DETECTION OF GENETICALLY MODIFIED SUGARCANE BY USING TERAHERTZ SPECTROSCOPY AND CHEMOMETRICS

J. Liu [1,3*], H. Xie [2], B. Zha [2], W. Ding [2], J. Luo [2], C. Hu [3]

[1] College of Food Science, Southwest University, Chongqing, 400715, China; e-mail: jjn.edu@outlook.com
[2] School of Electrical Engineering, Jiujiang University, Jiujiang Jiangxi 332005 China
[3] Guilin University of Electronic Technology, Guangxi Key Laboratory of Automatic Detecting Technology and Instrument, Guangxi, China

A methodology is proposed to identify genetically modified sugarcane from non-genetically modified sugarcane by using terahertz spectroscopy and chemometrics techniques, including linear discriminant analysis (LDA), support vector machine-discriminant analysis (SVM-DA), and partial least squares-discriminant analysis (PLS-DA). The classification rate of the above mentioned methods is compared, and different types of preprocessing are considered. According to the experimental results, the best option is PLS-DA, with an identification rate of 98%. The results indicated that THz spectroscopy and chemometrics techniques are a powerful tool to identify genetically modified and non-genetically modified sugarcane.

Keywords: *terahertz spectroscopy, genetically modified, spectroscopy, chemometrics.*

УСТАНОВЛЕНИЕ ГЕНЕТИЧЕСКИ МОДИФИЦИРОВАННОГО САХАРНОГО ТРОСТНИКА С ИСПОЛЬЗОВАНИЕМ ТЕРАГЕРЦОВОЙ СПЕКТРОСКОПИИ И ХЕМОМЕТРИИ

J. Liu [1,3*], H. Xie [2], B. Zha [2], W. Ding [2], J. Luo [2], C. Hu [3]

УДК 543.42:664.111

[1] Колледж продовольственной науки, Юго-западный университет, Чунцин, 400715, Китай; e-mail: jjn.edu@outlook.com
[2] Школа электротехники, Университет Цзюцзян, Цзюцзян, Цзянси, 332005, Китай
[3] Гуйлиньский университет электронных технологий, Гуанси, Китай

(Поступила 7 сентября 2016)

Предложена методология идентификации генетически модифицированного сахарного тростника, основанная на использовании терагерцовой спектроскопии и хемометрии, а также линейного дискриминантного анализа (LDA), векторного дискриминантного анализа (SVM-DA) и метода частичных наименьших квадратов (PLS-DA). Получено, что лучшие результаты дает PLS-DA, позволяющий идентифицировать вещество с 98% вероятностью. Показано, что терагерцовая спектроскопия и хемометрия – мощный инструмент, позволяющий различать генетически и негенетически модифицированный сахарный тростник.

Ключевые слова: *терагерцовая спектроскопия, генетически модифицированный, спектроскопия, хемометрия.*

Introduction. Genetical modification is a technology of transferring genes from one organism to another, which aims to increase biological resistance. However, genetically modified organisms (GMOs) [1–4] bear potential risks for human health and the environment. For these reasons, genetically modified products are rigorously regulated in the majority of countries in the world, and it is very important to develop effective methods for their rapid classification.

At present, the majority of detection methods for genetically modified organisms include the polymerase chain reaction (PCR) [5, 6], and enzyme-linked immune sorbent assay (ELISA) [7, 8], but these methods are expensive and complex. In order to avoid using these traditional methods, many researchers commit themselves to originating multivariate methods for genetically modified organisms, such as principal component analysis (PCA) [9, 10], partial least squares (PLS) [11, 12], support vector machine (SVM) [13–16], and linear discriminant analysis (LDA) [17–20].

Alcântara et al. distinguished genetically modified soybean grains from non-genetically modified grains by using principal component analysis and FT-MIR methods [21]. Luna, Silva et al. proposed a method to identify genetically modified soybean oil samples from traditional soybean oil samples using near infrared (NIR) spectroscopy and chemometrics techniques [22]. Aparicio et al. contrasted spectroscopy and high-performance liquid chromatography (HPLC) in the classification of adulteration in olive oil samples [23]. Koidis et al. developed a novel discriminate methodology for correctly labeling vegetable oils using spectroscopy and chemometrics techniques [24]. Nunes et al. used absorption spectroscopy and chemometrics method to evaluate the authenticity of quality parameters of edible oils and fats [25].

