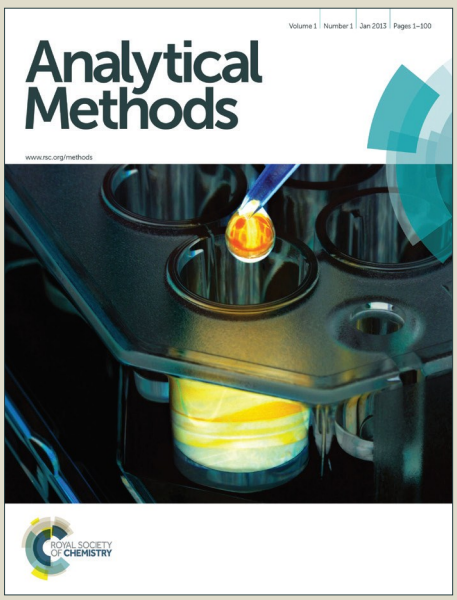
Analytical Methods

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Analytical Methods

RSCPublishing

ARTICLE

Optimal wavelengths selection for visible diffuse reflectance spectroscopy discriminating human and nonhuman blood species

Cite this: DOI: 10.1039/c4ay01673d

Received 16th July 2014,
Accepted 28th August 2014

DOI: 10.1039/c4ay01673d

www.rsc.org/methods

Linna Zhang [a,b], Meixiu Sun [c], Zhennan Wang [c], Hongxiao Li [c], Yingxin Li [c], Zhigang Fu[d], Yang Guan[d], Gang Li [a,b], Ling Lin [a *]

The species identification of human and nonhuman blood is an important and immediate challenge for forensic science, veterinary purposes, and wildlife preservation. Current methods used to identify the species of origin of a blood stain are limited in scope and destructive to the sample. We have previously demonstrated that visible diffuse reflectance spectroscopy combining PLS-DA method can realize the discrimination of human and nonhuman blood. Researches have proved the application of a proper wavelength variable selection prior to the model calibration can be greatly beneficial in providing more reliable and parsimonious model. Apart from improving the prediction ability, the using of variable selection will also reduce the experimental work. Moreover, the cost of a high-performance optical emission spectrometer and a supercontinuum white light laser source is comparatively high. In contrast, diode laser, fixed-filter spectrometer and diode arrays spectrometer, which are very common products, greatly cut the cost of measurement system. The key to use this kind of spectrometer is to find optimal wavelength combination for getting a fine calibration. In this paper, we used Equidistant Combination Multiple Linear Regression (ECMLR) method for wavelength selection. Compared with the results of full-spectrum PLS-DA, ECMLR method could enhance the performance of identified models. Delightedly, for time related validation, the prediction effect of ECMLR method was slightly better than full-spectrum PLS-DA method. The overall results sufficiently demonstrate the PLS-DA model constructed using selected wavelength variables by a proper wavelength variable method can be more effective and accurate.

1. Introduction

Interspecies blood analysis is an important part in analytical chemistry, forensic science, and biochemistry. Identifying the species of a blood stain is of great significance to the fields of forensic casework [1, 2], veterinary science, and wildlife preservation [3]. Currently, species identification can be realized by many analytical techniques. For a forensic case, it is very important to certify a blood sample to be human origin. Some commercially available kits contain antihuman haemoglobin (Hb) antibodies, which recognize and bind to human Hb [4]. These tests will show a positive result for human blood and a negative result for any other animal blood. There are also a few tests that can specifically identify the species of a blood sample. The double diffusion assays [5] and the Ouchterlony method can do the identification. High-performance liquid chromatography (HPLC) [6] and mass spectrometry (MS) [7] are emerging analytical techniques for discriminating blood species. However, most of these methods are destructive to the sample. Raman Spectroscopy method was also demonstrated to be efficient for identification of blood species [8-10]. Compared to these methods, our group has proved that visible diffuse reflectance spectroscopy combined with PLS-DA method can successfully discriminate human blood and nonhuman blood, which was non-destructive and fast [11].

The use of high-performance optical emission spectrometer and supercontinuum white light laser source makes the measurement system high-budget. Comparatively, diode laser, and diode arrays spectrometer and so on are more economical products. The key to realize a practical measurement system is seeking out an optimal wavelength combination. Moreover, model calibration (especially linear models) with the full-spectrum information is time-consuming. Furthermore, redundancy and collinearity are two widespread phenomena among the spectral data matrix, which would affect the accuracy and robustness of the model. The application of a proper wavelength variable selection prior to the model

calibration has been proved to be greatly beneficial for a more reliable and parsimonious model [12]. During the last few decades, many methods of wavelength variable selection have been developed.

