Analytical Methods

Sugar and acid content of Citrus prediction modeling using FT-IR fingerprinting in combination with multivariate statistical analysis

Seung Yeob Song a,1, Young Koung Lee b,1, In-Jung Kim a,c,*

a Faculty of Biotechnology, College of Applied Life Sciences, Jeju National University, Jeju 690-756, Republic of Korea
b 1 Bungtown Rd., Cold Spring Harbor Laboratory, NY 11724, USA
c Research Institute for Subtropical Agriculture and Biotechnology, SARI, Jeju National University, Jeju 690-756, Republic of Korea

A R T I C L E   I N F O

Article history:
Received 6 January 2015
Received in revised form 17 June 2015
Accepted 20 June 2015
Available online 22 June 2015

A B S T R A C T

A high-throughput screening system for Citrus lines were established with higher sugar and acid contents using Fourier transform infrared (FT-IR) spectroscopy in combination with multivariate analysis. FT-IR spectra confirmed typical spectral differences between the frequency regions of 950–1100 cm⁻¹, 1300–1500 cm⁻¹, and 1500–1700 cm⁻¹. Principal component analysis (PCA) and subsequent partial least square-discriminant analysis (PLS-DA) were able to discriminate five Citrus lines into three separate clusters corresponding to their taxonomic relationships. The quantitative predictive modeling of sugar and acid contents from Citrus fruits was established using partial least square regression algorithms from FT-IR spectra. The regression coefficients (R²) between predicted values and estimated sugar and acid content values were 0.99. These results demonstrate that by using FT-IR spectra and applying quantitative prediction modeling to Citrus sugar and acid contents, excellent Citrus lines can be early detected with greater accuracy.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Citrus originated in Southeast Asia and are characterized by fragrant flowers and edible juicy fruit. Citrus fruits are produced throughout the tropical and subtropical regions of the world. Major production region of Citrus in Korea is Jeju, a subtropical region. Citrus is a high-yield production fruit and accounts for about 25%, 680 thousand tons yearly, of the total yield production of all domestic fruit orchards. In Korea, Satsuma mandarin is major cultivar of Citrus and occupied 90% over.

Mutation-breeding technology is artificially induced genetic variation and eventually changes individual characteristics. It has been used worldwide for more than 80 years in various crop varieties since the 1960s (van Harten, 2008). Due to the development of novel research methods, it has become possible to apply metabolomics research to demonstrate qualitative and quantitative patterns including sugar, acid, bioactive compounds between samples at the metabolic level. Such studies primarily employ Fourier transform infrared (FT-IR) spectroscopy, near-infrared (NIR) spectroscopy, proton nuclear magnetic resonance (H NMR) spectroscopy, and mass spectrometry (MS). The data used in the analyses utilize a variety of sample identifications and are applied in metabolite marker development (Krishnan, Kruger, & Ratcliffe, 2005; Chen, Zhang, & Matsunaga, 2006). The FT-IR analysis method is a rapid and accurate classification tool that applies multivariate statistical analysis to spectra information (Gallardo-Velázquez, Osorio-Revilla, Zuñiga de Loa, & Rivera-Espinoza, 2009).

Therefore, metabolomics are being utilized not only in basic research such as the functional identification of metabolomic components, but also in various industrial applications including plant variety, agricultural origin, cultivation age, authenticity identification, standardization of the pharmaceutical industry, and other stability-related applications.

Partial least squares (PLS) regression is a representative type of multivariate statistical analyses that has been utilized in predictive modeling of various components using exact quantitative data from samples. Correlation analyses of spectra are also performed for the samples (Bastien, Vinzi, & Tenenhaus, 2005; Höskuldsson, 1988; Mevik & Wehrens, 2007; Wold, Sjöström, & Eriksson, 2001). The combination of these two methods has increased the possibility of practical industrial applications with very high precision and accuracy. These PLS regression technology applications include the prediction modeling of fatty acid components in hawthorn plant Camellia (Camellia oleifera) (Yuan, Wang, Chen, Zhou, & Ye, 2013); apple, orange, and peach juice...
carotenoid contents (Leopold, Leopold, Diehl, & Socaciu, 2011); sugar composition in orange (Magwaza et al., 2014); mushroom monosaccharide and polysaccharide contents (Chen et al., 2012); and meat decomposition and damage (Argyri et al., 2013).

