Journal of Innovative Optical Health Sciences
Vol. 11, No. 2 (2018) 1850005 (13 pages)

DOI: 10.1142/S1793545818500050

Moving-window bis-correlation coefficients method for visible and near-infrared spectral discriminant analysis with applications

Lijun Yao*, Weiqun Xu*, Tao Pan*,‡,¶ and Jiemei Chen†,§,¶
*Guangdong Provincial Key Laboratory of Optical Fiber Sensing and Communications
Department of Optoelectronic Engineering
Jinan University, Guangzhou, China
†Department of Biological Engineering
Jinan University, Guangzhou, China
‡tpan@jnu.edu.cn
§tchjm@jnu.edu.cn

Received 11 April 2017
Accepted 24 July 2017
Published 18 August 2017

The moving-window bis-correlation coefficients (MW-BiCC) was proposed and employed for the discriminant analysis of transgenic sugarcane leaves and β-thalassemia with visible and near-infrared (Vis–NIR) spectroscopy. The well-performed moving-window principal component analysis linear discriminant analysis (MW-PCA–LDA) was also conducted for comparison. A total of 306 transgenic (positive) and 150 nontransgenic (negative) leave samples of sugarcane were collected and divided to calibration, prediction, and validation. The diffuse reflection spectra were corrected using Savitzky–Golay (SG) smoothing with first-order derivative ($d = 1$), third-degree polynomial ($p = 3$) and 25 smoothing points ($m = 25$). The selected waveband was 736–1054 nm with MW-BiCC, and the positive and negative validation recognition rates (V_REC$^+$, V_REC$^-$) were 100%, 98.0%, which achieved the same effect as MW-PCA–LDA. Another example, the 93 β-thalassemia (positive) and 148 nonthalassemia (negative) of human hemolytic samples were collected. The transmission spectra were corrected using SG smoothing with $d = 1$, $p = 3$ and $m = 53$. Using MW-BiCC, many best wavebands were selected (e.g., 1116–1146, 1794–1848 and 2284–2342 nm). The V_REC$^+$ and V_REC$^-$ were both 100%, which achieved the same effect as MW-PCA–LDA. Importantly, the BiCC only required calculating correlation coefficients between the spectrum of prediction sample and the average spectra of two types of calibration samples. Thus, BiCC was very simple in algorithm, and expected to obtain more applications. The results first confirmed the feasibility of distinguishing β-thalassemia and normal control samples by NIR spectroscopy, and provided a promising simple tool for large population thalassemia screening.

¶Corresponding authors.

Keywords: Visible and near-infrared spectroscopic discriminant analysis; transgenic sugarcane leaves; β-thalassemia; moving-window bis-correlation coefficients; moving-window principal component analysis linear discriminant analysis.

1. Introduction

Near-infrared (NIR) spectroscopy as a simple and quick tool has been effectively utilized in various fields for quantitative and qualitative analysis, such as agriculture,[1–6] food,[7–9] environment,[10,11] biomedicine,[12–16] petroleum industry,[17] and so on.[18] Spectral discriminant analysis is a methodology that uses computer pattern recognition to identify and classify samples based on the collected spectral data. Instead of the quantitative analysis for some components in the samples, its bases are the spectral overall features that are the spectral similarity of the samples of the same types and the spectral differences between samples of the different types. Pattern recognition technology based on NIR spectral information is presently an important research area,[18] such as distinction of different melon genotypes,[2] identification of transgenic sugarcane leaves,[9] and classification of multiple online petroleum industrial products.[17]

Principal component analysis linear discriminant analysis (PCA–LDA) is the commonly well-performed method for spectral discriminant analysis.[2,9,17,18] The extraction of feature information and dimension reduction were performed based on a spectral data matrix that corresponds to variables (e.g., wavelengths) and samples. In general, the original data matrix for the entire scan range is directly subjected to PCA–LDA. However, this process has its drawbacks. Substances have their own specific molecular absorbance bands and spectral data that involve instrumental noise disturbance bands; hence, the raw data matrix usually contains considerable interference noise. Inappropriate wavelength selection will certainly affect the quality of extracted feature information. In summary, wavelength selection is required in the sense of mathematics, physics and chemistry.

The moving-window waveband screening is a well-performed method for wavelength selection that uses each waveband as a window, and wavelength optimization is achieved by varying the size and position of the window. In the spectral quantitative analysis, numerous experimental results[4,8,11,12,19] indicate that the moving-window waveband screening can extract information effectively, eliminate noise disturbances, and improve spectral predictive capability significantly. For spectral discriminant analysis, the moving-window waveband screening was integrated with PCA–LDA, called moving-window PCA–LDA (MW-PCA–LDA), and was applied to the nondestructive discriminant analysis of transgenic sugarcane leaves with visible and near-infrared (Vis–NIR) spectroscopy.[9]

The bases of discriminant analysis are the spectral similarity of the samples of the same types and the spectral differences between samples of the different types. If the spectral waveband selection is appropriate, the spectral correlation coefficient can also reflect similarities and differences. Thus, a simpler spectral discriminant analysis method named moving-window bis-correlation coefficients (MW-BiCC) was proposed in this study.

Along with the development of agricultural biotechnology, transgenic sugarcane breeding receives more and more attention. In transgenic sugarcane breeding, the traditional molecular biology detection technologies (i.e., enzyme-linked immunosorbent assay (ELISA)) are complicated and cannot meet the needs of large-scale production. It is significant to develop the nondestructive discriminant method of transgenic sugarcane leaves with Vis–NIR spectroscopy.

Another example, thalassemia is a hemolytic genetic disease, and it affects individuals from many parts of the world, including southern China, where it has a high prevalence and incidence and has caused serious health damage.[20] In China, the rates of gene carriers are as high as 24.13% and 11.07% in the population of Guangxi and Guangdong provinces, respectively.[21,22] Among them, β-thalassemia is a common type. Up to now, the disease cannot be cured, except through hematopoietic stem cell transplantation. The most fundamental prevention measures include premarital and prenatal thalassemia screening in a large population.

Hematologic phenotypic analysis is a first-line screening method for thalassemia diagnosis, which includes full blood cell (FBC) analysis and

hemoglobin component analysis (HCA). Among them, mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) are preliminary screening indicators for thalassemia, which are measured by FBC. The indicator hemoglobin A_2(HbA_2) is used to further distinguish β-thalassemia, which are measured by HCA. The phenotype-positive subjects for β-thalassemia are those with $MCH \leq 27.0\,pg$ (or $MCV \leq 80.0\,fL$) and $HbA_2 > 3.5\%$.[20–22] These procedures are relatively complicated because they require different chemical reagents and different measuring instruments. It is not suitable for thalassemia screening in large population. To the best of our knowledge, the use of NIR spectroscopy to directly discriminate β-thalassemia has not been reported yet. It is very significant to develop the rapid and simple discriminant method of large population thalassemia screening with NIR spectroscopy.

