Journal of Innovative Optical Health Sciences
Vol. 7, No. 6 (2014) 1450022 (7 pages)

DOI: 10.1142/S1793545814500229

The relevance study of effective information between near infrared spectroscopy and chondroitin sulfate in ethanol precipitation process

Lian Li*, Baoyang Ding*, Qi Yang*, Shang Chen*,
Huaying Ren†, Jinfeng Wang*, Hengchang Zang*,‡,
Fengshan Wang* and Lixuan Zang*

*School of Pharmaceutical Sciences and
National Glycoengineering Research Center
Shandong University, No. 44 Wenhuaxi Road
Jinan 250012, P. R. China

†School of Chemistry and Chemical Engineering
Shandong University, No. 27 Shandanan Road
Jinan 250010, P. R. China
‡zanghcw@126.com

Received 17 July 2013
Accepted 1 December 2013
Published 14 January 2014

Near infrared spectroscopy (NIRS) is based on molecular overtone and combination vibrations. It is difficult to assign specific features under complicated system. So it is necessary to find the relevance between NIRS and target compound. For this purpose, the chondroitin sulfate (CS) ethanol precipitation process was selected as the research model, and 90 samples of 5 different batches were collected and the content of CS was determined by modified carbazole method. The relevance between NIRS and CS was studied throughout optical pathlength, pretreatment methods and variables selection methods. In conclusion, the first derivative with Savitzky–Golay (SG) smoothing was selected as the best pretreatment, and the best spectral region was selected using interval partial least squares (iPLS) method under 1 mm optical cell. A multivariate calibration model was established using PLS algorithm for determining the content of CS, and the root mean square error of prediction (RMSEP) is 3.934 $g \cdot L^{-1}$. This method will have great potential in process analytical technology in the future.

Keywords: Chondroitin sulfate; near infrared spectroscopy; variable selection; pathlength.

1. Introduction

Near infrared spectroscopy (NIRS) is a rapid and nondestructive analytical technology, which is also treated as one of the most efficient process analytical tools to analyze complicated components.[1] The most prominent absorption bands of near infrared region are related to overtones and combinations of fundamental vibrations of C–H, N–H, O–H and S–H groups.[2] NIR spectra are composed of a large number of exploitable variables, but there are a large number of unwanted variables, causing colinearity, will reduce the predictive capability of the method.[3] Therefore, it is difficult to find a specific band for specific compound. It is very important to investigate the relevance of effective information between NIRS and objective compound. The factors which affect the relevance include optical pathlength, pretreatment, variables selection, etc. In order to get a good spectrum, selection of appropriate pathlength is very important especially for analysis of trace component.[4] Pretreatment methods are also essentially important. First derivative can eliminate irrelevant base-line drifting while second derivative can remove the linear background to near zero.[5] Savitzky–Golay (SG) smoothing can preserve the feature of raw spectra. Standard normal variate (SNV) transformation[6] seems to be suitable to remove the multiplicative interferences of scatter and particle size.[5] On the other hand, selection of variables will solve the collinearity between spectral variables and eliminate the information that are useless, as well as decrease the cost of instrument, and improve the interpretability of the results ultimately.[7]

In this study, chondroitin sulfate (CS) was selected as a substance model. It is a compound which is composed alternatively by D-glucuronic acid and differently sulfated residues of N-acetyl-D-galactosamine linked by β(1–3) bonds,[8–10] and the diversity of CS makes it have a wide range of applications in clinical. The production of CS includes whole processes from raw material to product storage. It generally refers to evaluation of raw materials, CS extraction, ultrafiltration, oxidation, precipitation, drying, grinding and packaging. Above all, ethanol precipitation which will lead to an unstable product quality and differences between batches is the most important unit used in industrial production[11–14] and the relevance between NIRS and CS is still unclear. Therefore, a CS ethanol precipitation process was selected as the research process for investigating the relevance between NIRS and CS.

This study was aimed at revealing the relevance of effective information between NIRS and CS and establishing a PLS model during the process of ethanol precipitation. The optical pathlength was investigated using 1 and 4 mm optical glass cells. Different pretreatment methods including SG smoothing, SNV transformation and derivatives were used to enhance the interpretability of the spectra. Variables selection methods including manual and interval partial least squares regression (iPLS) methods were used to select the best spectra region related to objective compound. Finally, a PLS model based on the best parameters above was constructed.

2. Materials and Methods

2.1. *Materials*

Crude CS from shark cartilage was obtained from Yantai Dongcheng Biochemical Limited Company of Shandong Province in China. D-glucuronic acid was purchased from Sigma-Aldrich Co. LLC. Ethanol was purchased from Tianjin Fuyu Fine Chemical Co., Ltd. Freshly prepared sulfuric acid with sodium borate was obtained by dissolving 4.77 g of sodium borate in 500 mL of sulfuric acid. Carbozole (CP) was from Tianjin Guangfu Fine Chemical Research Institute. All the other reagents were of analytical grade. Deionized water was purified by Milli-Q water system (Millipore Corp., Bedford, MA, USA).