Therefore, the absorption spectroscopy and chemometrics technique is a powerful tool to detect genetically modified organisms. Detecting genetically modified organisms has been applied in food monitoring because of its advantages, such as relative ease and excellent classification results. Owing to these advantages, the aim of this paper is to propose a novel approach to distinguish genetically modified sugarcane from non-genetically modified sugarcane using terahertz spectroscopy and chemometrics.

Chemometric methods. *Spectral range selection.* All of the absorption spectra of sugarcane samples are analyzed by discriminant analysis (DA). First of all, it is very important to select an appropriate spectrum range for investigation. Because the absorption range of sugarcane is between 0 and 1.5 THz, the spectrum data above 1.5 THz cannot be used to establish the model for calculation. In this paper we used a number of spectral ranges for the identification (Table 1). It is can be seen from Table 1 that the highest identification accuracy is 94.13% in the regions of 0.1–0.6, 0.1–1.0, 0.1–1.2, and 0.1–1.5 THz. In Fig. 1, the THz frequency domain spectra of the genetically modified sugarcane and non-genetically modified sugarcane overlap above 0.6 THz, which agrees with Table 1.

TABLE 1. Statistic Results of the Identification of Sugarcane Samples Using DA at Different Bands

Range, THz	Accuracy, %	Range, THz	Accuracy, %
0.1–0.2	61.32	0.1–0.9	87.44
0.1–0.3	65.67	0.1–1.0	94.13
0.1–0.4	63.71	0.1–1.1	88.36
0.1–0.5	84.24	0.1–1.2	94.13
0.1–0.6	94.13	0.1–1.3	90.85
0.1–0.7	88.79	0.1–1.4	91.58
0.1–0.8	86.93	0.1–1.5	94.13

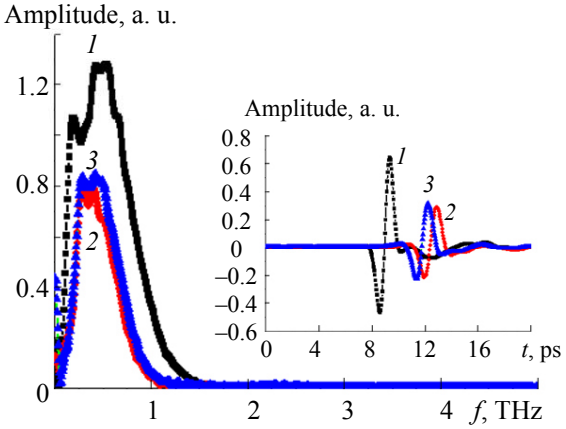

Fig. 1. Transmitted terahertz amplitude spectroscopy of genetically modified sugarcane (2), non-genetically modified sugarcane (3), and air (1). The inset indicates the time domain signals of the samples.

Spectral feature extracted. The THz spectral data of genetically modified sugarcane was collected from 0.1 to 1.5 THz in this paper, and 131,072 data points composed a spectral matrix whose size is 256×512. In order to reduce the amount of data, it is necessary to extract the feature spectrum. In this paper, PCA is utilized to extract the main information from the THz spectra of the sugarcane samples. PCA aims to reduce the dimensionality of spectral data and decrease the error. The data matrix X of PCA is defined as: $X = TP^T + E$, where T expresses the $n \times k$ scores matrix, P describes the $m \times k$ loadings matrix, and P is the error matrix.

Identification methods for terahertz spectral data. Different schemes are applied to identify the terahertz spectral data in this paper. The first scheme is based on LDA, the second scheme is based on the SVM-DA model, and the last scheme uses PLS-DA. All of the schemes are multivariate pattern recognition approaches.

Scheme 1: Linear discriminant analysis. LDA, or fisher linear discriminant (FLD), is a classical algorithm of pattern recognition. The basic idea of LDA is to map high-dimensional samples into the best identifiable vector space, and the aim of LDA is to achieve the effect of compression of the feature space dimensions and extract the classification information. Therefore, it is an effective method for feature extraction. In this paper, this method is utilized to reduce the dimensions of the spectral data.

