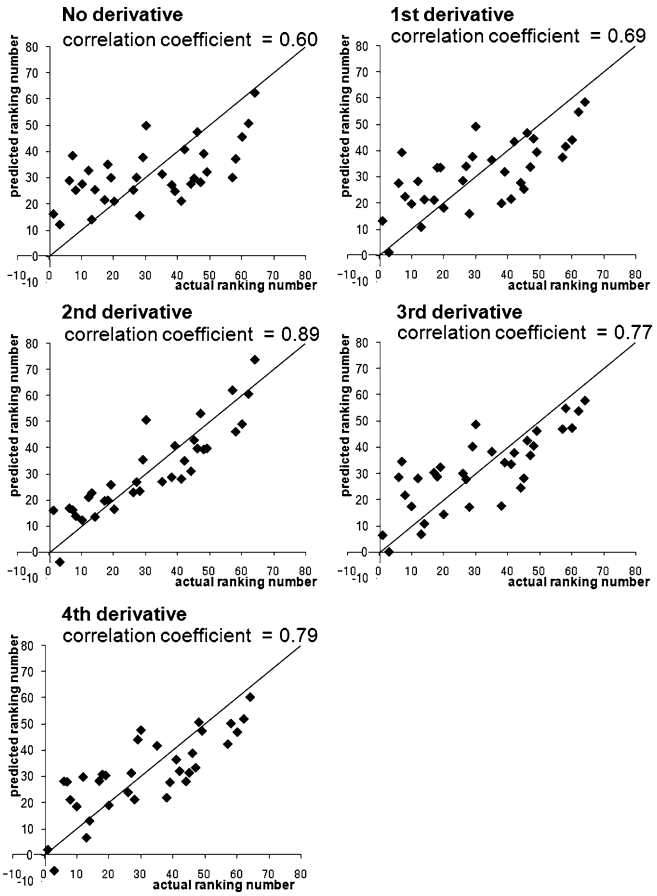

Figure 8. Prediction of ranking number by PLS. X axis indicates actual ranking number. Y axis indicates predicted ranking number by PLS models. The correlation coefficient between predicted numbers with actual numbers was 0.60 at no derivative, 0.69 at 1st derivative, 0.89 at 2nd derivative, 0.77 at 3rd derivative, and 0.79 at 4th derivative.

metabolic finger printing has been developed. Feature extraction by spearman's rank correlation coefficient, exploratory multivariate analysis by PCA and Making standard curve by PLS can be performed by using this software. The data size reduction by feature extraction give the decrease in the computing time and high accuracy of estimation. Additionally, we developed this software which is equipped with convenient to handle Graphical User Interface (GUI). We think that this software is very useful in the field of food metabolic finger printing.

Acknowledgements

The study was supported, in part, by Collaboration of Regional Entities for the Advancement of Technological Excellence from Japan Science and Technology Corporation (JST-CREATE).

References

Blanco M, Coello J, Iturriaga H, Maspoch S, Pezuela C (1998) Near-infrared spectroscopy in the pharmaceutical industry. *Analyst* 123: 135–150

Downey G, Boussion J (1996) Authentication of coffee bean variety by near-infrared reflectance spectroscopy of dried extract. *J Sci Food Agric* 71: 41–49

Downey G, Briandet R, Wilson RH, Kemsley EK (1997) Near- and mid-infrared spectroscopies in food authentication: coffee varietal identification. *J Agric Food Chem* 45: 4357–4361

Sirieix A, Downey G (1993) Commercial wheatflour authentication by discriminant analysis of near infrared reflectance spectra. *J Near Infrared Spectrosc* 1: 187–198

Osborne BG, Mertens B, Thompson M, Fearn T (1993) The authentication of Basmati rice using near infrared spectroscopy. *J Near Infrared Spectrosc* 1: 77–83

Evans DG, Scotter CNG, Day LZ, Hall MN (1993) Determination of the authenticity of orange juice by discriminant analysis of near infrared spectra: A study of pretreatment and transformation of spectral data. *J Near Infrared Spectrosc* 1: 33–44

Laporte MF, Martel R, Paquin P (1998) The near-infrared optic probe for monitoring rennet coagulation in cow's milk. *Int Dairy J* 8: 659–666

O'Callaghana DJ, O'Donnellb CP, Paynec FA (2000) On-line sensing techniques for coagulum setting in renneted milks. *J Food Eng* 43: 155–165

Giardina C, Cattaneo TMP, Barzaghi S (2003) Study of modifications in delactosated milk during shelf-life by NIR and FT-IR spectroscopy. *Milchwissenschaft* 58: 363–366

Guthrie J, Walsk K (1997) Non-invasive assessment of pineapple

and mango fruit quality using near infra-red spectroscopy. *Aust J Exp Agric* 37: 253–263

Buchanan BR, Honigs DE, Lee CJ, Roth W (1988) Detection of ethanol in wines using optical-fiber measurements and near-infrared analysis. *Appl Spectrosc* 42: 1106–1111

Dong J, Ma K., Voort FR, Ismail AA (1997) Stochiometric determination of hydroperoxides in oils by Fourier Transform Near-Infrared (FT-NIR) spectroscopy. *J AOAC Int* 80: 345–352

Navratil M, Cimander C, Mandenius C (2004) On-line multisensor monitoring of yogurt and filmjo1lk fermentations on production scale. *J Agric Food Chem* 52: 415–420

Twomey M, Downey G, McNulty P (1995) The potential of NIR spectroscopy for the detection of the adulteration of orange juice. *J Sci Food Agric* 67: 77–84

Williams P, Norris K (1990) *Near-Infrared Technology in the Agricultural and Food Industries*, second ed. American Association of Cereal Chemists, Minnesota

Osborne BG, Fearn T, Hindle PH (1993) *Practical Near-Infrared Spectroscopy with Applications in Food and Beverage Analysis*, Longman Scientific & Technical, Essex, England

Laasonen M, Harmia-Pulkkinen T, Simard CL, Michiels E, Rasanen M, Vuorela H (2002) Fast identification of Echinacea purpurea dried roots using near-infrared spectroscopy. *Anal Chem* 74: 2493–2499

Barnes RJ, Dhanoa MS, Lister SJ (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc* 43: 772–777

Ikeda T, Altaf-Ul-Amin Md, Parvin AK, Kanaya S, Yonetani T, Fukusaki E (2008) Predicting rank of Japanese green teas by derivative profiles of spectra obtained from fourier transform near-infrared reflectance spectroscopy. *J Comput Aided Chem* 9: 37–46

Savitzky A, Golay JE (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* 36: 1627–1639

Spearman C (1904) The proof and measurement of association between two things. *Amer J Psychol* 15: 72–101

Ikeda T, Kanaya S, Yonetani T, Kobayashi A, Fukusaki E (2007) Prediction of Japanese green tea ranking by fourier transform near-infrared reflectance spectroscopy. *J Agric Food Chem* 55: 9908–9912