In this study, sugar and acid contents were investigated in Citrus samples at early and harvesting stage obtained through a mutation-breeding method. We also established rapid classification and identification systems using multivariate statistical analysis of FT-IR spectra data. The ultimate goal of this work is to develop rapid content prediction modeling of sugar and acid in Citrus fruits at harvest time and establish a rapid selection system that can generate elite mutation Citrus lines with excellent sugar and acid contents before harvesting.

2. Materials and methods

2.1. Plant materials

Citrus unshiu Marc. Cv. Miyagawa-Wase and radiation-induced mutagenesis lines were used in this study. C. unshiu Marc. Cv. Miyagawa-Wase called as Satsuma mandarin was widely cultivated in the Jeju island in South Korea, and the mutagenized line was selected based on its different sugar and acid contents compared with C. unshiu Marc. Miyagawa-Wase. Mutagenized lines were identified by grafting with trifoliate orange after induction of mutagenesis using 60Co radiation exposure at the Department of Radiation Applied Science Institute in Jeju National University. In this study, we used five Citrus lines: C. unshiu Marc. Cv. Miyagawa-Wase as a control and four mutagenized lines. For statistical analysis, we collected between three and five samples and stored each at 4 °C for characteristic trait observation. In addition, the mutagenized Citrus lines were freeze-dried before the experiments, and the dried Citrus was ground and stored at −70 °C prior to analysis.

2.2. Measurement of sugar and acid contents

Citrus supernatants prepared using a juice extractor was injected into a sugar and acid content analysis device (NH-2000, HORIBA, Japan) to measure the contents. Sugar contents analysis using refractive index and acid contents analysis using electric conductivity. The refractive index was measured by using spectrum D of visible light. Statistical significances on comparisons of all results between control and mutant samples were made by using a nonparametric test (Wilcoxon test) with p < 0.05 (SPSS, ver. 12.0; SPSS Inc., Chicago, IL, US).

2.3. FT-IR spectrum analysis and multivariate statistical analysis

Each 20-mg Citrus powder sample in C. unshiu Marc. Cv. Miyagawa-Wase and the four mutagenized lines were placed in tubes and mixed with 200 μL methanol. Individual tubes containing extracted samples were incubated in a 50 °C water bath for 20 min, and the supernatant was transferred into a new tube after centrifugation at 10,000×g for 15 min. To eliminate cell debris, the collected supernatant was repeatedly centrifuged and carefully transferred to a new tube for storage at −20 °C prior to undergoing FT-IR spectroscopy.

FT-IR measurement was recorded by a Tensor 27 (Bruker Optics GmbH, Ettlingen, Germany) and analyzed by DTGS (deuterated triglycine sulfate). Extracted each 5-μL sample was aliquoted in a 384-well ZnSe plate and dried on a 37 °C hot plate for 20 min. FT-IR in the dried ZnSe plates were measured by a Tensor 27 equipped with a high-efficiency HTS-XT automation device (Bruker Optics GmbH). The spectra of each sample were used to analyze the average spectra of 128 repeated measurements in the range of 400–4000 cm⁻¹ and with a 4 cm⁻¹ interval of spectral resolution.

For the statistical analysis, the FT-IR spectrum of each sample was measured from three individual replicates. All spectra operations including measurements and the program for data conversion were investigated using OPUS Lab (ver. 6.5, Bruker Optics Inc.). For the multivariate statistical analysis of FT-IR spectra data, we first used the R program (version 2.15.0, Auckland, New Zealand) for preprocessing the spectra of the data baseline correction, normalization, and mean centering of the FT-IR spectra. To standardize baseline correction, absorbance of the FT-IR spectrum analysis was adjusted to 0 at both end points, 800–1800 cm⁻¹, and each spectrum was normalized to the same area to minimize experiment error. Later, we followed a mean centering process, and preprocessing spectrum data were used as standardized data for multivariate statistical methods after applying quadratic differential equation. Processed FT-IR spectrum data were submitted to PCA (principal component analysis) in the R program (version 2.15.0) using a NIPALS algorithm (Wold, 1966) and PLS-DA analysis (partial least square discriminant analysis) (Fiehn et al., 2000; Trygg, Holmes, & Lundstedt, 2007). HCA (hierarchical clustering analysis) was performed using the score obtained through PCA and PLS-DA, and a hierarchical dendrogram of each sample was generated from UPGMA (unweighted pair group method with arithmetic mean analysis) using a similarity index that measured Euclidean distance.