Scheme 2: methods of identification using SVM-DA. SVM is a learning machine based on the theory of Vapnik–Chervonenkis (VC) and structural risk minimization (SRM). By seeking the minimum risks, it aims to improve the generalization ability of the learning machine, minimize empirical risks and the confidence limit, and obtain a good statistics with fewer samples. The purpose of SVM is to achieve the best generalization ability. SVM-DA is used for regression and the classification of multiple classes.

Scheme 3: partial least squares-discriminant analysis. PLS-DA is a multivariate statistical analysis method used for discriminant analysis. The principle of PLS-DA is training the characteristics of different samples, producing a training set, and verifying the credibility of it. The purpose of PLS-DA is to ameliorate the separation of samples in different groups. This technique usually uses 0 and 1 to predicate different classes. Therefore, when the sample value is closer to 0, the sample will belong to a particular class.

Figures of merit. When proposing a new identification methodology, it is important to confirm the figures of merit. In the identification approach for validation, there are several criteria for success. In order to establish the identification method for genetically modified sugarcane, the discrimination techniques of LDA, SVM-DA, and PLS-DA are compared in the aspect of statistical parameters (sensitivity, specificity, precision, and misclassification error) [26, 27]:

$$\text{sensitivity} = G_{MC}/(G_{MC} + G_{MIN}),\ \text{specificity} = N_{GMC}/(N_{GMC} + N_{GMIG}),$$

$$\text{precision} = G_{MC}/(G_{MC} + N_{GMIG}),\ \text{misclassification error} = (N_I/T_N)\times 100\%,$$

where G_{MC} is the proportion of genetically modified samples that were correctly classified, N_{GMIG} is the proportion of non-genetically modified samples that were incorrectly identified as genetically modified samples, N_{GMC} is the proportion of non-genetically modified samples that were classified correctly, G_{MIN} is the proportion of genetically modified samples that were incorrectly classified as non-genetically modified samples, N_I is the number of samples incorrectly classified, and T_N the total number of samples.

Experimental. The genetically modified and non-genetically sugarcane samples with a purity of above 99% are supplied by Sigma-Aldrich Shanghai Trading Co., Ltd. A total of 100 sugarcane samples (20 genetically modified sugarcane samples and 30 non-genetically modified sugarcane samples) of similar sizes are prepared. All genetically modified sugarcane samples are distinguished and labeled as genetically modified by the manufacturers.

Spectra are obtained with a terahertz time-domain spectrometer. Each spectrum consists of an average of 30 scans collected from the spectrometer. In order to rectify the measurement error, the spectral data of 50 samples were pretreated by normalization. All spectra of the samples are adjusted to the baseline before being converted into ASCII format. All of the chemometrics methods are accomplished by using TQ Analyst V8.0 (Thermo Nicolet Corporation, Madison, WI, USA).

Results and discussion. *Spectral analysis.* Figure 2 shows the terahertz absorption spectra of genetically modified sugarcane. It can be seen from Fig. 2 that the absorption spectra can be divided into three regions. Region I indicates the absorption peak at 0.55 THz caused by the symmetric and asymmetric structure of the C-H bond. Region II presents the absorption peak at 0.75 THz that most likely resulted in the angle of the chemical bond of the C-H3 and C-H2. Region III points out the absorption peak at 1.43 THz, related to the structure of the C-O and C=O bond.

Figure 3 indicates the absorption spectra of genetically modified and non-genetically modified sugarcane. It is practically impossible to discover any difference between the absorption spectra of genetically modified and non-genetically modified sugarcane. Thus, it is difficult to identify genetically modified and non-genetically modified sugarcane by the absorption spectroscopy method. So, chemometric tools, including PCA, LDA, SVM-DA, and PLS-DA, are adopted to identify genetically modified and non-genetically modified sugarcane samples.

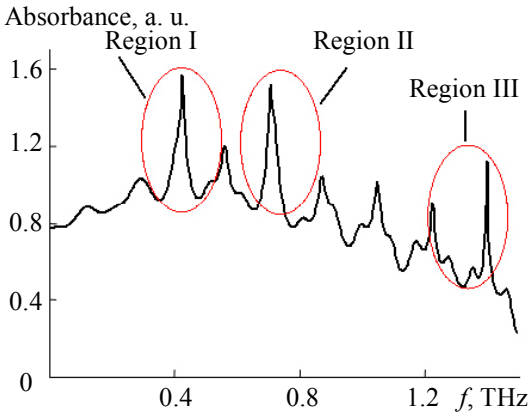

Fig. 2. Terahertz absorption spectroscopy of a genetically modified sugarcane.