In this study, the Vis–NIR spectroscopic discriminant analysis of transgenic sugarcane leaves and β-thalassemia were taken as the examples to evaluate the performance of the proposed MW-BiCC, and the MW-PCA–LDA method was also performed for comparison. On the other hand, Savitzky–Golay (SG) smoothing,[4,9,13,23] which is an efficient spectral preprocessing method with a wide scope of application and a variety of different smoothing modes, was used first for the pretreatment of diffuse reflectance spectral data.

2. Materials and Methods

2.1. *Samples and reference methods*

2.1.1. *Sugarcane leaf samples*

The sugarcane strains of Xintaitang (ROC), which are the widely cultivated varieties of sugarcane in southern China, were adopted. The leaves of the field-planting sugarcane strains in the elongation stage were collected as experimental samples. Some of the sugarcane receptors contained both *Bacillus thuringiensis* and *Bialaphos resistance* genes, and the others were nontransgenic sugarcane strains. In chronological order, the samples were divided into two groups. The first group (300 samples) collected on the first day was used for modeling, whereas the second group (156 samples) collected on the second day was used for validation. The second group excluded in modeling was used to validate and achieve an objective evaluation.

One leaf was randomly collected from each sugarcane strain. The collected samples were washed and stored at room temperature for 2 h to dry naturally and equilibrate to the experimental environment before collection of the Vis–NIR diffuse reflectance spectra. Each leaf was cut into several neat little leaves. The leaflets were flattened and overlaid into a round sample cup (face up), so that the spot of light could be covered. They were used for the measurement of diffuse reflectance spectra at room temperature.

Enzyme-linked immunosorbent assay (ELISA) was applied to check the genes and expression for each sample. The instruments were an ELISA kit BT-Cry1Ab/1Ac (AGDIA, USA) and a microplate reader iMark (Bio-rad, USA). After the above measurement, the first group included 200 transgenic (positive) and 100 nontransgenic (negative) samples of sugarcane leaves; while the second group included 106 transgenic and 50 nontransgenic samples of sugarcane leaves.

2.1.2. *Human hemolytic solution samples*

A total of 241 human peripheral blood samples were collected from the same hospital and placed in 0.2% ethylenediaminetetraacetic acid-containing tubes. As the blood samples were collected and used in this study, the informed consent of all individual participants was obtained. Experiments were performed in compliance with the relevant laws and institutional guidelines and approved by local medical institution, which obtained the informed consent from all subjects. The β-thalassemia indicators (MCH, MCV and HbA_2) of these samples were measured via two existing clinical methods from the same hospital. In which, the MCH, and MCV values were measured with a BC-3000Plus Blood Cell Analyzer (Shenzhen Mairui, China) using FBC count method, and the HbA_2 values were measured with a VARIANT™ Hemoglobin Testing System (Bio-Rad Laboratories, USA) using high-pressure liquid chromatography analysis method. According to the cutoff values ($MCH \leq 27.0\,pg$ or $MCV \leq 80.0\,fL$ and $HbA_2 > 3.5\%$) for β-thalassemia, the 241 samples are identified as 93 β-thalassemia (positive) and 148 nonthalassemia (negative).

Thalassemia is a hemoglobin disease, which is closely related to the erythrocyte. Erythrocyte is the tangible part of the blood. When the light penetrates the erythrocytes, scattering occurs, affecting

the accuracy of spectral analysis. If the distilled water is added to the blood, due to osmotic pressure, the blood cell membrane will rupture and become hemolysate solution sample. This process is called hemolysis of distilled water. Since the homogeneity of the hemolytic solution sample is significantly better than that of the whole blood sample, the interference of the light scattering is reduced. In the previous study,[14] the quantitative analysis for thalassemia indicators (MCH and MCV) has achieved very good effect when the $2\times$ dilute hemolytic solution samples were adopted for NIR measurement. The $2\times$ dilute hemolytic solution samples were also used in this study. To do this, one volume of peripheral blood sample was diluted with an equal volume of distilled water to rupture the erythrocytes, and the hemolysate samples were obtained.

2.2. *Spectral experiments*

The spectra were collected by the XDS Rapid Content™ grating spectrometer (FOSS, Denmark). The scanning range was a part of visible and whole NIR regions of 400–2498 nm equipped with a silicon detector (400–1100 nm) and a plumbous sulfide detector (1100–2498 nm), respectively. The spectra were recorded with 4-nm spectral resolution with a 2-nm wavelength interval, and a total of 15 scans were averaged for every spectrum to overcome the inhomogeneity.

The spectrometer was equipped with a diffuse reflection accessory for the measurement of sugarcane leaf samples, and a transmission accessory with a 2-mm cuvette for the measurement of human hemolytic solution samples. In order to achieve stability, each sample was measured three times and the mean spectrum of the measurements was used for modeling and validation. The spectra were measured at $25\pm1^{\circ}$C and $46\pm1\%$ relative humidity.

2.3. *Calibration, prediction, and validation processes*

The Kennard–Stone (KS) algorithm[24,25] is an effective partition method for sample experiment. A "distance" between every two samples was first defined (e.g., Euclidean distance or Mahalanobis distance). A small distance value indicates a high similarity between two samples, and vice versa indicates low similarity. Based on the defined distance, the KS algorithm effectively selects an appropriate subset of samples which can uniformly and sufficiently represent the entire sample space.

2.3.1. *Sugarcane leaf samples*

The first group of samples was used for modeling. Using the KS algorithm, the separate divisions for calibration and prediction sets were performed for positive and negative groups to achieve uniformity and representativeness. The modeling samples (200 positive and 100 negative) were divided into calibration (100 positive and 50 negative) and prediction (100 positive and 50 negative) sets. Parameter optimization was performed according to the prediction recognition rate. The posterior group of samples (106 positive and 50 negative) excluded in the modeling was used for validation.

2.3.2. *Human hemolytic solution samples*

First, 96 samples were randomly selected as the validation set (58 negative and 38 positive). Then, the remaining 145 samples were used as modeling set (90 negative and 55 positive). Using the KS algorithm, the modeling samples were divided into calibration (45 negative and 28 positive) and prediction (45 negative and 27 positive) sets.

2.4. *Spectral discriminant analysis methods with moving-window waveband screening*

2.4.1. *Proposed bis-correlation coefficients method*

The BiCC method in any fixed waveband is introduced specifically as follows:

First, the spectral waveband was assumed to contain N consecutive wavelengths. The average spectra of negative and positive calibration samples were calculated and denoted as $(A_1^-, A_2^-, \ldots, A_N^-)$ and $(A_1^+, A_2^+, \ldots, A_N^+)$, respectively, which could be used as the characteristic spectra of negative and positive calibration samples.