2.2. *Ethanol precipitation process*

About 14 g CS and 4 g NaCl were dissolved in 200 mL of deionized water. After that, 800 mL ethanol was gradually added through a peristaltic pump in 10 mL·min^{-1} and 1 mL supernatant was collected for analysis. The first 7 samples were collected every 3 min, and then 8 samples were collected every 5 min, and last 3 samples were collected every 10 min. There were 18 samples for each batch and 5 normal batches were repeated in all.

2.3. *Reference method*

The content of CS was determined using carbazole method[15] with modification. Some parameters

including precision, repeatability, stability and recovery of method were determined. All samples were loaded onto a 96-well microplate (200 μL/well) and the absorbance was measured at 540 nm with a microplate reader (RIO-RAD, Model-680).[16]

2.4. *NIR analyzer*

Fourier transform near infared spectrometer (Antaris II, Thermo Fisher, USA) with InGaAs detector were used to collect the spectra. The room temperature and humidity was steady in the laboratory. All samples were loaded in optical cells (1 and 4 mm optical length) after centrifugation, and each spectrum was the average of 32 scans with a resolution of 4 cm^{-1}. The spectral range was from 10,000 to 4000 cm^{-1}, and background spectrum (32 scans) was taken before the measurement of every sample. Data analysis including pretreatment and variable selection was performed by Matlab (7.10.0 R2010a, The Math Works, Inc., USA) and PLS toolbox (7.0.2, Eigenvector Research, Inc., USA).

2.5. *Sample dividing*

The samples from five batches were collected under the same condition, and were divided into two sets including calibration set and validation set with random method. A total of 54 samples from 3 batches were used to construct the calibration set and the remaining 36 samples of 2 batches were used as a validation set which is used to validate the model's predictive ability.

2.6. *Model efficiency estimation*

In order to characterize the prediction ability of a created PLS model, the coefficient of determination of calibration set (R_c^2), coefficient of determination of validation set (R_p^2), root mean square error of calibration (RMSEC) and root mean square error of prediction (RMSEP) were used. The equations were as follows:

$$R_c^2 = 1 - \frac{\sum_{c=1}^{n} (Y_c - \hat{Y}_c)^2}{\sum_{c=1}^{n} (Y_c - Y_l)^2}, \tag{1}$$

$$R_p^2 = 1 - \frac{\sum_{p=1}^{m} (Y_p - \hat{Y}_p)^2}{\sum_{p=1}^{m} (Y_p - Y_k)^2}, \tag{2}$$

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{c=1}^{n} (\hat{Y}_c - Y_c)^2}{n}}, \tag{3}$$

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{p=1}^{m} (\hat{Y}_p - Y_p)^2}{m}}, \tag{4}$$

where Y_c is the reference result from reference method of calibration set for sample c, $\hat{Y}_c$ is the calculated value of calibration set for sample c from NIRS, Y_l is the mean value of Y_c from calibration set, n is the number of calibration set sample, Y_p is the reference result from reference method of validation set for sample p, $\hat{Y}_p$ is the calculated value of validation set for sample p from NIRS, Y_k is the mean value of Y_p from validation set, m is the number of calibration set sample.

3. Results and Discussion

3.1. *Determination of CS in supernatant*

For quantitative consideration, the calibration equation was established as follows: $y = 0.0109x - 0.0587$ ($R = 0.9997$).

The precision of the reference method was carried out by continuous measuring of the CS absorbance of a single sample for six times with a microplate reader at 540 nm. The Relative standard deviation (RSD) value of absorbance was 0.12%.

Six parallel determination of CS content of single sample with a microplate reader at 540 nm gave a RSD value of 2.26% for repeatability.

The stability was carried out by measuring the sample (which was used for the determination of repeatability) content at 0, 10, 20, 30, 60, 90 and 120 min, separately, with a RSD value of absorbance 0.59%.

The accuracy of the method was confirmed by spike recovery test. The recovery is calculated according to the following equation: $R = (m_s - m_b)/m_c$, where m_s is the content of the CS sample and D-glucuronic acid, m_b is the content of the CS standard sample and m_c is the content of D-glucuronic acid. Three levels (0.3, 0.5 and 0.7 mL) of CS sample solution were used for recovery (shown in Table 1), and the RSD value is 2.26%.

The trend chart of CS content changing in five batches was shown in Fig. 1, and the initial content

Table 1. Sample solution preparation for recovery determination.