At present, spectral wavelength selection method can be classified into two categories: the continuous mode and the discrete mode. Continuous mode selects adjacent wavelengths, whose typical case was moving window partial least squares (MWPLS) [13, 14]. However, spectral co-linearity may appear in adjacent wavelengths, which may result in evaluation distortion. Discrete mode usually selects the non-adjacent wavelengths, which was designed to minimize co-linearity problems. Multiple linear regressions (MLR) can be directly employed to discrete wavelength combination [15]. Equidistant combination multiple linear regressions (ECMLR), are equidistant wavelength selection method with quasi-continuous mode [16]. Equidistant combination multiple linear regressions improve the conditioning of MLR by minimizing co-linearity effects when the data gap selection is appropriate. An appropriate quasi-continuous wavelength combination can be conveniently screened from a continuous waveband along with a large amount of wavelengths.

Based on the achievements of ECMLR method and MWPLS method, in this paper, we tried to use both methods to do the wavelength selection, reducing the experimental work and contributing to a more reliable calibration model.

2. Materials and methods

2.1 Experimental Materials, Instruments

A supercontinuum white light laser source (Superke Compact, NKT, Denmark), a visible spectrometer (QE65000, Ocean Optics, USA) and a fiber probe were used to get the diffuse reflectance spectra. The spectrometer has a spectral range of 350–1150 nm, 1044 wavelengths in total. The details had been introduced in reference [11].

Sixty blood samples from macaque, rat, chicken, pig and guinea pigs were formally delivered by Institute of Zoology, Chinese Academy of Sciences. Fifteen human blood samples were formally delivered by Tianjin People's Hospital, which was the same with the blood samples in reference [11]. All experiments performed were in compliance with relevant laws, as well as with the guidelines of Institute of Zoology, Chinese Academy of Sciences, Tianjin People's Hospital and State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University. All the above mentioned institutes have approved the experiments. The volunteers had given their consent for the experiments. The blood samples were measured within 48 hours (after 24 hours) after the samples acquired. Before the measurement, the samples were prepared with ice. Each sample was prepared by placing about 1 mL on a circular sample dish. Each sample was measured twenty times, with integration time of ten milliseconds.

All data preparation and construction of statistical models were performed with MATLAB 8.2.0. The preprocessing of the spectral data was the same with reference [11].

2.2 Division of calibration set and prediction set

All 75 samples were divided into calibration set and prediction set. To get stable prediction results, the calibration set and the prediction set were randomly divided for fifty times. The calibration set consisted of ten human blood samples randomly chosen from all fifteen human blood samples and thirty animal blood samples—twenty macaque randomly chosen from all twenty-nine macaque blood samples, ten guinea pig blood samples randomly chosen from all seventeen guinea pig blood samples. The prediction set consisted of thirty-five unknown spectra–nine macaque, seven guinea pig, five human samples (the species within the model) and twelve rat, one chicken, one pig samples (the species out of the model). The blood samples used in the prediction model were all the rest blood samples except for the calibration set. Calibration models were established for each division, and model prediction effects (e.g. root mean square error of prediction set, RMSEP) in fifty different divisions were calculated and then averaged. Based on the averaged RMSEP values, the stable optimal model was selected.

2.3 MWPLS method

MWPLSR is a wavelength interval selection method for multicomponent spectra analysis. Briefly, MWPLSR builds a series of PLS models in a window that moves over the whole spectral region and then locates useful spectral intervals in terms of the least complexity of PLS models reaching a desired error level [13]. Take the position and length of wavebands as well as the PLS factor into consideration, the search parameters are set as follows: (1) the beginning wavelength (B), (2) the number of wavelengths (N), and (3) the number of PLS factors (F). The PLS models are established for any combination (B, N, F). The optimal model is selected according to the minimum RMSEP.