Second, for the kth prediction sample, the spectrum was denoted as $(A_{1,k}, A_{2,k}, \ldots, A_{N,k})$, two correlation coefficients between the spectrum and the average spectra of two types of calibration samples were further calculated and expressed as R_k^- and R_k^+, as shown in formulas (1) and (2), where $A_{\mathrm{Ave},k}$, A_{Ave}^- and A_{Ave}^+ were the mean values of

$(A_{1,k}, A_{2,k}, \ldots, A_{N,k})$, $(A_1^-, A_2^-, \ldots, A_N^-)$ and $(A_1^+, A_2^+, \ldots, A_N^+)$, respectively:

$$\mathrm{R}_k^- = \frac{\sum_{i=1}^{N}(A_i^- - A_{\mathrm{Ave}}^-)(A_{i,k} - A_{\mathrm{Ave},k})}{\sqrt{\sum_{i=1}^{N}(A_i^- - A_{\mathrm{Ave}}^-)^2 \sum_{i=1}^{N}(A_{i,k} - A_{\mathrm{Ave},k})^2}}, \tag{1}$$

$$\mathrm{R}_k^+ = \frac{\sum_{i=1}^{N}(A_i^+ - A_{\mathrm{Ave}}^+)(A_{i,k} - A_{\mathrm{Ave},k})}{\sqrt{\sum_{i=1}^{N}(A_i^+ - A_{\mathrm{Ave}}^+)^2 \sum_{i=1}^{N}(A_{i,k} - A_{\mathrm{Ave},k})^2}}. \tag{2}$$

The different value between the BiCC was further calculated and denoted as $\Delta\mathrm{R}_k$, as shown in formula (3), if $\Delta\mathrm{R}_k > 0$, then the kth prediction sample was determined as a negative sample; if $\Delta\mathrm{R}_k \leq 0$, then the kth prediction sample was determined as a positive sample.

$$\Delta\mathrm{R}_k = \mathrm{R}_k^- - \mathrm{R}_k^+. \tag{3}$$

Finally, referring to the known types of the prediction samples, the prediction recognition rate was calculated easily and denoted as P_REC.

2.4.2. *PCA–LDA method*

PCA–LDA is a commonly well-performed method for spectral discriminant analysis.[1,2,9,17,18]

First, the PCA was performed based on the absorbance matrix of the calibration set, and then the extracted principal components are sorted by variance from large to small. The preceding several principal components based on the cumulative contribution rate mainly reflect the information of the original data, while the following principal components reflect the noise disturbances. The purpose of this study is the spectral discrimination analysis of the samples. In order to determine the surface classification and facilitate visualization, the three-dimensional PCA model with the first three principal components was adopted in the next LDA procedure. The detailed procedure can be found in the previous studies.[1,18]

2.4.3. *MW-BiCC and MW-PCA–LDA methods*

The BiCC and PCA–LDA combined with moving-window waveband screening were called MW-BiCC and MW-PCA–LDA, respectively.

Considering the position and length of the wavebands, the search parameters of moving-window waveband screening were set as follows: (1) initial wavelength (I) and (2) number of wavelengths (N). For the sugarcane leaves, dataset adopted the whole scanning region (400–2498 nm), I was set to $I \in \{400, 402, \ldots, 2498\}$. To lessen the workload and ensure representativeness, N was set as $N \in \{3, 4, \ldots, 100\} \cup N \in \{3, 4, \ldots, 100\} \cup \{100, 110, \ldots, 200\} \cup \{220, 240, \ldots, 860\} \cup \{1050\}$. For the human hemolytic solution, dataset adopted the whole NIR region (780–2498 nm), I was set to $I \in \{780, 782, \ldots, 2498\}$, and N was set as $N \in \{3, 4, \ldots, 100\} \cup \{100, 110, \ldots, 200\} \cup \{220, 240, \ldots, 860\}$. For each waveband that corresponds to a combination of parameters (I, N), the BiCC and PCA–LDA models were established, and the optimal parameters can be preferred according to the maximum P_REC.

Furthermore, the validation samples excluded in the modeling procedure were applied to validate the selected models screening by MW-BiCC and MW-PCA–LDA. According to the genuine genotypes type of each validation samples and the number of correctly recognized validation samples, the recognition rate can be calculated easily, and was denoted as V_REC. The validation recognition rates of positive and negative samples were calculated and denoted as V_REC$^+$ and V_REC$^-$, respectively.

The computer algorithms for the two methods were designed using MATLAB V7.6.

3. Results and Discussion

The Vis–NIR spectra of 306 positive and 150 negative samples of sugarcane leaves in the whole scanning region (400–2498 nm) are shown in Fig. 1. The NIR spectra of 93 positive and 148 negative samples of human hemolytic solution in the whole NIR region (780–2498 nm) are shown in Fig. 2. As shown in Figs. 1 and 2, the spectra of negative and positive samples were overlapping, thereby resulting in no obvious spectral differences for direct discriminant analysis. Figures 1 and 2 show that the baseline deviations (drifts) of the spectra of different samples are serious.

3.1. *Full spectral models*

For the sugarcane leaves' dataset, the BiCC and PCA–LDA models with the whole scanning region (400–2498 nm) were established. The corresponding modeling effects are summarized in Table 1. The P_REC values of the two methods were 80.0% and

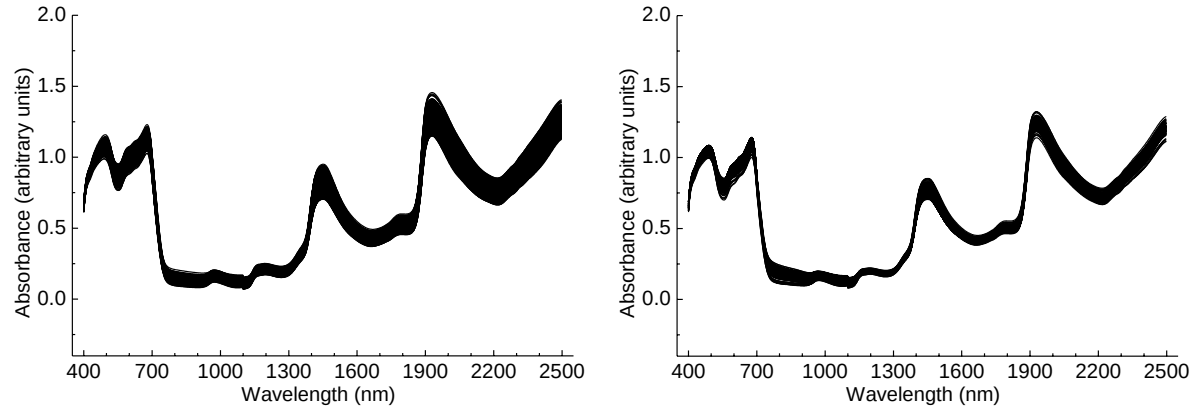

Fig. 1. Vis–NIR spectra of sugarcane leaf samples for (a) 306 positive and (b) 150 negative.

