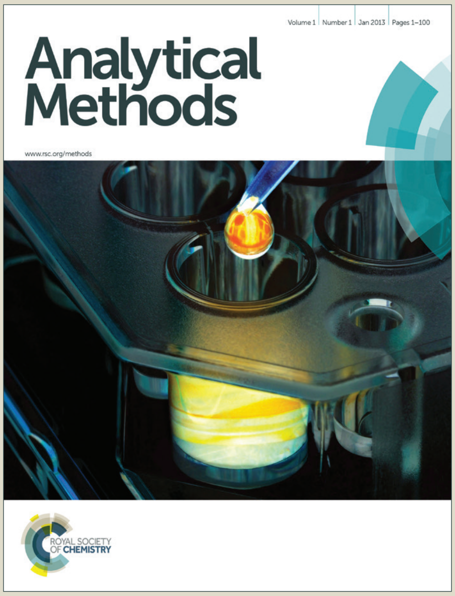

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analytical Methods
www.rsc.org/methods
ROYAL SOCIETY OF CHEMISTRY

Analytical Methods

Dynamic Article Links ►

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

PAPER

Vis−NIR Wavelength Selection for Non-Destructive Discriminant Analysis of Breed Screening of Transgenic Sugarcane

Haosong Guo,[1] Jiemei Chen,[1,]* Tao Pan,[1,]* Jihua Wang,[1,2] Gan Cao[1,2]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Savitzky−Golay (SG) method and moving-window waveband screening are applied to a coupling model of principal component (PCA) and linear discriminant analysis (LDA). An SG-pretreatment-based method (MW−PCA−LDA) for spectral pattern recognition is proposed, which is successfully employed for the non-destructive recognition of transgenic sugarcane leaves using visible (Vis) and near-infrared (NIR) diffuse reflectance spectroscopy. A Kennard-Stone-algorithm-based process of calibration, prediction and validation in consideration of uniformity and representative was performed to produce objective models. A total of 456 samples of sugarcane leaves in the elongating stage were collected from a planted field. These samples were composed of 306 transgenic samples containing both bacillus thuringiensis (Bt) and biolaphos resistance (Bar) genes, and 150 non-transgenic samples. According to the spectral recgonition effects, two parallel optimal SG modes were selected. The one of 1st order derivative, 3rd degree polynomial and 25 smoothing points was taken as an example to pretreat the diffuse reflectance spectra. Based on MW−PCA−LDA method, the optimal waveband was 768 nm to 822 nm, the optimal PC combination was PC_1−PC_3 and the corresponding validation recognition rates of transgenic and non-transgenic samples achieved 99.1% and 98.0%, respectively. The results show that Vis−NIR spectroscopy combined with SG pretreatment and MW−PCA−LDA method can be used for accurate recognition of transgenic sugarcane leaves and provides a quick and convenient means of screening transgenic sugarcane breeding for large-scale agricultural production.

Introduction

Sugarcane is the major sugar crop and cane sugar accounts for approximately 70% of total world sugar production. China's sugar production ranks third in the world. In addition to sugar refining, sugarcane is also used for the production of paper and fuel ethanol. Sugarcane is usually grown under conditions of high temperature and humidity and its productivity is challenged by a wide array of biotic and abiotic stresses, with insects being one of the major causes of economic losses. With the development of agricultural biotechnology, transgenic sugarcane breeding is increasingly receiving attention. This increases the yield, resistance and added value of sugarcane by transferring insect-resistant and herbicide-tolerant genes into the sugarcane[1]. The microorganism bacillus thuringiensis (Bt), a gram-positive, spore-forming soil bacterium, produces a crystalline parasporal body during sporulation, which shows biocidal activity against some invertebrate orders such as lepidopteran, dipteran, and coleopteran insects at larval stage, as well as against nematodes. And biolaphos resistance (Bar) gene, which was cloned from streptomyces hygroscopicus, can code phosphinothricin acetyltransferase to avoid the damage that L-phosphinothricin from herbcide could do to the plants[2]. In transgenic sugarcane breeding, it is necessary to determine whether the exogenous gene is successfully expressed in a sugarcane plant. Molecular biology detection technologies, such as polymerase chain reaction and enzyme-linked immunosorbent assay (ELISA), are mainly used for genetic screening[3]. These methods are complicated, require expertise and cannot meet the needs of large-scale production. It is therefore of applied value to develop a simple and rapid method of transgenic sugarcane breeding screening.

Near-infrared (NIR) electromagnetic radiation has frequencies between the visible (Vis) and mid-infrared ranges. It primarily reflects absorption of overtones and combinations of vibrations of X−H functional groups (such as C−H, O−H and N−H). Because NIR absorption strength is weak, most sample types can be measured directly without preprocessing. This rapid, simple and non-destructive technique therefore has obvious advantages and is commonly used in many areas, including agriculture[4–6], food[7], environment[8] and biological medicine[9–12]. Furthermore, NIR spectra capture absorption information from the protein molecules related to genetic variations and this is applied in the fields of genetic disorders[13] and genetically modified crops[14, 15].