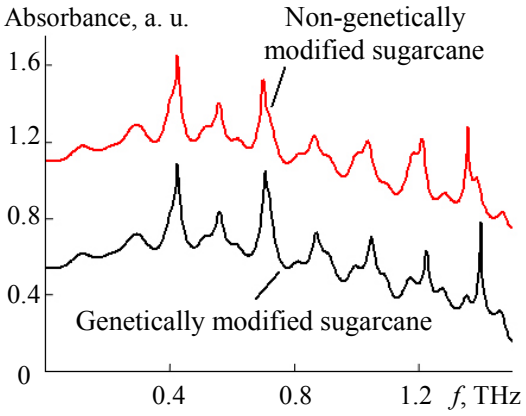

Fig. 3. Terahertz spectra of a genetically modified sugarcane and non-genetically modified sugarcane.

Principal component analysis. The PCA is used to extract feature information of sugarcane samples. PCA indicates that the first two eigenvectors apprehend more than 97.41% of the total variance. Figure 4 gives the score value of 50 samples by using PCA. The PCA score indicates that the samples are divided into two groups, but not all samples are separated correctly. The main reason is that PCA is not a reorganization technique but a feature information extract technique.

In Fig. 5, some samples, labeled 34, 48, and 50 (genetically modified samples), have a high value of Q-residuals versus Hotelling T^2. Hence, these samples should be excluded from the classification model owing to the negative effect that they can exert on the classification. As a consequence, 47 samples (17 genetically modified sugarcane samples and 30 non-genetically modified sugarcane samples) are used in the discriminant analysis.

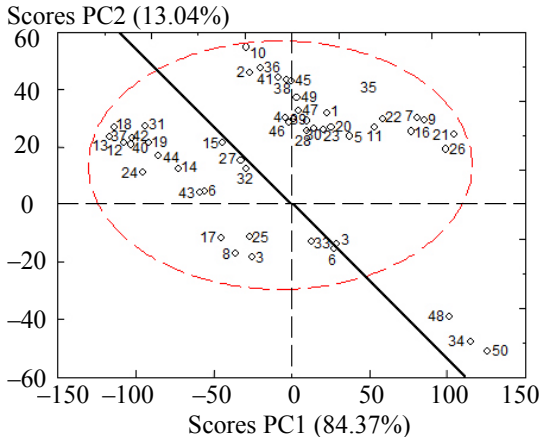

Fig. 4. Score plots of PC1 (84.37% variance) and PC2 (13.04% variance) for the sugarcane samples.

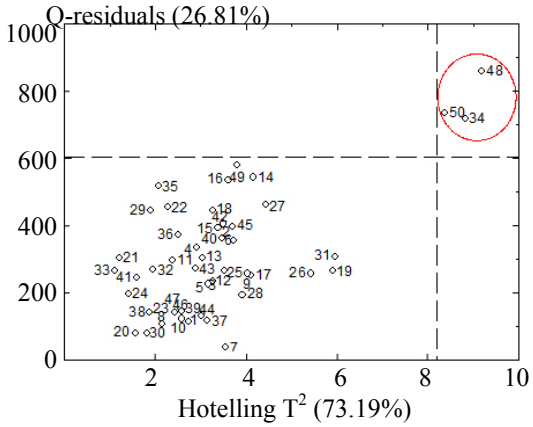

Fig. 5. The value of Q-residuals versus Hotelling T^2 of the samples.

In this study, all samples are divided into two groups: the calibration set and the validation set. The calibration set includes 27 samples (8 genetically modified sugarcane samples and 19 non-genetically modified sugarcane samples), and the validation set includes 20 samples (6 genetically modified sugarcane samples and 14 non-genetically modified sugarcane samples).

Linear discriminant analysis. For this technique, the first three principal components are picked for LDA. The different data preprocessing techniques, including mean centering (MC), multiplicative signal correction (MSC), first Savgol derivative and second Savgol derivative, are used to built different classification models. Table 2 exhibits the identification rate of each method. It can be seen from Table 2 that only the preprocessing method of MSC can reach a recognition rate of 100% and all other methods are more or less misjudging genetically modified ones in the test and validation sets.