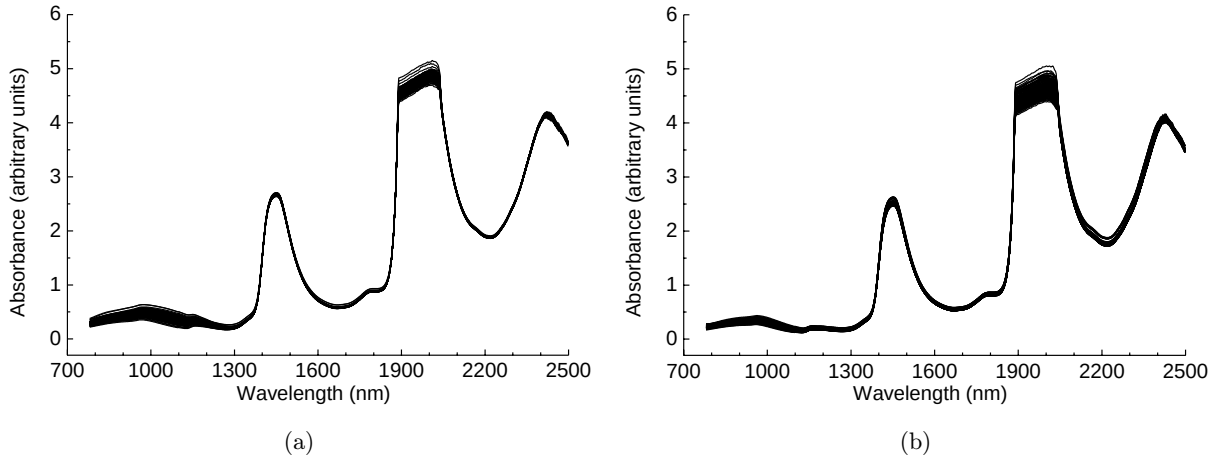

Fig. 2. Vis–NIR spectra of human hemolytic solution samples for (a) 93 positive and (b) 148 negative.

75.3%, respectively. The results show that the spectral identification was unsatisfactory when using the spectroscopy data without pretreatment.

For the human hemolytic solution dataset, the BiCC and PCA–LDA models with the whole NIR region (780–2498 nm) were also established. The corresponding modeling effects are summarized in Table 2. The *P*_REC values of the two methods were 90.3% and 88.9%.

In the following, the spectral data were preprocessed, and then the modeling was performed.

Table 1. Modeling effects of the BiCC and PCA–LDA models with the whole scanning region for the sugarcane leaf samples.

Method	Waveband (nm)	*N*	*P*_REC
Without pretreatment			
BiCC	400–2498	1050	80.0%
PCA–LDA			75.3%
With SG smoothing			
BiCC	400–2498	1050	92.7%
PCA–LDA			92.0%

Notes: *N*: number of wavelengths; *P*_REC: prediction recognition rate; BiCC: bis-correlation coefficients; PCA–LDA: principal component analysis linear discriminant analysis.

3.2. *Full spectral models with SG smoothing*

3.2.1. *Sugarcane leaves' dataset*

The parameters of SG smoothing include order of derivatives (d), degree of polynomial (p) and number of smoothing points (m, odd). In the previous study,[9] the SG smoothing mode with first-order derivative, third-degree polynomial and 25 smoothing points ($d = 1$, $p = 3$ and $m = 25$) was used and achieved a better prediction effect of PCA–LDA model for the sugarcane leaf samples.

Table 2. Modeling effects of the BiCC and PCA–LDA models with the whole NIR region for the human hemolytic solution samples.

Method	Waveband (nm)	N	P_REC
Without pretreatment			
BiCC	780–2498	860	90.3%
PCA–LDA			88.9%
With SG smoothing			
BiCC	780–2498	860	94.4%
PCA–LDA			93.1%

Notes: N: number of wavelengths; P_REC: prediction recognition rate; BiCC: bis-correlation coefficients; PCA–LDA: principal component analysis linear discriminant analysis.

In the present study, the SG mode ($d = 1, p = 3$ and $m = 25$) was still applied to BiCC model.

The corresponding SG derivative spectra are shown in Fig. 3. The baseline deviations (drifts) of the spectra of different samples are significantly reduced. In addition, as shown in Figs. 3(c) and 3(d), some difference in positive and negative samples was observed in the waveband of 736–1054 nm. The BiCC and PCA–LDA models with SG smoothing pretreatment were further established. The corresponding modeling effects are also summarized in Table 1. For the BiCC and PCA–LDA models with SG smoothing pretreatment, the predictive discrimination rates (P_REC) were improved to 92.7% and 92.0%, respectively. These results show that the SG smoothing can reduce spectral noise and improve spectral recognition ability. But, the full spectral models adopted 1050 wavelengths with high parameter complexity.

3.2.2. *Human hemolytic solutions' dataset*

In the previous study,[13] the SG smoothing mode with first-order derivative, third-degree polynomial