*[1]Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Educational Institutes, Jinan University, Huangpu Road West 601, Tianhe District, Guangzhou 510632, China. [2]Research Institute of Crops, Guangdong Academy of Agricultural Sciences, Jinying Road 28, Tianhe District, Guangzhou 510640, China. *E-mail: tpan@jnu.edu.cn (T. Pan) and tchjm@jnu.edu.cn (J. M. Chen), Tel: +86-20-85224379.*

In parallel with NIR spectroscopy and chemometrics developments, non-destructive NIR spectroscopy shows substantial potential as a pattern recognition tool for transgenic sugarcane breeding screening. Differences in protein molecule structures between transgenic and non-transgenic sugarcane contain large numbers of X−H functional groups that have significant NIR absorption. However, the application of NIR spectroscopy to transgenic sugarcane breeding screening has not previously been proposed.

Linear discriminant analysis (LDA) is an effective method of pattern recognition. To overcome high-dimension and co-linearity (singularity) problems of NIR spectral data, principal component analysis (PCA) is commonly used to reduce dimensionality and eliminate data singularity. The processed data are then input to an LDA model for spectral pattern recognition. The PCA−LDA method is successfully used in many applications of spectral discriminant analysis[16, 17].

Non-destructive spectroscopic detection of crops has obvious application advantages, but also offers a challenge to calculation methods. Because crops, such as sugarcane leaves, are complex systems with multiple components, their absorption spectra can be disturbed by various types of noise. To improve the horizontal and vertical quality of spectral data, spectral preprocessing and wavelength selection are necessary to eliminate noise and to extract information, respectively.

Savitzky−Golay (SG) method[18] is an effective spectral preprocessing method with a wide scope of application and a variety of different SG modes[7, 8, 11, 19–21]. The moving-window partial least squares (MW−PLS) method[9] has proven effective for waveband selection in spectroscopic quantitative analysis [4, 5, 7, 8, 10, 13]. SG method combined with moving-window waveband screening is applied to PCA−LDA models to improve pattern recognition by simultaneously optimising both smoothing modes and wavebands. In the present study, an MW−PCA−LDA algorithm platform is established and applied to non-destructive Vis−NIR spectroscopic recognition of transgenic sugarcane leaves.

Materials and methods

Experimental materials, instruments and measurement methods

Transgenic sugarcane materials: Transgenic sugarcane strains contained both Bt and Bar genes genetically modified from three types of sugarcane receptors, ROC 20th, ROC 22th and Yuetang No.00-236, giving a total of 306 samples. *Non-transgenic sugarcane materials*: Non-transgenic sugarcane strains of seven types, ROC 1st, ROC 2nd, ROC 3rd, ROC 4th, ROC 20th, ROC 22th and Yuetang NO.00-236, gave a total of 150 samples. ELISA was used to check the integrity of copies of the genes introduced during the breeding phase and the expression of the exogenous gene was guaranteed. The equipment used was ELISA kit BT-Cry1Ab/1Ac (AGDIA, Inc., USA) and microplate reader iMark (Bio-rad, Inc., USA).

All sugarcane plants were field-grown to the elongation stage. A total of 456 samples of sugarcane leaves were collected, comprising the 306 transgenic samples (positive) with Bt and Bar genes and 150 non-transgenic samples (negative), and at least one leaf was collected from each plant. The collected samples were cleaned and stored at room temperature for 2 h to equilibrate to the experimental environment before collection of the Vis−NIR diffuse reflectance spectra.

Spectra were collected using an XDS Rapid Content^{TM} grating spectrometer (FOSS, Denmark) equipped with a diffuse reflection accessory and a round sample cell. The scanning range spanned 400 nm to 2498 nm with a 2 nm wavelength gap, which includes the entire NIR region and a large part of the visible region. Wavebands of 400 nm to 1100 nm and 1100 nm to 2498 nm were selected for Si and PbS detection, respectively. The samples were placed directly in the diffuse reflection accessory. Each sample was measured in triplicate, and the mean value of three measurements was used for modeling. The spectra were measured at 25 ±1 ℃ and 46% ±1% relative humidity.

Sample division and calibration, prediction and validation process

A framework of calibration, prediction and validation based on uniformity and representative was developped to produce objective models. Some samples were randomly selected from all samples as validation samples and were not subjected to the modeling optimisation process. The remaining samples were used as modeling samples and were divided into calibration and prediction sets using Kennard-Stone (K−S) algorithm[22]. The MW−PCA−LDA models were established for calibration and prediction sets, and model parameters were optimised depending on the recognition rate of prediction set. Finally, the selected optimal models were revalidated against the validation samples.

K−S algorithm is a well-performed method for sample division in experiment planning[22, 23]. The goal of the K-S algorithm is to select a maximally diverse subset from a large set of candidate samples, so the subset can represent the whole sample space uniformly and sufficiently. The algorithm assumes that one can define a ‘distance’ between two samples, which is low when the two samples are similar and high otherwise.