TABLE 2. Performance of Each Method in the Identification of the Sugarcane Samples Using LDA

Parameter	Calibration				Validation			
	MC	MSC	1st derivative	2nd derivative	MC	MSC	1st derivative	2nd derivative
N_I	8	0	3	2	6	0	1	2
T_N	27	27	27	27	20	20	20	20
Misclassification error, %	29.63	0.0	11.11	7.4	30	0.0	5	10
G_{MC}	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
N_{GMIG}	0.47	0.0	0.16	0.09	0.68	0.0	0.17	0.11
N_{GMC}	0.39	1.0	0.73	0.77	0.19	1	0.73	0.75
G_{MIN}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sensitivity	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Specificity	0.45	1.0	0.82	0.90	0.22	1.0	0.81	0.87
Precision	0.68	1.0	0.86	0.92	0.60	1.0	0.85	0.90

Support vector machine-discriminant analysis. For this algorithm, different classification models are established by different data preprocessing methods, including MC, MSC, first Savgol derivative, and second Savgol derivative, which are performed with available software. All preprocessing methods can obtain a recognition rate of 100%, except derivatives methods. The recognition results of SVM-DA are shown in Table 3.

TABLE 3. Performance of Each Method in the Identification of the Sugarcane Samples Using SVM-DA

Parameter	Calibration				Validation			
	MC	MSC	1st derivative	2nd derivative	MC	MSC	1st derivative	2nd derivative
N_I	0	0	6	8	0	0	3	3
T_N	27	27	27	27	20	20	20	20
Misclassification error, %	0.0	0.0	22.2	29.6	0.0	0.0	15	15
G_{MC}	1.0	1.0	0.35	0.21	1.0	1.0	0.67	0.67
N_{GIMG}	0.0	0.0	0.06	0.04	0.0	0.0	0.17	0.17
N_{GMC}	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
G_{MIN}	0.0	0.0	0.59	0.73	0.0	0.0	0.24	0.24
Sensitivity	1.0	1.0	0.37	0.22	1.0	1.0	0.74	0.74
Specificity	1.0	1.0	0.94	0.96	1.0	1.0	0.85	0.85
Precision	1.0	1.0	0.85	0.84	1.0	1.0	0.79	0.79

Partial least squares-discriminant analysis. The same as SVM-DA, different data preprocessing methods (MC, MSC, first Savgol derivative, and second Savgol derivative) are employed to assemble different identification models. Table 4 presents the identification rate of each method. According to the Table 4, the PLS-DA method can obtain a discrimination rate of 100 %.

TABLE 4. Performance of Each Method in the Identification of the Sugarcane Samples Using PLS-DA

Parameter	Calibration				Validation			
	MC	MSC	1st derivative	2nd derivative	MC	MSC	1st derivative	2nd derivative
N_I	0	0	0	0	0	0	0	0
T_N	27	27	27	27	20	20	20	20
Misclassification error, %	0	0	0	0	0	0	0	0
G_{MC}	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
N_{GIMG}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N_{GMC}	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
G_{MIN}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sensitivity	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Specificity	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Precision	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Conclusion. Terahertz spectroscopy and chemometrics technique is a powerful tool to detect genetically modified sugarcane. Terahertz spectroscopy and chemometrics technique can provide the advantage of avolding time-consuming, and costly chemical and sensory analyses. Distinguishing genetically modified samples by this technique is meritorious, and this research indicates the potential of terahertz spectroscopy with chemometrics for genetically modified organisms. The aim of further researches is to establish more powerful discrimination models for other genetically modified organisms.

Acknowledgment. This work is supported by the Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (No.YQ16204) and the Science project of Jiangxi Education (No. GJJ161067).