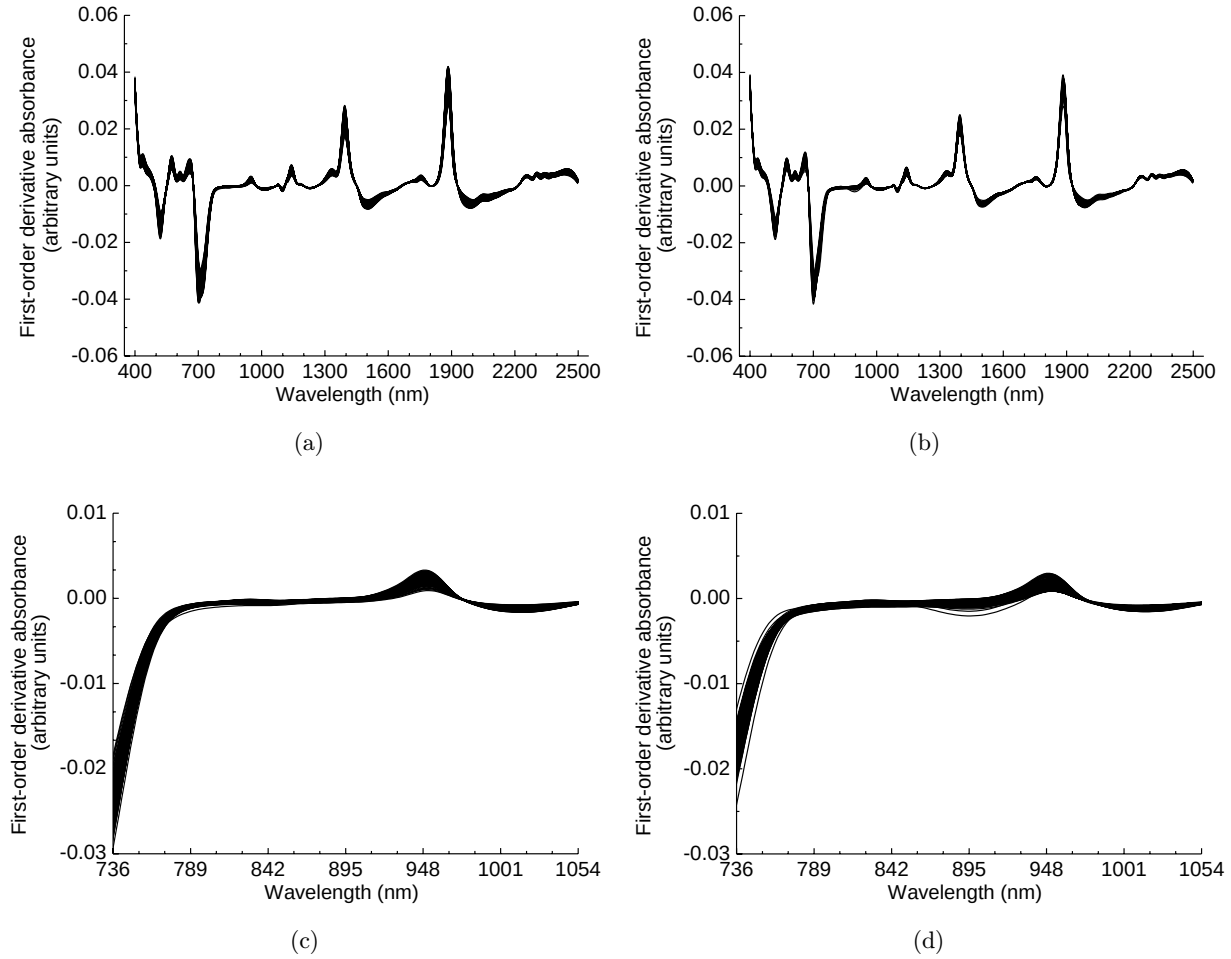

Fig. 3. SG derivative spectra of sugarcane leaves samples for (a) 306 positive (400–2498 nm), (b) 150 negative (400–2498 nm), (c) 306 positive (736–1054 nm) and (d) 150 negative (736–1054 nm).

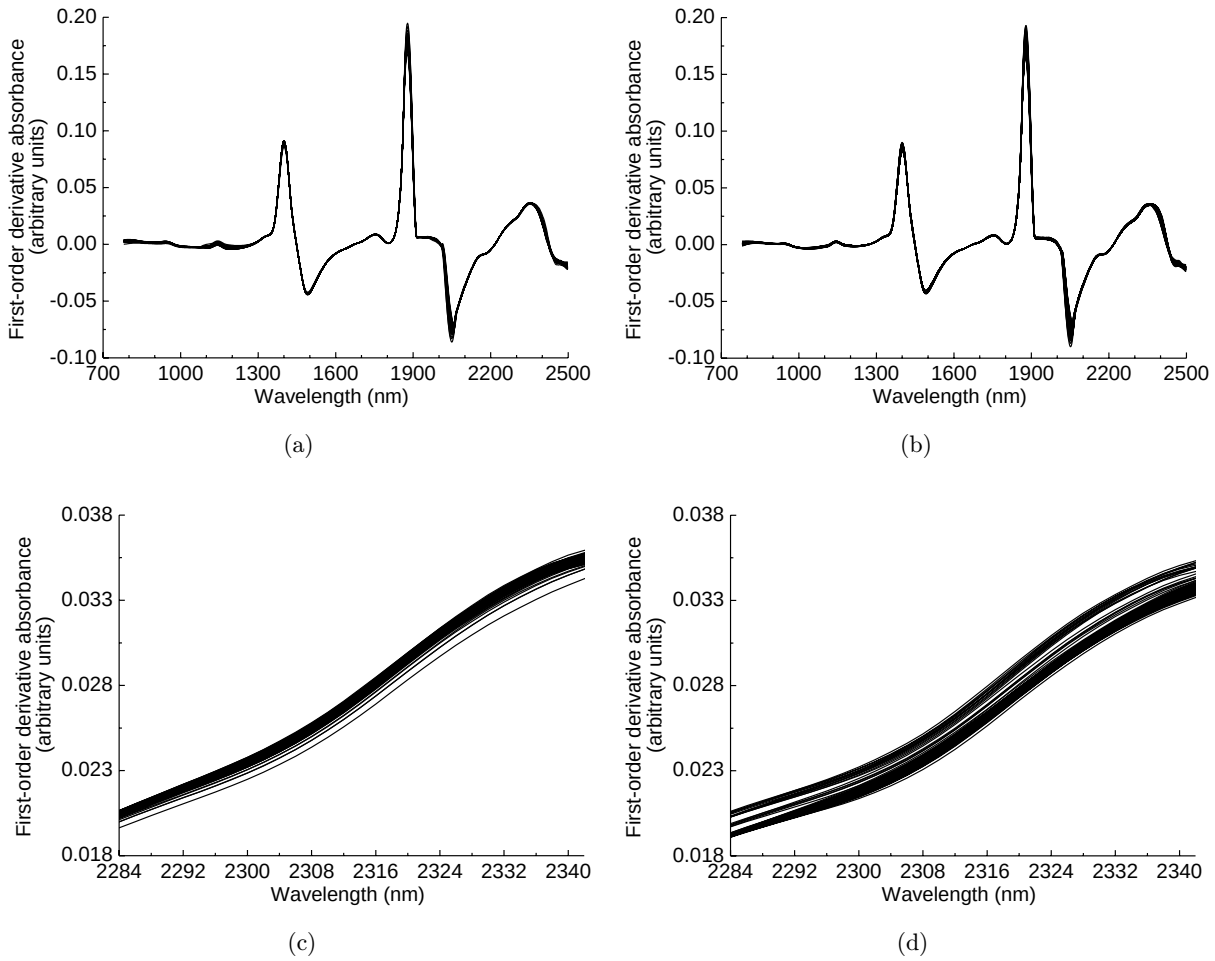

Fig. 4. SG derivative spectra of human hemolytic solution samples for (a) 93 positive (780–2498 nm), (b) 148 negative (780–2498 nm), (c) 93 positive (2284–2342 nm) and (d) 148 negative (2284–2342 nm).

and 53 smoothing points ($d = 1$, $p = 3$ and $m = 53$) was used and achieved a better prediction effect of PLS model for the human serum samples. In the present study, the SG mode ($d = 1$, $p = 3$ and $m = 53$) was tried to BiCC model.

The corresponding SG derivative spectra are shown in Fig. 4. The baseline deviations (drifts) of the spectra of different samples are significantly reduced. In addition, as shown in Figs. 4(c) and 4(d), some difference in positive and negative samples was observed in the waveband of 2284–2342 nm. The BiCC and PCA–LDA models with SG smoothing pretreatment were also further established. The corresponding modeling effects are also summarized in Table 2. For the BiCC model with SG smoothing, the *P*_REC was improved to 94.4%. For the PCA–LDA model with SG smoothing, the *P*_REC was improved to 93.1%.

In order to extract efficient information, and eliminate noise, the waveband optimizations were further carried out by MW-BiCC and MW-PCA–LDA methods after the SG smoothing.