In this study, the calibration samples were selected from all of the modeling samples using K−S algorithm while the remainings were used as prediction samples. For the spectral data $A^{(i)}$ (absorbance vector) of n samples with K variables (wavelength),

$$A^{(i)} = (A_1^{(i)}, A_2^{(i)}, \cdots, A_K^{(i)}), i = 1, 2, \cdots, n \cdot \qquad (1)$$

The Euclidean distance is employed to be the distance measure between any two samples *i*, *j*, which is difined as the follows:

$$D_{i,j} = \sqrt{\sum_{\mu=1}^{K} (A_\mu^{(i)} - A_\mu^{(j)})^2} \cdot \qquad (2)$$

Given this distance measure, the K−S algorithm works as follows. One starts with the modeling set being the candidate set. The first two selected samples are the two candidates with the maximal pair distance. All subsequent samples are selected in an iterative way until the number of selected samples reaches the desired number of calibration samples. In one iteration, every candidates' minimum distances to all of the previously selected are calculated, the one with the largest minimum distance is selected.

To ensure modeling representativeness and integrity, all calibration, prediction and validation sets must contain negative and positive samples, so the negative and positive samples should be divided into these three sets. The specific procedure was as follows. Firstly, for 150 negative samples, 50 samples were randomly selected for validation. The remaining 100 samples were used as modeling samples. Fifty samples were further selected from modeling samples using K−S algorithm as calibration samples while the remainings as predicion samples. Secondly, for 306 positive samples, 106 samples were randomly selected for validation. The remaining 200 samples were used as modeling samples. One hundred samples were further selected from modeling samples using K−S algorithm as calibration samples while the remainings as predicion samples. Finally, the positive and negative samples used for validation were merged into one validation set (156 samples). Similarly, the positive and negative samples used for calibration and prediction were merged into calibration (150 samples) and prediction (150 samples) sets, respectively. Fig. 1 shows the type and number of samples in the calibration, prediction and validation sets.

SG method

SG method's parameters include order of derivatives *d*, degree of polynomial *p* and number of smoothing points *m* (odd)[19]. Any *m* consecutive spectral data are considered as a smoothing window and the data in the window were fitted using a polynomial function. The values of *d*th derivative at the centre of the window can be then calculated and expressed as a linear combination of all data within the window, in which the coefficients are uniquely determined and called SG smoothing coefficients. The *d*th derivative spectra were obtained by moving the window across the entire spectral range. Each combination of parameters corresponds to an SG mode and a set of SG smoothing coefficients. In the original paper of SG method[19], the paremeters *d*, *p*, *m* were set to be: *d* = 0, 1, 2, 3, 4, 5; *p* = 2, 3, 4, 5, 6; $m = 5, 7, \cdots, 25$, which corresponded to a total of 117 SG modes.

Considering that the absolute values of the fourth and fifth derivatives were very small (which meant a large amount of spectral information was missing), the SG modes using these derivatives were not used for screening in this study. Furthermore, if both the wavelength gap and number of smoothing points were small, then the smoothing window was narrow and the information in the window for smoothing was insufficient, and it was difficult to get satisfactory preprocessing effects. Hence, it was necessary to expand the number of smoothing points *m*. In this paper, *m* was expanded to $5, 7, \cdots, 51$ (odd). The corresponding polynomial coefficients of the new modes were determined based on the original method[18]. A total of 264 SG modes were obtained.

Taking the SG mode with 1st order derivative, 3rd degree polynomial and 25 smoothing points as an example, namely (*d*, *p*, *m*)= (1, 3, 25), the calculation process of SG smoothing coefficient is presented as follows. In fact, 3rd degree polynomial and 25 smoothing points are employed to calculate the 1st derivative spectra. Firstly, serial number *i* of 25 continuous wavelengths in the window is $i = 0, \pm 1, \pm 2, \cdots, \pm 12$, the corresponding absorbance value is A_i, 3rd degree polynomial is defined by the following equation:

$$f_i = \sum_{k=0}^{3} b_{3k} i^k = b_{30} + b_{31} i + b_{32} i^2 + b_{33} i^3 \cdot \qquad (3)$$

Polynomial coefficients b_{3k}, *k*=0, 1, 2, 3, are fitted using A_i. Then the value of the 1st order derivative of the centre point (*i*=0) is calculated as the follows:

$$\left(\frac{df_i}{di} \right) = b_{31} \cdot \qquad (4)$$

Therefore, only b_{31} needs to be determined. According to least square method, there are the equations:

$$\frac{\partial}{\partial b_{3r}} \left[\sum_{i=-12}^{12} \left(\sum_{k=0}^{3} b_{3k} i^k - A_i \right)^2 \right] = 0, r = 0, 1, 2, 3 \text{,} \qquad (5)$$

which can be reduced to:

$$\sum_{k=0}^{3} b_{5k} \left(\sum_{i=-12}^{12} i^{r+k} \right) = \sum_{i=-12}^{12} i^r A_i, r = 0, 1, 2, 3 \cdot \qquad (6)$$