REFERENCES

1. Jianjun Liu, Lili Mao, Jinfeng Ku, Jun He, *Opt. Quantum Electron*., **48**, No. 2, 167–173 (2016).
2. Jianjun Liu, Zhi Li, Fangrong Hu, Tao Chen, Yong Du, Haitao Xin, *J. Appl. Spectrosc*., **82**, No. 1, 104–110 (2015).
3. Jianjun Liu, Zhi Li, Fangrong Hu, Tao Chen, Yong Du, Haitao Xin, *Opt. Spectrosc*., **118**, No. 1, 175–180 (2015).
4. Jianjun Liu, Zhi Li, *Optik*, **125**, No. 23, 6867–6869 (2014).
5. K. Nakamura, H. Akiyama, N. Kawano, T. Kobayashi, K. Yoshimatsu, J. Mano, K. Kitta, K. Ohmori, A. Noguchi, K. Kondo, R. Teshima, *Food Chem*., **141**, 2618–2624 (2013).
6. M. Vaitilingom, H. Pijnenburg, F. Gendre, P. Brignon, *J. Agr. Food Chem*., **47**, 5261–5266 (1999).
7. G. Shan, S. K. Embrey, B. W. Schafer, *J. Agr. Food Chem*., **55**, 5974–5979 (2007).
8. G. Liu, W. Su, Q. Xu, M. Long, J. Zhou, S. Song, *Food Control*, **15**, 303–306 (2004).
9. Y. C. Shen, P. F. Taday, M. C. Kemp, *Proc. SPIE*, **5619**, 82–89 (2004).
10. A. D. Burnett, W. H. Fan, P. C. Upadhya, J. E. Cunningham, M. D. Hargreaves, T. Munshi, H. G. Edwards, E. H. Linfield, A. G. Davies, *Analyst*, **134**, 1658–1668 (2009).
11. J. El Haddad, F. de Miollis, J. Bou Sleiman, L. Canioni, P. Mounaix, B. Bousquet, *Anal. Chem*., **86**, 4927–4933 (2014).
12. H. Wu, E. J. Heilweil, A. S. Hussain, M. A. Khan, *J. Pharm. Sci*., **97**, 970–984 (2008).
13. Jianjun Liu, Lanlan Fan, *Int. J. Light Electron Opt*., **127**, 1957–1961 (2016).
14. Jianjun Liu, Zhi Li, Fangrong Hu,Tao Chen, Aijun Zhu, *Opt. Quantum Electron*., **47**, No. 2, 313–322 (2015).
15. Jianjun Liu, Zhi Li, Fangrong Hu, Tao Chen, Aijun Zhu, *Optik*, **125**, No. 23, 6914–6919 (2014).
16. Li Tiejun, Liu Jianjun,Shao Guifeng, Fan Lanlan, *Opt. Spectrosc*., **120**, No. 4, 660–665 (2016).
17. S. Shrestha, F. Kazama, T. Nakamura, *J. Hydroinform*., **10**, 43–56 (2008).
18. C. Kandemir-Cavas, E. Nasibov, *Turk. J. Biol*., **37**, 54–61 (2012).
19. Jianjun Liu, Zhi Li, Fangrong Hu, Tao Chen, Aijun Zhu, Yong Du, Haitao Xin, *Optik*, **126**, No. 19, 1872–1877 (2015).
20. L. Xie, Y. Ying, T. Ying, H. Yu, X. Fu, *Anal. Chim. Acta*, **584**, 379–384 (2007).
21. G. B. Alcântara, A. Bsrison, M. S. Santos, L. P. S. Santos, J. F. F. Toledo, A. G. Ferreira, *Orbital – Electron. J. Chem*., **2**, 41–52 (2010).
22. M. D. Salvador, A. M. Inarejos-Garcia, S. Gómez-Alonso, G. Fregapane, *Food Res. Int*., **50**, 250–258 (2013).
23. R. Aparicio, M. T. Morales, R. Aparicio-Ruiz, N. Tena, D. L. García-González, *Food Res. Int*., **54**, 2025–2038 (2013).
24. A. Koidis, M. T. Osorio, S. A. Haughey, C. T. Elliott, *Food Res. Int*., **60**, 66–75 (2014).
25. C. A. Nunes, *Food Res. Int*., **60**, 255–261 (2014).
26. R. G. Brereton, *Chemometrics for Pattern Recognition*, Chichester, John Wiley and Sons Ltd. (2009).
27. M. J. C. Pontes, R. K. H. Galvão, M. C. U. Araújo, P. N. T. Moreira, O. D. P. Neto, G. E. José, T. C. B. Saldanha, *Chemometr. Intell. Lab. Syst.*, **78**, 11–18 (2005).