2.4 ECMLR method

Equidistant combination multiple linear regression (ECMLR) is a method for equidistant discrete wavelengths selection method [16]. The parameters of ECMLR are as follows: (1) the beginning wavelength (B), (2) the number of wavelengths (N), (3) the number of wavelength gaps (G). The search range of ECMLR may cover the entire scanning region. The Sketch map for equidistant combination multiple linear regression was shown in Fig.1. The whole spectra were from 349 nm to 1148 nm. When the selected (B, N, G) combination was (500, 5, 30), the selected wavelength were 500 nm, 530 nm, 560 nm, 590 nm and 620 nm exactly, which was shown as circles in Fig. 1. Models with each combination of any beginning wavelength, number of wavelengths, and number of wavelength gaps were validated according to the prediction effect of the PLSDA. In ECMLR, the combination with the smallest RMSEP was

selected. In practical measurement systems, the less numbers of wavelengths were needed, the easier the system will be. Therefore, the maximum number of wavelengths was constrained in this algorithm to avoid selecting the smallest RMSEP with a relatively high number; i.e., the selected wavelengths number by validation method must not be larger than the maximum number [17]. In this study, we set the maximum number of wavelengths to be 16.

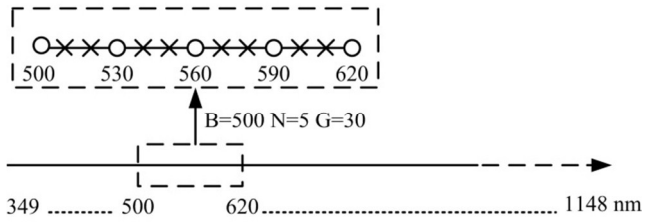


Fig.1. Sketch map for equidistant combination partial linear regression, with the circles being the selected wavelengths

3. Results

3.1 Wavelength selected

At a rough ratio of 1:1, all 75 blood samples were divided into calibration set (40 samples) and prediction set (35 samples). Fifty divisions were randomly generated. We tried to use the MWPLS method and ECMLR method for wavelength selection. According to Section 2.2, all samples were divided into calibration set and prediction set for 50 times, and calibration models were established for each division. For each combination of model parameters, such as for the same parameter (B, N, F) of MWPLS method and for the same parameter (B, N, G) of ECMLR method, the RMSEPs of the models were calculated for all 50 divisions, and then the mean value of them were calculated and denoted by $RMSEP_{AVE}$. In this paper, $RMSEP_{AVE}$ was chosen as the evaluation indicator for the optimization of model parameters. Namely, according to the minimum $RMSEP_{AVE}$, the corresponding optimal model parameter (B, N, and F) or (B, N, G) was selected.

For MWPLS, we set B from the first wavelength to the 640th wavelength (349-848 nm), N from 16 to 400, F from 1 to 16. The MWPLS wavelength selection was applied based on $RMSEP_{AVE}$ of each combination of (B, N, F). The wavelength selected by MWPLS method was from the 124th wavelength, and the following 318 wavelengths. Fig.2 [11] showed the mean diffuse reflectance spectra of animal and human and the selected wavelengths by MWPLS method, using black dot marks for animal and red dot marks for human, with all wavelengths from 448 nm to 698 nm.

For ECMLR, we set B from the first wavelength to the 240th wavelength (349-540 nm), N from 1 to 16, G from 1 to 50. The ECMLR wavelength selection was applied based on $RMSEP_{AVE}$ of each combination of (B, N, G). The wavelengths selected by ECMLR method was from the 20th wavelength, at intervals of 37 wavelengths, 12 wavelengths in total—365 nm, 394 nm, 424 nm, 454 nm, 483 nm, 513 nm, 542 nm, 571 nm, 600 nm, 629 nm, 657 nm, 686 nm. Fig.3 [11] showed the mean diffuse reflectance spectra of animal and human and the selected wavelengths by ECMLR method, using circle marks for animal and diamond marks for human.

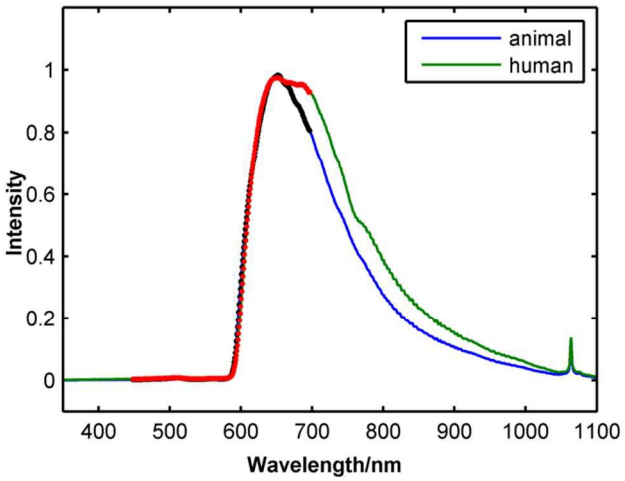


Fig.2. MWPLS wavelength selection on the mean spectra of animal and human: the selected wavelengths by MWPLS method, using black dot marks for animal, and red dot marks for human, with all wavelengths from 448 nm to 698 nm