This a system of constant coefficient linear equations about b_{3k} (*k*= 0, 1, 2, 3), whose coefficient determinant is not zero. So b_{31} can be uniquely determined and represented as the linear combination of A_i as the follows:

$$b_{31} = \sum_{i=-12}^{12} \lambda_i A_i \text{,} \qquad (7)$$

in which, λ_i is called SG smoothing coefficient. The 25 obtained smoothing coefficients are presented as the follows: 1.7379, 0.4843, −0.4800, −1.1814, −1.6461, −1.9005, −1.9708, −1.8834, −1.6645, −1.3404, −0.9374, −0.4819, 0, 0.4819, 0.9374, 1.3404, 1.6645, 1.8834, 1.9708, 1.9005, 1.6461, 1.1814, 0.4800, −0.4843, −1.7379.

Other SG modes were calculated in the similar way. In this study, the specific computer algorithms platform was built up by using MATLAB v7.6, the combination of smoothing coefficients for every SG mode was obtained. All of the 264 SG modes were employed in modeling, and the most appropriate SG mode was screened based on modeling effects.

Waveband screening with moving window

Consecutive spectral data for *N* adjacent wavelengths were designated as a window. By moving and varying the size of the window in a predetermined search region of the spectrum, each analytical waveband corresponding to a window was determined for modeling. By considering the position and length of the wavebands, the search parameters were set as follows: (1) initial wavelength (*I*) and (2) number of wavelengths (*N*).

The search range covered the entire scan region from 400 nm to 2498 nm using 1050 wavelengths. Further, *I* was set to $I \in \{400, 402, \cdots, 1050\}$. To reduce workload and ensure representativeness, *N* was set to

$$N \in \{1, 2, \cdots, 50\} \cup \{60, 70, \cdots, 200\} \cup \{220, 240, \cdots, 860\} \cup \{1050\}$$

PCA-LDA method

The PCA−LDA calibration and prediction modeling process was as follows: (1) Principal component analysis was performed based on a matrix of the absorbance spectra of the calibration set, and the loading and score matrices were obtained. (2) By projecting the high-dimensional spectral data to a two-dimensional principal component plane and plotting the results, it is possible to visualise the structure of the investigated data set. Specifically, based on spectra of the calibration set, the first, second and third components (PC_1, PC_2 and PC_3) with maximum contribution values were derived and normalised. Any two of these together defined a model plane, namely PC_1−PC_2, PC_1−PC_3

and PC_2−PC_3 planes. (3) Linear discriminant analysis was performed on each principal component plane based on spectra of calibration set. A straight cut-off line was determined which optimally classified transgenic and non-transgenic sugarcane leaves samples. (4) On the basis of absorbance matrices of the prediction set and the acquired loading matrices of the calibration set, the score matrix of the prediction set was calculated. The first, second and third components of prediction samples were derived and normalised. According to the cut-off line, the genotypes of the prediction samples were further recognised. (5) Referring to genuine genotypes of the prediction samples, the prediction recognition rate was calculated, and the optimal combination of principal components was determined according to the recognition rate.

Optimization frame of MW-PCA-LDA and evaluation indices

Step 1 Screening of SG modes based on PCA−LDA models for the entire scan region: (1) the derivative spectra corresponding to each SG mode were calculated for the entire scan region, 400 nm to 2498 nm. (2) PCA−LDA was established on the basis of each SG mode. The recognition rate (REC) among prediction samples were calculated and denoted as *P*_REC, which is defined by the following equation:

$$P\_\mathrm{REC} = \frac{\tilde{N}_{\mathrm{Pre}}}{N_{\mathrm{Pre}}} \times 100\ \%, \tag{8}$$

Where N_{Pre} is the number of prediction samples and $\tilde{N}_{\mathrm{Pre}}$ is the number of correctly recognised prediction samples. The optimal SG mode's parameters (*d*, *p* and *m*) and the combination of principal components for the PCA−LDA model were selected to maximise *P*_REC.

Step2 Screening of wavebands with moving-window mode: Using derivative spectra by the optimal SG mode selected in *Step 1*, MW−PCA−LDA models were established for all combinations of parameters *I*, *N* and combinations of principal components, and the corresponding *P*_RECs were calculated. The optimal waveband was sequentially selected according to the maximum *P*_REC.

Step3 Model validation: The optimal MW−PCA−LDA model was validated using the validation samples that were excluded from the modeling optimisation process. The validation recognition rate of transgenic and non-transgenic samples were calculated and denoted as V_REC^{+} and V_REC^{-}, respectively, which are defined by the following equations:

$$V\_\mathrm{REC}^{+} = \frac{\tilde{N}_{\mathrm{Val}}^{+}}{N_{\mathrm{Val}}^{+}} \times 100\ \%, \tag{9}$$

$$V\_\mathrm{REC}^{-} = \frac{\tilde{N}_{\mathrm{Val}}^{-}}{N_{\mathrm{Val}}^{-}} \times 100\ \%, \tag{10}$$

where N_{Val}^{+} and N_{Val}^{-} are the numbers of transgenic and non-transgenic samples in the validation set, respectively, and $\tilde{N}_{\mathrm{Val}}^{+}$ and $\tilde{N}_{\mathrm{Val}}^{-}$ are the numbers of correctly recognised transgenic and non-transgenic samples in the validation set, respectively.