3.3. *MW-BiCC models*

3.3.1. *Sugarcane leaves' dataset*

The average spectra of negative and positive samples are shown in Fig. 5. Notable differences were observed between the two spectra, particularly around the three peaks at 678, 1450, and 1928 nm and four valleys at 552, 800, 1666, and 2216 nm. The absorption of positive samples around the valley at 800 nm was remarkably lower than that of the negative samples; whereas, at the three peaks and other three valleys, the positive samples absorption was remarkably higher than that of the negative

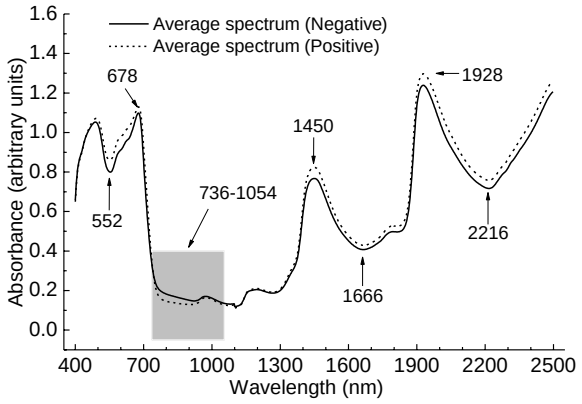

Fig. 5. Average spectra of positive and negative samples of sugarcane leaves.

samples. These differences indicate the feasibility of identifying transgenic and nontransgenic samples.

Using the MW-BiCC method based on the SG derivative spectra, the selected optimal I and N were 736 nm and 160, respectively. The corresponding waveband was 736–1054 nm, which covered part of the Vis–NIR combined region. As shown in Table 3, the corresponding prediction recognition rate (P_REC) achieved 98.0%, which was evidently higher than one of the BiCC models (see Table 1). Moreover, a minority of wavelengths (i.e., $N = 160$) was used in the selected model, so the scope of the wavelength was significantly narrowed. The selected waveband (736–1054 nm) contained the spectral valley at 800 nm, and the absorption of positive samples was remarkably lower than that of the negative samples in the selected waveband (seen in Fig. 5). This region was related to the fourth overtone of C–H (CH and CH_2) and the third overtone of O–H (H_2O and Ar–OH).[26] In fact, transgenic sugarcane leaves contain BT and BR proteins expressed by *Bacillus thuringiensis* and *Bialaphos resistance* genes.[27,28]

Table 3. Modeling effects of the selected MW-BiCC and MW-PCA–LDA models with SG smoothing for the sugarcane leaf samples.

Method	Waveband (nm)	N	P_REC
MW-BiCC	736–1054	160	98.0%
MW-PCA–LDA	756–1094	170	97.3%

Notes: N: number of wavelengths; P_REC: prediction recognition rate; MW-BiCC: moving-window bis-correlation coefficients; MW-PCA–LDA: moving-window principal component analysis linear discriminant analysis.

These two proteins do not exist in nontransgenic sugarcane leaves. Therefore, there are some differences in the molecular structure of the positive and negative transgenic sugarcane leaf samples. The protein molecules contain a large amount of hydrogen-containing groups, which have absorption information in the NIR region.

3.3.2. *Human hemolytic solutions' dataset*

The average spectra of negative and positive samples are shown in Fig. 6. The differences were observed between the two spectra, particularly around the two peaks at 1450 nm and 1950 nm and three valleys at 1126, 1666, and 2216 nm. The positive samples absorption was remarkably higher than that of the negative samples. These differences indicate the feasibility of identifying positive and negative samples. In fact, the β-thalassemia is caused by partial or total mutations that reduce or abolish the synthesis of β-globin chains of the hemoglobin molecule, which will result in hemolytic anemia.[20–22] And the blood molecules in the nonthalassemia sample do not have these changes. Therefore, there are some differences in the molecular structure of the positive and negative samples.

Using the MW-BiCC method based on the SG derivative spectra, a lot of wavebands were selected and their P_REC achieved 100%. The saturate absorption regions appear around the two peaks at 1450 nm and 1950 nm (seen in Fig. 6), implying a strong absorption of water molecules and high noise levels, which need to be avoided. Around the three valleys at 1126, 1666 and 2216 nm, the selected wavebands whose P_REC achieved 100% and contained the least wavelengths were 1116–1146 nm

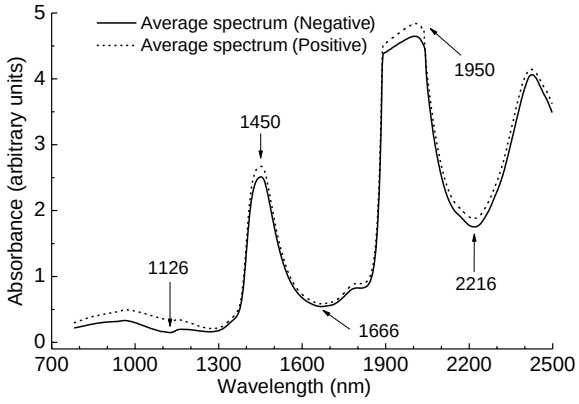

Fig. 6. Average spectra of positive and negative samples of human hemolytic solutions.

($N = 16$), 1794–1848 nm ($N = 28$) and 2284–2342 nm ($N = 30$), respectively. They are located in the third overtone, the first overtone and the combination regions, respectively. These models have the best predictive effect and the lowest parameter complexity.

3.4. *MW-PCA–LDA models*

3.4.1. *Sugarcane leaves' dataset*

By using the MW-PCA–LDA method based on the SG derivative spectra, the selected optimal I and N values were 756 nm and 170, respectively. The corresponding waveband was 756–1094 nm, which covered part of the Vis–NIR region.

As shown in Table 3, the corresponding prediction recognition rate (P_REC) achieved 97.3% with $N = 170$. Therefore, the selected MW-PCA–LDA model was significantly better than the PCA–LDA model (see Table 1) in two aspects of prediction performance and parameter complexity.

The position of the selected wavebands (756–1094 nm) was also in the spectral valley at 800 nm, and the two selected wavebands screening by MW-BiCC and MW-PCA–LDA methods were consistent similarly. The MW-BiCC achieved slightly better modeling effect (P_REC), but adopted less wavelengths ($N = 160$).

3.4.2. *Human hemolytic solutions' dataset*

Using the MW-PCA–LDA method based on the SG derivative spectra, a lot of wavebands were also selected and their P_REC achieved 100%. Around the three valleys at 1126, 1666, and 2216 nm, the selected wavebands whose P_REC achieved 100% and contained the least wavelengths were 1088–1266 nm ($N = 90$), 1634–1852 nm ($N = 110$) and 2180–2318 nm ($N = 70$), respectively. However, these wavebands almost covered the corresponding wavebands obtained by the MW-BiCC and contained more wavelengths.

3.5. *Validation*

3.5.1. *Sugarcane leaves' dataset*

The validation samples (106 positive and 50 negative samples) excluded in the modeling procedure were used for validating the selected MW-BiCC (736–1054 nm) and MW-PCA–LDA (756–1094 nm) models with SG derivative spectra. Table 4 summarizes the obtained values of V_REC, V_REC$^+$, and V_REC$^-$. As shown in Fig. 7(a), the validation samples' plot on the two-dimensional (2D) principal component space was clearly classified into two groups using the different values (ΔR) of BiCC. Moreover, as shown in Fig. 8(a), the validation samples' plot on the three-dimensional (3D) principal component space was clearly classified into two groups. Only one negative sample was misjudged as a positive sample (false positive).