2.4. Sugar-acid content prediction of Citrus samples with PLS modeling

To predict sugar and acid contents in Citrus samples, a sugar and acid content prediction model was generated from FT-IR spectra data measured by qualification data from C. unshiu Marc. Cv. Miyagawa-Wase and four mutagenized lines. The X variable indicates FT-IR spectra data, and two Y variables from the same samples represent the sugar and acid contents measured as estimated quantitative data using HORIBA (NH-2000, HORIBA, Japan). The R program (version 2.15.0) was used for PLSR analysis (partial least square regression). To improve the accuracy of predictive modeling, cross-validation was applied by each X variable and the two Y variables. The total dataset was divided into two parts, training set and test set. The training set data was obtained from 15 samples and the test set data from 10 samples. Test set data were used in the prediction model. The sugar and acid content prediction of each Citrus sample was carried out using the developed prediction model. To measure the accuracy of content prediction modeling from each Citrus sample, the actual measurement and prediction values of sugar and acid content were examined by calculating correlation coefficients using linear regression analysis.

3. Results and discussion

3.1. Sugar and acid content in Citrus fruits

Sugar and acid contents were measured at two different time points (August and November) using a HORIBA device (Table 1). The overall sugar content ranges in Citrus lines in August were 6.2–6.3°Brix between C. unshiu Marc. Cv. Miyagawa-Wase and four mutagenesis-induced lines. The differences in general sugar contents between the control and mutagenesis lines were relatively low and not significantly different in samples collected in August. This was because the Citrus water content had begun to evaporate, whereas sugar content began to accumulate in this period. While all the samples had similar sugar contents, acid varied among
samples. *C. unshiu* Marc. Cv. Miyagawa-Wase (control) had the highest acid content at 2.19%, and the M4 line had the lowest acid content at 1.75%. There was a noticeable difference in the acid contents between *C. unshiu* Marc. Cv. Miyagawa-Wase and the four mutantized lines. The M2 line had a value of 8.4°Brix (similar sugar content compared with the control), and M3 and M4 had approximately 1°Brix higher than control (9.7 and 9.5°Brix, respectively). In addition to sugar content, the data also showed that there was a difference in the acidities between *C. unshiu* Marc. Cv. Miyagawa-Wase and the mutagenized M lines. The acid content in *C. unshiu* Marc. Cv. Miyagawa-Wase was 0.64%, while those in M1, M2, M3, and M4 were 0.99%, 0.79%, 0.61%, and 0.65%, respectively (Table 1). These data were consistent with those published by Song, Heo, and Kim (2014), who recently reported that sugar content in mutagenized *C. unshiu* Marc. Cv. Miyagawa-Wase lines were divided into two parts, with the top being high and the bottom being low, by calculating correlations based on the sugar and acid content difference (Table 1) and sample locations in the PCA score plot. The plot revealed that *C. unshiu* Marc. Cv. Miyagawa-Wase with its relatively low sugar content was separately located in the middle of the Y-axis. However, mutagenized M1 lines with the highest sugar content were located in a separate cluster between the M2 and M3 lines.

### Table 1

Quantitative analysis of sugar and acid content from *Citrus* samples.

<table>
<thead>
<tr>
<th>Sample Line number</th>
<th>August Sugar content (°Brix)</th>
<th>November Sugar content (°Brix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>6.2 ± 0.13</td>
<td>8.4 ± 0.42</td>
</tr>
<tr>
<td>M1</td>
<td>6.2 ± 0.29</td>
<td>10.3 ± 1.88</td>
</tr>
<tr>
<td>M2</td>
<td>6.2 ± 0.53</td>
<td>8.4 ± 0.23</td>
</tr>
<tr>
<td>M3</td>
<td>6.3 ± 0.38</td>
<td>9.7 ± 0.83</td>
</tr>
<tr>
<td>M4</td>
<td>6.3 ± 0.26</td>
<td>9.5 ± 0.18</td>
</tr>
</tbody>
</table>

Data represent the mean ± SD, all samples were measured in five replicates.

* Values are significantly different from corresponding C (*P* < 0.05).