The computer algorithms for the abovementioned method were designed using the MATLAB version 7.6 software.

Results and discussion

PCA-LDA based on the entire scanning region without SG pretreatment

The Vis−NIR diffuse reflection spectra of 456 samples of sugarcane leaves for the entire scan region (400 nm to 2498 nm), covering the entire NIR region and part of visible region, are shown in Fig. 2. And Fig. 3 shows the average spectra of the transgenic and non-transgenic sugarcane leaves samples. Obvious differences are found from a visual observation of the two spectra, especially around three spectral peaks at 678 nm, 1450 nm and 1928 nm and four spectral valleys at 552 nm, 800 nm, 1666 nm and 2216 nm. The peak at 678 nm represents pigment absorption where a gap is indicated between these lines. In Fig. 3, it is interestingly revealed that, around the valley at 800 nm, the absorption of transgenic sugarcane leaves is remarkably lower than that of non-transgenic sugarcane leaves, however, at those three peaks and other three valleys, the absorption of transgenic sugarcane leaves is remarkably higher than that of non-transgenic sugarcane leaves. Even though chlorophyll content might partially explain the observed spectral difference between types of leaf, the difference might arise from other facts.

For comparison, PCA−LDA models without pretreatment of SG method were first established using the entire scan region. The *P*_REC is 81.3%. The result shows that spectral recognition without pretreatment of SG method is unsatisfactory and it is necessary to pretreat the spectral data.

Screening of SG modes

A total of 264 different SG modes were used to preprocess each spectrum to establish PCA−LDA models based on the entire scan region. The modeling effects (*P*_REC) of local optimal models corresponding to each *m* (number of smoothing points) are plotted in Fig. 4 distinguished by different *d* (orders of derivative). Furthermore, the *P*_REC and parameters of local optimal models corresponding to each *d* are summarised in Table 1. For the case of 1st order of derivative (*d*=1), there are two parallel local optimal models. The two models are just the global optimal models. For the cases of 2nd , 3rd order derivative and no derivative (*d*=0, 2, 3), there are also multiple models tied for local optimal models with poor modeling effects, so only the models with the least *m* are listed in Table 1. As shown in Table 1 and Fig. 4, the maximum *P*_REC was 93.3%. The spectral recognition is obviously better than that without SG pretreatment. Derivatives are an alternative, among others, to correct baseline deviations (drifts) caused by the multiplicative light scattering observed in diffuse reflectance measurements. SG method uses a smoothing of the spectra prior to calculating the derivative in order to decrease the detrimental effect on the signal-to-noise ratio that conventional finite-difference derivatives would have.[21] These results indicate that SG method can reduce spectrum noise and enhance spectral recognition ability, which can be attributed more to the use of derivative than to the smoothing. There are also differences in spectral recognition for different SG modes, so screening of SG modes is necessary.

Table 1 Comparison of PCA−LDA modeling effects with and without SG pretreatment based on the entire scan region (400 nm to 2498 nm)

d	*p*	*m*	PCC	*P*_REC
No SG pretreatment	--	--	1-2	81.3%

0	2, 3	5	1-2	81.3%
1	3, 4	25	1-3	**93.3%**
	5, 6	43	1-3	
2	2, 3	31	1-3	92.7%
3	3, 4	45	1-3	92.0%

Note: *d*: order of derivatives; *p*: degree of polynomial; *m*: number of smoothing points; PCC: principal component combination.

Screening of wavebands with MW-PCA-LDA

The entire scan region, 400 nm to 2500 nm, contains numerous wavelengths (i.e. N = 1050), which leads to high modeling complexity. To extract further information, reduce model complexity and improve spectral recognition, waveband optimisation was performed with MW−PCA−LDA, which is illustrated by the example of the optimal SG mode with d = 1, p = 3 and m = 25. The SG derivative spectra of 456 samples for this mode are shown in Fig. 5.

Using the methods mentioned in *Step 2*, the optimal values of I and N are 768 and 28, respectively. The corresponding waveband was 768 nm to 822 nm, which covers part of the Vis−NIR combined region. The optimal PC combination is PC_1−PC_3. The corresponding P_REC is 98.0%. The result is obviously better than that obtained from the full scan region. In addition, only a small number of wavelengths (i.e. N = 28) are adopted in the selected model and model complexity is significantly reduced. The position of the selected waveband is also shown in Fig. 3 and Fig. 5, which is just located in the interesting spectral valley at 800 nm . This region is related to fourth overtones of C−H (CH and CH_2) and third overtones of O-H (H_2O and Ar−OH).[24]

For fixed I and changing N, the local optimal model corresponding to a single parameter I (initial wavelength) was selected according to maximum P_REC; P_REC values for all I are shown in Fig. 6(a). For fixed N and changing I, the local optimal model corresponding to a single parameter N (number of wavelengths) was selected according to maximum P_REC; P_REC values for all N are shown in Fig. 6(b).