3.5.2. *Human hemoytic solutions' dataset*

The validation samples (38 positive and 58 negative samples) excluded in the modeling procedure were used for validating the selected MW-BiCC (1116–1146, 1794–1848 and 2284–2342 nm) and MW-PCA–LDA (1088–1266, 1634–1852 and 2180–2318 nm) models with SG derivative spectra. Their V_REC, V_REC$^+$, and V_REC$^-$ values were all 100%.

For simplicity, the BiCC model with 1116–1146 nm and PCA–LDA model with 1088–1266 nm were taken as examples. As shown in Figs. 7(b) and 8(b), the validation samples' plot on the 2D and

Table 4. Validation effects of the selected MW-BiCC and MW-PCA–LDA models for the sugarcane leaf samples.

Method	Waveband (nm)	N	V_REC	V_REC$^+$	V_REC$^-$
MW-BiCC	736–1054	160	99.4%	100%	98.0%
MW-PCA–LDA	756–1094	170	99.4%	100%	98.0%

Notes: N: number of wavelengths; P_REC: prediction recognition rate; V_REC: validation recognition rate; V_REC$^+$: validation recognition rate of positive samples; V_REC$^-$: validation recognition rate of negative samples; MW-BiCC: moving-window bis-correlation coefficients; MW-PCA–LDA: moving-window principal component analysis linear discriminant analysis.

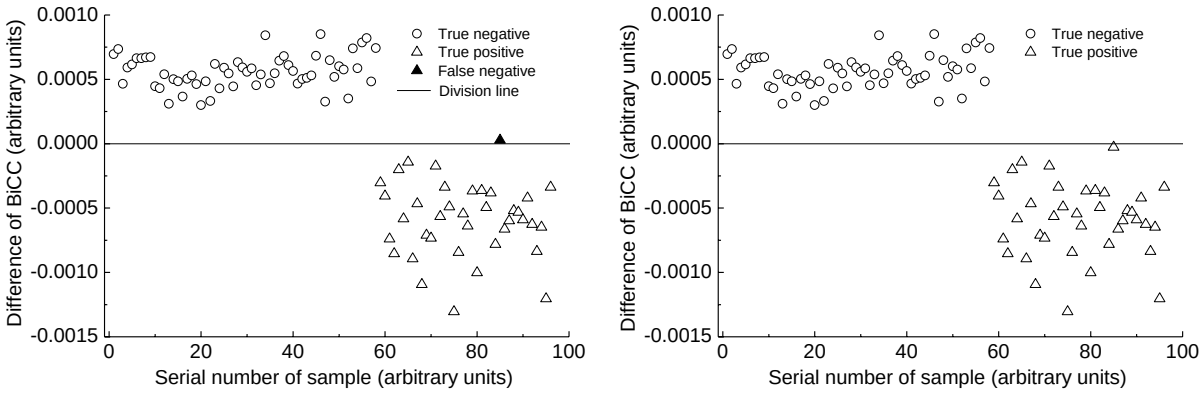

Fig. 7. 2D diagrams of validation samples classified as positive and negative with MW-BiCC for (a) sugarcane leaves and (b) human hemolytic solutions.

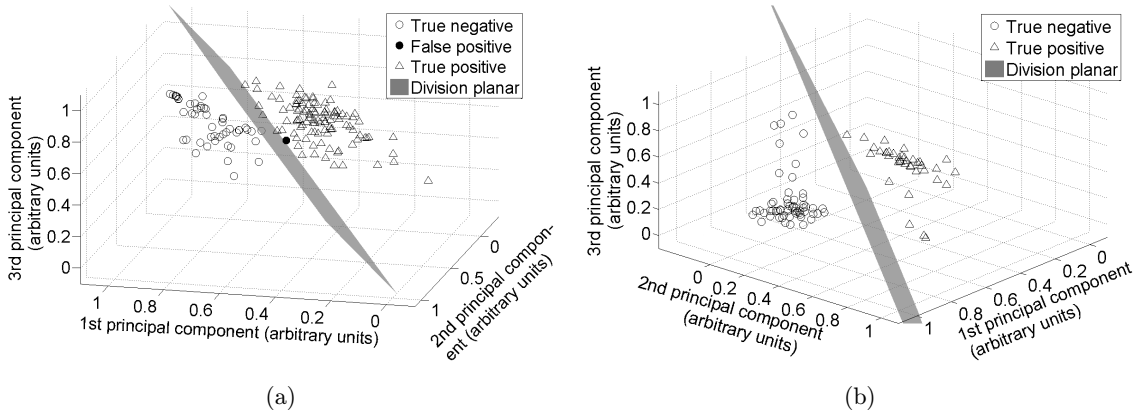

Fig. 8. 3D diagrams of validation samples classified as positive and negative with MW-PCA–LDA for (a) sugarcane leaves and (b) human hemolytic solutions.

the 3D diagrams were clearly classified into two groups, respectively.

The results indicated that the moving-window waveband screening applied to BiCC and PCA–LDA models could effectively extract spectral information, eliminate noise disturbances, and significantly improve spectral pattern recognition capability. Furthermore, the number of used wavelengths could be largely reduced.

The results also showed that the selected MW-BiCC model achieved the same validation effect (V_REC, V_REC$^+$, V_REC$^-$) as the selected MW-PCA–LDA model. But, the PCA–LDA method comprised a number of procedures, such as calculating the loading and score matrices, determining the expression of the cutoff planar or line, and selecting the optimal principal component combination. For the BiCC, the average spectra of the negative and positive calibration samples were calculated first. Then, the correlation coefficients between the spectrum of the prediction sample and the two average spectra were calculated; the type of the sample could be determined according to the size of the correlation coefficients. Therefore, the BiCC was very simple in terms of algorithm.

On the other hand, the experimental results first confirmed the feasibility of distinguishing β-thalassemia and normal control samples by NIR spectroscopy of human hemolytic solutions.

4. Conclusion

A novel spectral discriminant analysis method (i.e., MW-BiCC) was proposed based on Vis–NIR spectroscopy. The different types of samples (e.g., transgenic and nontransgenic sugarcane leaves,

human hemolytic solutions of β-thalassemia and normal control) have different molecular structures, so that their spectral characteristics are also different in the specific waveband. The BiCC method used the average spectrum to define the spectral characteristics and the correlation coefficient to evaluate the difference. Through the waveband selection, the MW-BiCC method highlighted the differences between spectral characteristics of different types of samples (e.g., Figs. 3 and 4), and achieved accurate spectral discriminant analysis.

The experimental results of the spectral discriminant analysis of transgenic sugarcane leaves and β-thalassemia indicated that the MW-BiCC was an efficient method. In the modeling process, the MW-BiCC achieved slightly better prediction effect (P_REC) than the MW-PCA–LDA; whereas in the validation process, the two methods had the same prediction effect (V_REC, V_REC$^+$, V_REC$^-$). Importantly, the BiCC required only calculating the correlation coefficients between the spectrum of prediction sample and average spectra of the two types of the calibration samples. Thus, the BiCC was very simple in terms of algorithm, and expected to obtain more applications.