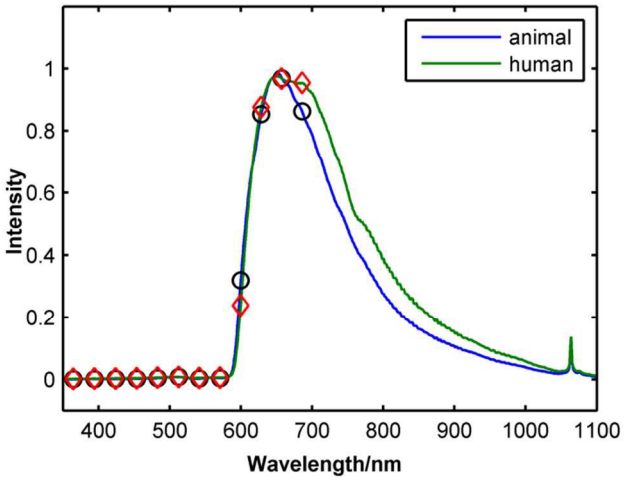


Fig.3. ECMLR wavelength selection on the mean spectra of animal and human: the black circles on the blue line and the red diamonds on the red line were all the selected wavelengths, including 365 nm, 394 nm, 424 nm, 454 nm, 483 nm, 513 nm, 542 nm, 571 nm, 600 nm, 629 nm, 657 nm, 686 nm

3.2 Prediction result

We showed the prediction results of full-spectrum method, MWPLS method and ECMLR method in table 1. The factor used in full-spectrum model was eighteen, which was described in reference [11]. The prediction accuracy of the full-spectrum method was 96.58% for the average of the whole 50 times validations. The result of MWPLS method was 97.20%, and ECMLR method was 99.65%. All three methods performed well, and the ECMLR method did the best. The prediction effect of ECMLR method was shown in Fig. 4, in which the

blue diamonds represented the animal samples in the training set (samples 1-30) and the prediction set (samples 41-56), the black diamonds represented the prediction set (the species out of the model, samples 62-75), and the red circles represented the human samples in the training set (samples 31-40) and the prediction set (samples 57-61).

Table 1 Summary of the prediction effect of three methods.

Wavelength selected	Prediction ability
Full wavelength (F=18)	96.58%
MWPLS (124^{th}, 318, 14)	97.20%
ECMLR (20^{th}, 12, 37)	99.65%

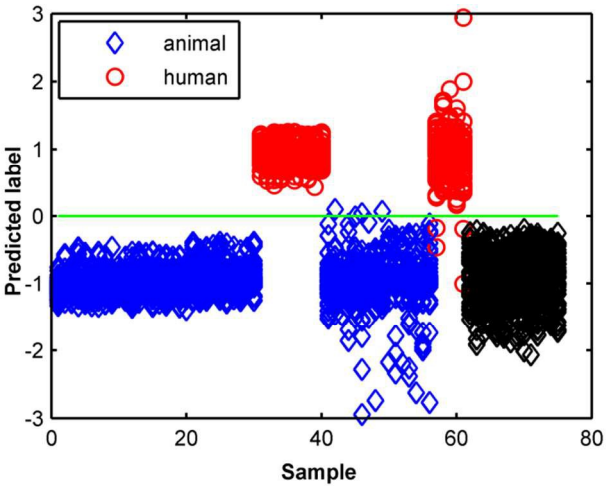


Fig.4. Predicted values for the prediction set with the wavelengths selected by ECMLR method: the blue diamonds represented the animal samples in the training set (samples 1-30) and the prediction set (samples 41-56), the black diamonds represented the prediction set (the species out of the model, samples 62-75), and the red circles represented the human samples in the training set (samples 31-40) and the prediction set (samples 57-61)

These results illustrate the ability of these methods to identify human and nonhuman blood samples, and both wavelength selection methods work better than full-spectrum method. This demonstrates that both wavelength selection methods were useful in constructing more reliable models.