Fig. 6(a) and 6(b) show the maximum P_REC achieved when I = 768 nm and N = 28. These data may serve as a valuable reference for designing splitting systems for spectroscopic instruments. Local optimal models with prediction parameters close to those of the global optimal model remain good choices. These models address restrictions such as cost and material properties as well as the position and number of wavelengths in instrument design.

The results show that SG method combined with moving-window waveband screening can be applied to PCA−LDA models to well improve spectral pattern recognition. In fact, there are also other well-performed methods for wavelength selection, such as competitive adaptive reweighted sampling (CARS)[25], Monte Carlo uninformative variable elimination (MC−UVE)[26], randomization test (RT)[27, 28] and so on. These methods could also be combined with the algorithm platform proposed in this paper. Considering the limitation of article length, further discussion was omitted.

Model validation

The randomly selected validation samples, which were excluded in the modeling optimisation process, were used to validate the optimal MW−PCA−LDA model (d = 1, p = 3, m = 25, I = 766, N = 28 and PC_1−PC_3). The validation process was based on the methods mentioned above. From a total of 156 validation samples, 154 were correctly recognised while one of the positives and one of the negatives were wrongly recognised. The validation recognition rates V_REC$^+$ and V_REC$^-$ achieved 99.1% and 98.0%, respectively. As shown in Fig. 7, the validation samples plot on the principal component plane PC_1−PC_3 and are clearly classified into two groups.

The genotypes of samples that recognised by Vis-NIR spectral analysis are highly consistent with those of genetic diagnosis, indicating that spectral prediction is highly accurate for determining negative and positive samples in transgenic sugarcane breeding screening. The proposed wavelength selection may also provide valuable references for designing specialised spectrometers.

Conclusions

SG pretreatment method combined with moving-window waveband screening is applied to a PCA−LDA model, and a SG-pretreatment-based method (MW−PCA−LDA) for spectral pattern recognition is proposed, which is successfully used for the non-destructive recognition of transgenic sugarcane leaves using Vis-NIR spectroscopy. In addition, a Kennard-Stone-algorithm-based process of calibration, prediction and validation in consideration of uniformity and representative was performed to produce objective models.

Vis-NIR spectral prediction is highly accurate for determining genotypes of sugarcane leaves samples. The proposed wavelength selection may also provide valuable references for designing specialised spectrometers.

Comparing with conventional methods, the proposed method is rapid and simple. This technique is a potential and promising tool for transgenic sugarcane breeding screening of large-scale agricultural production.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61078040), the Science and Technology Project of Guangdong Province of China (No. 2012B031800917), and the Science and Technology Project of Guangzhou of China (No. 2014Y2-00002).

References

1 X. Y. Zhang, B. P. Yang and S. Z. Zhang. *Molecular Plant Breeding*, 2007, **5**, 155–159.

2 A. Arencibia, R. I. Vazquez, D. Prieto, P. Tellez, E. R. Carmona, A. Coego, L. Hernandez, G. A. Riva and G. Selman-Housein, *Mol.Breeding*, 1997, **3**, 247–255.

3 G. Y. Wang, W. X. Fan, B. H. Chen, J. W. Zhang and S. D. Han, *Food Science*, 2008, **29**, 698–705.

4 H. Z. Chen, T. Pan, J. M. Chen and Q. P. Lu, *Chemometr. Intell. Lab.*, 2011, **107**, 139–146.

5 T. Pan, M. M. Li and J. M. Chen. *Appl. Spectrosc.*, 2014, **68**, 263–271.

6 T. Pan, Z. T. Wu and H. Z. Chen, *Chinese J. Anal. Chem.*, 2012, **40**, 920–924.

7 Z. Y. Liu, B. Liu, T. Pan and J. D. Yang, *Spectrochim. Acta A*, 2013, **102**, 269–274.

8 T. Pan, Z. H. Chen, J. M. Chen and Z. Y. Liu, *Anal. Methods*, 2012, **4**, 1046–1052.

9 J. H. Jiang, R. J. Berry, H. W. Siesler and Y. Ozaki, *Anal. Chem.*, 2002, **74**, 3555–3565.

10 Y. P. Du, Y. Z. Liang, J. H. Jiang, R. J. Berry and Y. Ozaki, *Anal. Chim. Acta*, 2004, **501**, 183–197.

11 J. Xie, T. Pan, J. M. Chen, H. Z. Chen and X. H. Ren, *Chinese J. Anal. Chem.*, 2010, **38**, 342–346.

12 Z. Y. Liu and T. Pan, *Optics Preci. Eng.*, 2012, **20**, 2171–2175.

13 T. Pan, J. M. Liu, J. M. Chen, G. P. Zhang and Y. Zhao, *Anal. Methods*, 2013, **5**, 4355–4362.

14 L. J. Xie, Y. B. Ying, T. J. Ying, H. Y. Yu and X. P. Fu, *Anal. Chim. Acta*, 2007, **584**, 379–384.

15 L. J. Xie, Y. B. Ying and T. J. Ying, *J. Agric. Food Chem.*, 2007, **55**, 4645–4650.

16 L. Errikson, E. Johansson, W. N. Kettaneh, J. Trygg, C. Wikstrom and S. Wold, *Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications*, 2nd Ed, Umetrics Academy, Umea Sweden, 2006.