It is worth mentioning that the experimental results first confirmed the feasibility of distinguishing β-thalassemia and normal control samples by NIR spectroscopy of human hemolytic solutions. Compared with the conventional methods, the NIR method is rapid and simple; it is a promising tool for β-thalassemia screening in large population prevention and control program.

The selected waveband provided valuable reference in designing small and dedicated spectrometer for the large-scale application of NIR method. The BiCC can be combined with other wavelength selection methods for more applications.

Acknowledgment

This work was supported by the Science and Technology Project of Guangdong Province of China (Nos. 2014A020213016 and 2014A020212445).

References

1. P. Williams, K. Norris, *Near-Infrared Technology in the Agricultural and Food Industries*, American Association of Cereal Chemists, USA (2001).
2. Z. Seregely, T. Deák, G. D. Bisztray, "Distinguishing melon genotypes using NIR spectroscopy," *Chemometr. Intell. Lab.* **72**, 195–203 (2004).
3. R. A. Viscarra Rossel, D. J. J. Walvoort, A. B. McBratney, L. J. Janik, J. O. Skjemstad, "Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties," *Geoderma* **131**, 59–75 (2006).
4. H. Z. Chen, T. Pan, J. M. Chen, Q. P. Lu, "Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods," *Chemometr. Intell. Lab.* **107**, 139–146 (2011).
5. T. Pan, M. M. Li, J. M. Chen, "Selection method of quasi-continuous wavelength combination with applications to the near-infrared spectroscopic analysis of soil organic matter," *Appl. Spectrosc.* **68**, 263–271 (2014).
6. T. Pan, Y. Han, J. M. Chen, L. J. Yao, J. Xie, "Optimal partner wavelength combination method with application tonear-infrared spectroscopic analysis," *Chemometr. Intell. Lab.* **156**, 217–223 (2016).
7. J. Y. Chen, H. Zhang, R. Matsunaga, "Rapid determination of the main organic acid composition of raw Japanese Apricot fruit juices using near-infrared spectroscopy," *J. Agric. Food Chem.* **54**, 9652–9657 (2006).
8. Z. Y. Liu, B. Liu, T. Pan, J. D. Yang, "Determination of amino acid nitrogen in tuber mustard using near-infrared spectroscopy with waveband selection stability," *Spectrochim. Acta A, Mol. Biomol. Spectrosc.* **102**, 269–274 (2013).
9. H. S. Guo, J. M. Chen, T. Pan, J. H. Wang, G. Cao, "Vis–NIR wavelength selection for nondestructive discriminant analysis of breed screening of transgenic sugarcane," *Anal. Methods* **6**, 8810–8816 (2014).
10. A. C. Sousa, M. M. L. M. Lucio, O. F. Bezerra Neto, G. P. S. Marcone, A. F. C. Pereira, E. O. Dantas, W. D. Fragoso, M. C. U. Araujo, R. K. H. Galvao, "A method for determination of COD in a domestic wastewater treatment plant by using near-infrared reflectance spectrometry of seston," *Anal. Chim. Acta* **588**, 231–236 (2007).
11. T. Pan, Z. H. Chen, J. M. Chen, Z. Y. Liu, "Near-Infrared spectroscopy with waveband selection stability for the determination of COD in sugar refinery wastewater," *Anal. Methods* **4**, 1046–1052 (2012).
12. J. H. Jiang, R. J. Berry, H. W. Siesler, Y. Ozaki, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared

and near-infrared spectroscopic data," *Anal. Chem.* **74**, 3555–3565 (2002).
13. J. Xie, T. Pan, J. M. Chen, H. Z. Chen, X. H. Ren, "Joint optimization of Savitzky–Golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose," *Chin. J. Anal. Chem.* **38**, 342–346 (2010).
14. T. Pan, J. M. Liu, J. M. Chen, G. P. Zhang, Y. Zhao, "Rapid determination of preliminary thalassaemia screening indicators based on near-infrared spectroscopy with wavelength selection stability," *Anal. Methods* **5**, 4355–4362 (2013).
15. Y. Han, J. M. Chen, T. Pan, G. S. Liu, "Determination of glycated hemoglobin using near-infrared spectroscopy," *Chemometr. Intell. Lab.* **145**, 84–92 (2015).
16. L. J. Yao, N. Lyu, J. M. Chen, T. Pan, J. Yu, "Joint analyses model for total cholesterol and triglyceride in human serum with near-infrared spectroscopy," *Spectrochim. Acta A, Mol. Biomol. Spectrosc.* **159**, 53–59 (2016).
17. M. Kim, Y. H. Lee, C. Han, "Real-time classification of petroleum products using near-infrared spectra," *Comput. Chem. Eng.* **24**, 513–517 (2000).
18. L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, S. Wold, *Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications*, Umetrics, Sweden (2006).
19. X. L. Long, G. S. Liu, T. Pan, J. M. Chen, "Waveband selection of reagent-free determination for thalassemia screening indicators using Fourier transform infrared spectroscopy with attenuated total reflection," *J. Biomed. Opt.* **19**, 1–11 (2014).
20. R. Galanello, A. Eleftheriou, J. Traeger-Synodinos, *Prevention of Thalassaemias and other Haemoglobin Disorders*, Team up Creations Ltd, Nicosia Cyprus (2003).
21. X. M. Xu, Y. Q. Zhou, G. X. Luo, C. Liao, "The prevalence and spectrum of α and β thalassaemia in Guangdong Province: Implications for the future health burden and population screening," *J. Clin. Pathol.* **57**, 517–522 (2004).
22. F. Xiong, M. Sun, X. Zhang, R. Cai, "Molecular epidemiological survey of haemoglobinopathies in the Guangxi Zhuang Autonomous Region of southern China," *Clin. Genet.* **78**, 139–148 (2010).
23. A. Savitzky, M. J. E. Golay, "Smoothing and differentiation of data by simplied least squares procedures," *Anal. Chem.* **36**, 1627–1639 (1964).
24. R. W. Kennard, L. A. Stone, "Computer-aided design of experiments," *Technometrics* **11**, 137–149 (1969).
25. D. D. Claeys, T. Verstraelen, E. Pauwels, C. V. Stevens, M. Waroquier, V. V. Speybroeck, "Conformational sampling of macrocyclic alkenes using a Kennard-Stone-based algorithm," *J. Phys. Chem. A* **114**, 6879–6887 (2010).
26. J. Workman, L. Weyer, *Practical Guide to Interpretive Near-infrared Spectroscopy*, CRC Press, USA (2008).
27. J. H. Wang, M. Q. Zhang, G. Cao, "Evaluation of the agronomic characteristics of transgenic sugarcane resistant to herbicide," *Guangdong Agric. Sci.* **9**, 23–24 (2011).
28. X. M. Li, M. X. Wang, T. H. Qin, C. Yang, Q. An, J. Zhang, "Genetic Transformation of Bt (cry1Ab) Gene into Sugarcane (*Saccharum officinarum L.*) Mediated by Agrobacterium tumefaciens," *Biotechnol. Bull.* **2**, 100–105 (2013).