3.3 Further validation

In order to further demonstrate the prediction ability of the wavelengths selected by MWPLS method and ECMLR method, we added blood samples within 24 hours, including eleven guinea pig blood samples and twenty-seven human blood samples. Six guinea pig blood samples, randomly chosen from eleven guinea pig blood samples, and fifteen human blood samples, randomly chosen from twenty-seven human blood samples, were added to the training dataset. Five guinea pig blood samples and twelve human blood samples were added to the prediction dataset. The blood samples added in the prediction model were all the rest blood samples except for the samples added in the calibration set.

Fifty divisions of calibration dataset and validation dataset were randomly generated. And the prediction results of full-spectrum, MWPLS method and ECMLR method are shown in Table 2. The prediction accuracy of the full-spectrum method was 95.62% for the average of the whole fifty times validations. The results of MWPLS method and ECMLR method were 96.38% and 97.23%, respectively. All three methods were validated to be great for discrimination of human and nonhuman species, and the ECMLR method did the best. The prediction results of validation by ECMLR methods was shown in Fig. 5, in which the blue diamonds represented the animal samples in the training set (samples 1-36) and the prediction set (samples 62-77), the black diamonds represented the prediction set (the species out of the model, samples 95-113), and the red circles represented the human samples in the training set (samples 37-61) and the prediction set (samples 78-94). So it was further confirmed that the prediction effect of wavelength selection methods was better than full-spectrum method.

Table 2 Summary of the prediction effect of three methods (section 3.3)

Wavelength selected	Prediction ability
Full wavelength (F=18)	95.62%
MWPLS (124^{th}, 318, 14)	96.38%
ECMLR (20th, 12, 37)	97.23%

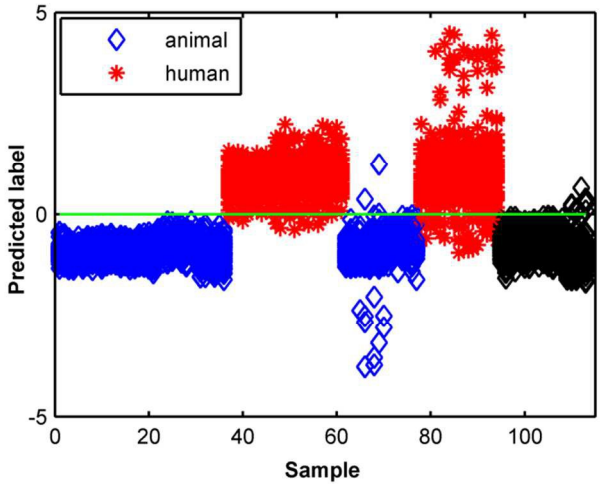


Fig.5. Predicted values for the prediction set with the wavelengths selected by ECMLR methods in further validation part: the blue diamonds represented the animal samples in the training set (samples 1-36) and the prediction set (samples 62-77), the black diamonds represented the prediction set (the species out of the model, samples 95-113), and the red circles represented the human samples in the training set (samples 37-61) and the prediction set (samples 78-94)

4. Discussion

Given that blood is complex with multiple components, the spectroscopic analysis of major components in blood has to mitigate noise disturbance. In addition, appropriate wavelength selection and stability are two key objectives. And prediction

effects and parameters were sensitive to the differences of dividing calibration and prediction sets, which yields unreliable results. In this study, a portion of samples was randomly selected as a validation set.

From the prediction analysis, the wavelength selection methods were beneficial to the model for discriminating the blood species. MWPLS method and ECMLR method both have many advantages, and also they have different features. A salient advantage of MWPLSR is that the calibration model is very stable against the interference from environmental factors. Moreover, the selection of spectral intervals in terms of the least model complexity enables the reduction of the size of a calibration sample set in calibration modeling. However, for fixed-filter spectrometer measurement systems, the less numbers of wavelengths were used, the easier the system will be. Therefore, MWPLS method was not the optimal choice. Equidistant combination multiple linear regression yields an equidistant discrete wavelength combination which improves the conditioning of MLR by minimizing co-linearity effects when the data gap selection is appropriate. A proper quasi-continuous wavelength combination can be effectively screened from a continuous waveband along with various numbers of wavelengths. The advantages of ECMLR are a low degree of freedom and low computational complexity, which inherits the merits of both continuous mode and the discrete mode.