17 L. P. Hu, L. Zhang, Y. Li, L. M. Zhang and J. D. Wang, *Chinese J. Anal. Chem.*, 2007, **35**, 345–349.

18 A. Savitzky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1627–1637.

19 T. Pan, A. Hashimoto, M. Kanou, K. Nakanishi and T. Kameoka, *Bioprocess Biosyst. Eng.*, 2003, **26**, 133–139.

20 T. Pan, A. Hashimoto, M. Kanou, K. Nakanishi and T. Kameoka, *Jpn. J. Food Eng.*, 2004, **5**, 22–31.

21 A. Rinnan, F.v.d. Berg and S. B. Engelsen, *Trac-trend Anal. Chem.*, 2009, **28**, 1201–1222.

22 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.

23 D. D. Claeys, T. Verstraelen, E. Pauwels, C. V. Stevens, M. Waroquier and V. V. Speybroeck, *J. Phys. Chem. A*, 2010, **114**, 6879–6887.

24 J. Workman Jr. and L. Weyer, *Practical Guide to Interpretive Near-Infrared Spectroscopy*, CRC Press, Boca Raton, USA, 2008.

25 H. D. Li, Y. Z. Liang, Q. S. Xu and D. S. Cao, *Ana. Chim. Acta.*, 2009, **648**, 77–84.

26 W. S. Cai, Y. K. Li and X. G. Shao, *Chemometr. Intell. Lab.*, 2008, **90**, 188-194.

27 E. S. Edgington and P. Onghena, *Randomization Tests*, 4th Ed, CRC Press, Boca Raton, USA, 2007.

28 S. Wiklund, D. Nilsson, L. Eriksson, M. Sjostrom, S. Wold and K. Faber, *J. Chemometrics*, 2007, **21**, 427–439.

Figure legend

Figure. 1 Type and number of samples in the calibration, prediction and validation sets.

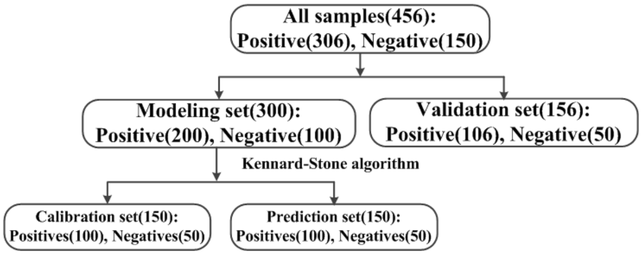

All samples(456):
Positive(306), Negative(150)
Modeling set(300):
Positive(200), Negative(100)
Validation set(156):
Positive(106), Negative(50)
Kennard-Stone algorithm
Calibration set(150):
Positives(100), Negatives(50)
Prediction set(150):
Positives(100), Negatives(50)

Figure. 2 Vis—NIR diffuse reflection spectra of 456 samples of sugarcane leaves.

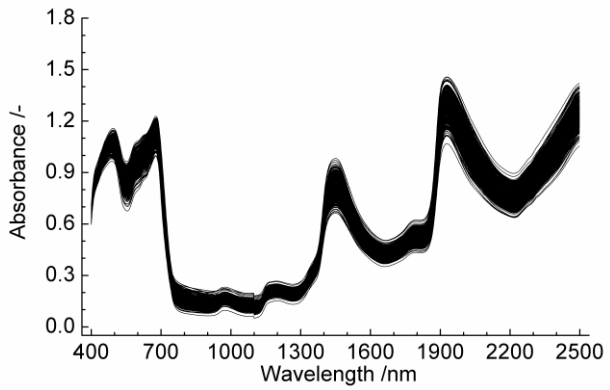


Figure. 3 Average spectra of the transgenic and non-transgenic sugarcane leaf samples.

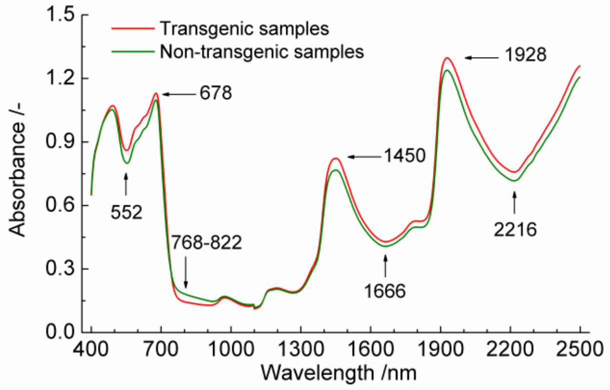


Figure. 4 *P*_REC corresponding to each number of smoothing points distinguished by different orders of derivative.