### 3.2. FT-IR spectra analysis in *Citrus* samples

Each FT-IR spectra site (at 950–1100 cm⁻¹, 1300–1500 cm⁻¹, and 1500–1700 cm⁻¹) of a chemical compound provides quantitative and qualitative information such as the presence of a carbohydrate family containing a monosaccharide complex polysaccharide (C–O ring, C–C and C–O–C, the region of 950–1100 cm⁻¹); organic acids containing phosphodiester bonds (CH₂ and P=O, the region of 1300–1500 cm⁻¹); and phosphorus from nucleic acids phospholipids, and amino acids and amide bonds I and II of protein (P=O, C=O, N–H, C–C and C=C, the region of 1500–1700 cm⁻¹), respectively (Fig. 1) (D’Souza et al., 2008; Dumas & Miller, 2003; Lopez-Sanchez, Ayora-Canada, & Molina-Diaz, 2010; Parker, 1983; Wolkers, Oliver, Tablin, & Crowe, 2004; Yee, Benning, Phoenix, & Ferris, 2004). Therefore, quantitative and qualitative differences in FT-IR spectra represented significant qualitative and quantitative differences in amino acids, proteins, fatty acids, and compounds of carbohydrate-related families. Therefore, FT-IR spectra analysis of *Citrus* samples is a useful and practical tool for breeders to select good lines or strains with favorable levels of specific chemical compound. It will also allow breeders to observe quantitative and qualitative alterations in major metabolites.

### 3.3. FT-IR spectrum data and multivariate statistical analysis in *Citrus*

To better understand FT-IR spectra variability, PCA of FT-IR was performed by generating two-dimensional plots between control and mutant *Citrus* samples using two principal components, PC1 and PC2. The PCA analysis result from FT-IR spectra data showed that the overall variation scores of PC1 and PC2 were 63.4% and 15.5%, respectively. This result indicates that overall variability was 78.9% when PC1 and PC2 were combined (Fig. 2A). The PCA score plot with a high explanation power of PC1 and PC2 was examined, and the scatter plots comparing *C. unshiu* Marc. Cv. Miyagawa-Wase and the mutagenized M lines were divided into two parts, with the top being high and the bottom being low, by the PC2 score. Based on the PCA plot, *C. unshiu* Marc. Cv. Miyagawa-Wase lines (control) were broadly located in PC1 but located in the middle of the PC2 axis, and mutagenized M2 lines that had similar sugar contents as the control were closely located together in the *C. unshiu* Marc. Cv. Miyagawa-Wase group. M3 and M4 lines were individually located in the bottom and top parts based on their PC2 scores. M1 lines, containing the highest sugar contents were clustered between M2 and M3 lines (Fig. 2A). To understand biological information based on the spatial distribution of position from the PCA score plot in *Citrus* sample lines, we calculated correlations based on the sugar and acid content difference (Table 1) and sample locations in the PCA score plot. The plot revealed that *C. unshiu* Marc. Cv. Miyagawa-Wase with its relatively low sugar content was separately located in the middle of the Y-axis. However, mutagenized M1 lines with the highest sugar content were located in a separate cluster between the M2 and M3 lines.
lines. We found a close correlation between sample sugar content and the spatial distribution location in the PCA score plot.

To investigate the critical region of the FT-IR spectra, we examined which area determined PC1 and PC2 as it plays an important role in Citrus group classification and sample clustering based on the content of functional components in the PCA score plot. Loading value analysis revealed that the 1300–1500 cm\(^{-1}\) and 1500–1700 cm\(^{-1}\) regions of FT-IR spectra were important in determining the classification of critical PC1 into top and bottom sections. In addition, the 950–1100 cm\(^{-1}\) region played a significant role in PC2 for discriminating between wild-type Citrus and the M lines (Fig. 2B). The PCA score from the FT-IR spectra region coincided with the largely different regions observed in the representative Citrus fruit FT-IR spectra (Fig. 1). The coincident result is thought to play an important role at the metabolic scale level, and differences in amides I, II and carbohydrate group compounds can distinguish between wild-type Citrus samples and M lines. In particular, PC2 played an important role in the classification and discrimination of Citrus samples containing different levels of functional compounds. The qualitative and quantitative differences in the compounds of the carbohydrate family significantly impacted PC2, implying that quantitative changes in the primary metabolic components of sugar family-related contents strongly correlated with qualitative and quantitative alterations of secondary metabolic components.

PLS-DA analysis achieved more accurate cluster discrimination between wild-type and M Citrus samples than PCA analysis (Fig. 3). The cluster boundary of an individual Citrus sample was reduced more than with PCA analysis, and replicates belonging to the same species were more closely clustered than following PCA analysis. PLS-DA analysis places similar samples in nearby spatial locations, which results in better identification ability. Our findings clearly demonstrate that PLS-DA analysis is a better tool for Citrus discrimination than PCA analysis.