5. Conclusions

Optical wavelength selection method makes it possible for the visible diffuse reflectance measurement system simplified, and makes it more economical to apply on screening blood samples for customs supervision demands. Both MWPLS method and ECMLR method were applied to establish calibration models for discriminating human and nonhuman blood species. The prediction performance of ECMLR wavelength selection method was proved to be better than MWPLS method. Furthermore, the model used only 12 wavelengths, at intervals of 37 wavelengths, 12 wavelengths in total—365 nm, 394 nm, 424 nm, 454 nm, 483 nm, 513 nm, 542 nm, 571 nm, 600 nm, 629 nm, 657 nm, 686 nm, thus reducing method complexity substantially. The span of the optimal ECMLR wavelengths was almost in the whole visible light region. In addition, ECMLR method was in the ‘‘quasi-continuous’’ mode, which could easily undergo spectral pre-processing to improve predictive capability comparing with other discrete-mode wavelength selection method. Thus, the ECMLR method has great potential in practical application and instrument design. These findings provide valuable reference for the design of specialized quasi-continuous spectrometers.

Acknowledgements

We are grateful to the help from the Chinese Academy of Medical Science & Peking Union Medical College Institute of Biomedical Engineering, Institute of Zoology, Chinese Academy of Science and No. 254 Hospital of People’s Liberation Army. This research was supported by National High-Tech R&D Program of China (863 Program: 2015AA021105). This project was also supported by Tianjin Application Basis & Front Technology Study Programs (No. 11JCZDJC17100 and No. 14JCZDJC33100).

Notes and References

[a] State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin 300072
[b] Tianjin Key Laboratory of Biomedical Detecting Techniques & Instruments, Tianjin University, Tianjin 300072
[c] Institute of Biomedical Engineering, Chinese Academy of Medicl Sciences & Peking Union Medical College, Tianjin 300192
[d] Medical Examination Centre, No.254 Hospital of People's Liberation Army, Tianjin 300142
Corresponding author: linling@tju.edu.cn.

1 J. Duarte, M. T. Pacheco and R. Z. Machado. Mol Biol (Noisy-le-grand), 2002, **48**, 585.
2 N. Terada, N. Ohno and S. Saitoh. J Struct Biol, 2008, **163**, 147.
3 R. K. Stroud, W. J. Adrian In Noninfectious Diseases of Wildlife. Iowa State University Press, 1996.
4 R. P. Spalding. In Forensic Science: An Introduction to Scientific and Investigative Techniques CRC Press: Boca Raton, FL, 2003.
5 R. Li. Forensic Biology CRC Press: Boca Raton, FL, 2008.
6 H. Inouel, F. Takabe, O. Takenaka. Int. J. Legal Med, 1990, **104**, 9.
7 E. O. Espinoza, N. C. Lindley, K. M. Gordon, J. A. Ekhoff. Anal Biochem, 1999, **268**, 252.
8 K. De Wael, Forensic Science International, 2008, **180**, 37.
9 K. Virkler and I. K. Lednev, Analytical and Bioanalytical Chemistry, 2009, **396**, 525.
10 G. McLaughlin, K. C. Doty, I. K. Lednev, Forensic Science International, 2014, **238**, 91.
11 L. Zhang, M. Zhou, X. Li, G. Li, L. Lin. Analytical Methods, 2014, **6**, 9419.
12 X. Zhang, W. Li, B. Yin, W. Chen, D.P. Kelly, X. Wang, K. Zheng, Y. Du. Spectrochim. Acta A, 2013, **114**, 350.
13 J. H. Jiang, R. J. Berry, H. W. Siesler, Y. Ozaki. Anal. Chem, 2002, **74**, 3555.
14 H. Chen, T. Pan, J. Chen. Chemometrics and Intelligent Laboratory Systems, 2011, **107**, 139.
15 A. J. O’Neil, R. D. Jee, A. C. Moffat. Analyst, 1998, **123**, 2297.
16 T. Pan, M. Li, J. Chen. Applied Spectroscopy, 2014, **68**, 263.
17 S. Kasemsumran, Y. P. Du, K. Murayama. Analytica Chimica Acta, 2004, **512**, 223.