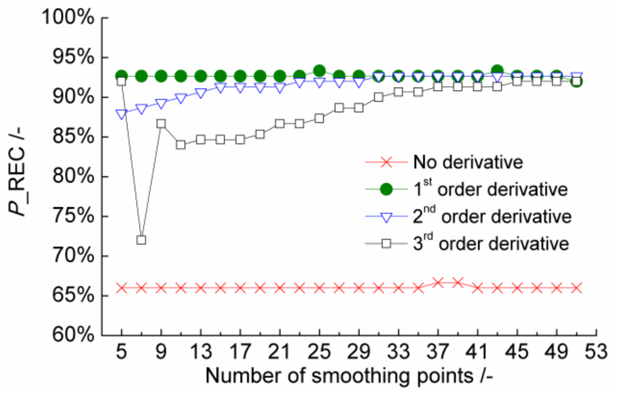


Figure. 5 SG derivative spectra of 456 samples with first-order derivative, third degree polynomial and twenty-five smoothing points.

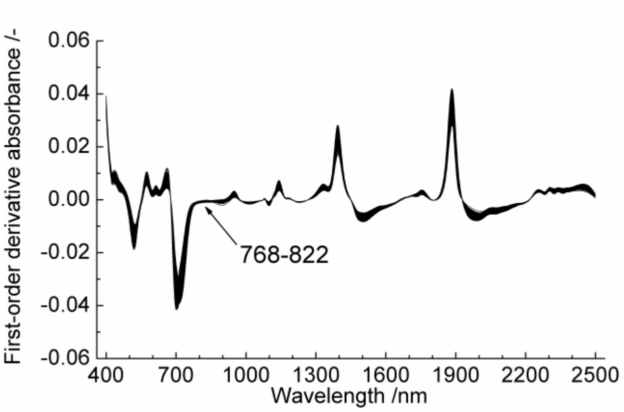


Figure. 6 *P*_REC corresponding to (a) initial wavelength *I* and (b) number of wavelengths *N*.

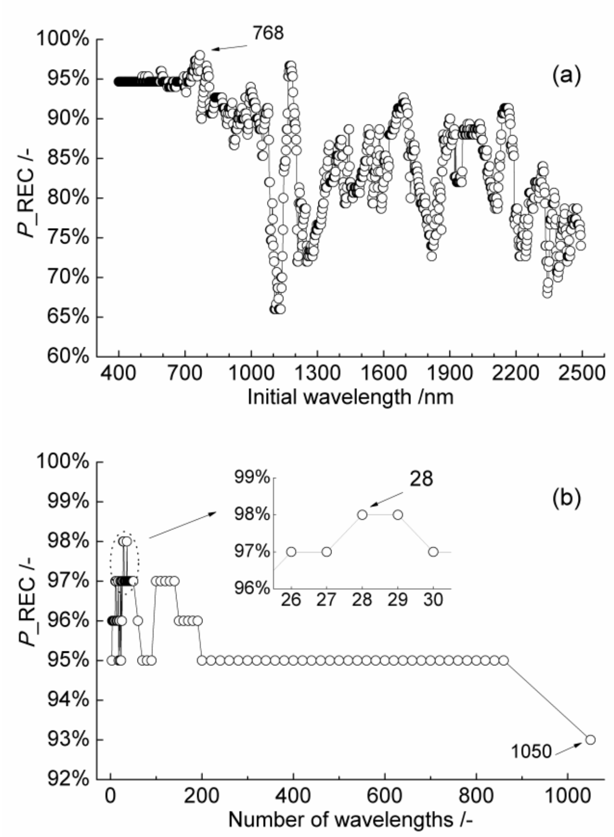
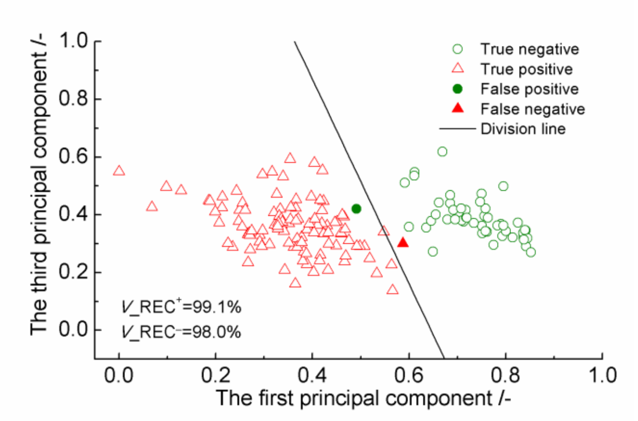


Figure. 7 Validation recognition of the optimal model with first-order derivative, third degree polynomial and twenty-five smoothing points on principal component plane PC_1−PC_3 based on the waveband from 768 nm to 822 nm.


The third principal component /-
○ True negative
△ True positive
● False positive
▲ False negative
Division line
V_REC⁺=99.1%
V_REC⁻=98.0%
The first principal component /-