*C. unshiu* Marc. Cv. Miyagawa-Wase with a relatively low sugar content was clustered in a group in the middle of the axis on the PLS-DA score plot, and M1 lines with the highest sugar content formed a discrete cluster in the right upper region along with M3 lines. M2 lines with similar sugar contents as *C. unshiu* Marc. Cv. Miyagawa-Wase were closely located near *C. unshiu* Marc. Cv. Miyagawa-Wase on the PL-DA score plot. In addition to PCA, PLS-DA also showed a close correlational relationship between the spatial distribution positions of these samples and the sugar-acid contents. PCA dendrograms based on the PCA and PLS-DA of FT-IR data were generated (Fig. 4). Based on the dendrogram analysis, it was possible to estimate the phylogenetic relationships among five Citrus samples (Fig. 4). The PCA dendrogram showed that four M lines (M1–M4) were mainly centered on *C. unshiu* Marc. Cv. Miyagawa-Wase. This is because all four lines were induced from *C. unshiu* Marc. Cv. Miyagawa-Wase. Overall, five sample lines were divided into two distinct subclades (Fig. 4). It was determined that M1–M3 lines had a closer phylogenetic relationship with *C. unshiu* Marc. Cv. Miyagawa-Wase than M4 among the five sample lines (Fig. 4A). The phylogenetic relationship pattern from the PLS-DA dendrogram was similar.

Based on the PLS-DA analysis, the *C. unshiu* Marc. Cv. Miyagawa-Wase and M2 lines had the closest phylogenetic relationship, and the M1 and M3 lines were closer with regard to characteristic traits. However, M4 lines, which were separately grouped from the control line, had a similar phylogenetic relationship pattern in both PCA and PLS-DA analyses. It was a relatively low phylogenetic relationship of characteristic trait. This result showed that the mutation induced a different phenotype;
therefore it is possible to estimate the phylogenetic relationship using PCA and PLS-DA analyses to compare the wild-type and mutagenized M lines.

3.4. Sugar-acid content prediction of PLS regression modeling from FT-IR spectra

Quantitative prediction PLS modeling from individual Citrus fruits was established using a PLSR algorithm from FT-IR spectra data and sugar and acid contents measured with a sugar-acid analysis device (Fig. 5). For sugar content, the correlation coefficients ($R^2$) were 0.99 after performing a regression analysis using the measured component and estimated component values of the same sample from FT-IR spectra over PLS modeling (Fig. 5). Fig. 5B shows that correlation coefficient for acid content between estimated and measured component values measured by linear regression analysis was also $R^2 = 0.99$. The $R^2$ values between the measured and estimated values of sugar and acid contents were also 0.99. These results indicated that the quantitative predictions of sugar and acid contents could be predicted solely from the FT-IR spectra data of Citrus fruits with approximately 90% accuracy. In addition, the reported $R^2$ values for rind fructose, glucose, sucrose, and total carbohydrates were 0.8, 0.79, 0.77, and 0.78, respectively (Magwaza et al., 2014). Thus sugar-acid content prediction modeling in this study could be adapted to other Citrus species such as sweet orange and lemon.

The present results demonstrate that the correlation coefficient of sugar-acid content, represented by complex components, was higher than that of individual component correlation coefficients in previous research. These findings indicate that the sugar and acid content of Citrus could be estimated before harvesting, and more accurate predictive component modeling is expected for future PLS. For example, sample standardization could further improve PLS modeling accuracy, as could increasing the number of samples used in models, and the utilization of samples with large differences in sugar-acid contents. PLS modeling is easy to apply, costs relatively little, and enables the rapid prediction of sugar-acid components when using the modeling described here.

4. Conclusion

The sugar-acid content prediction modeling technique described in this study enables the use of an evaluation tool for high-quality trait characteristics for a majority of unidentified and unstandardized Citrus samples. This could be a useful rapid selection method for the early quality evaluation of Citrus samples or outstanding Citrus lines. Such methods will accelerate the direct
development of Citrus variety breeding with increased sweetness ratios due to high sugar and low acid contents. Although the modeling showed the successful results, further studies with more mutant lines if any, are necessary to improve the modeling accuracy.

Acknowledgment

This work was supported by a research grant from Jeju National University of Republic of Korea in 2010. We thank Golden Seed Project and Research Institute for Subtropical Agriculture and Biotechnology for providing the Citrus materials.

References