	1	2	3	4	5	6	7	8	9
Standard CS sample solution V (mL)	0.3	0.3	0.3	0.5	0.5	0.5	0.7	0.7	0.7
D-glucuronic acid solution V (mL)	0.7	0.7	0.7	0.5	0.5	0.5	0.3	0.3	0.3
Sulfuric acid with sodium borate solution V (mL)	3	3	3	3	3	3	3	3	3
Carbozole solution V (mL)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

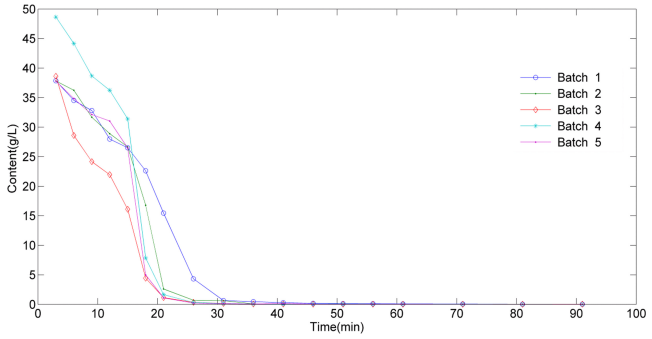


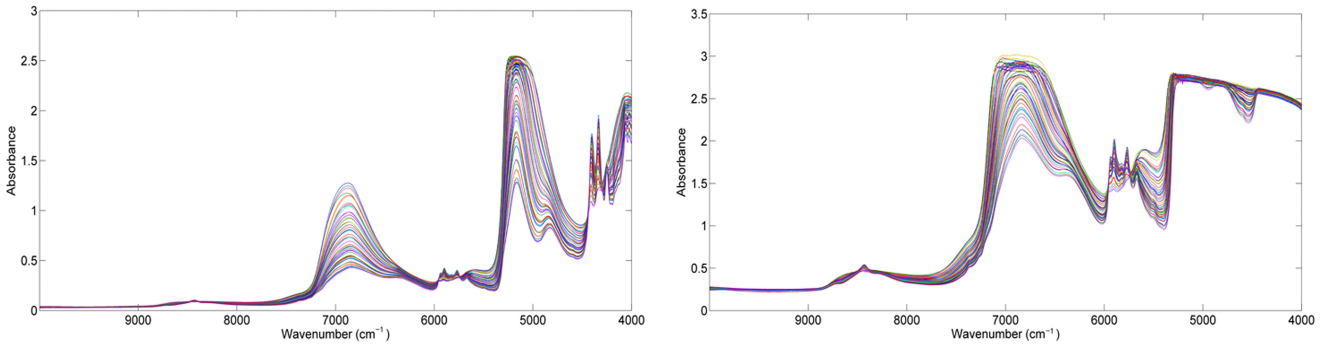
Fig. 1. The content trajectory of CS during ethanol precipitation process.

of the fourth batch was relatively high, which was because CS content of this batch was high in the original raw materials.

3.2. *Spectra interpretation and pretreatment*

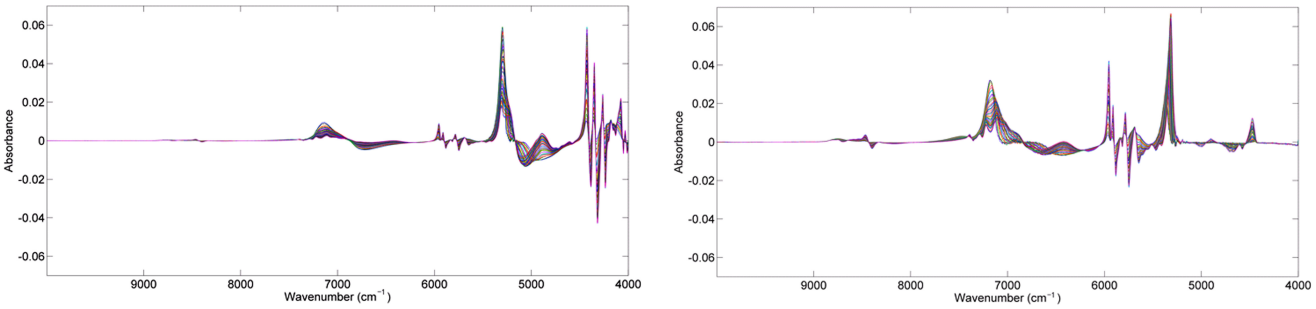
Figure 2 is the raw spectra of CS in the ethanol precipitation process based on different optical pathlength. According to Figs. 2(a) (optical pathlength is 1 mm) and 2(b) (optical pathlength is 4 mm), the raw spectra were mainly dominated by C–H and O–H,[17] which can be seen in the range from around 4500 to 4300 cm^{-1} as the C–H combination vibration, around 5400 to 5100 cm^{-1} as O–H stretching vibration and bending vibration, 5950 to 5850 cm^{-1} as C–H first overtone vibration, 7000 to 6800 cm^{-1} as O–H first overtone vibration.[18]

In order to get more information from the spectra, proper data pretreatment is necessary before calibration.[3] Four different data pre-processing methods were utilized, including SG smoothing, SG smoothing with first and second derivative (all of them with a window width 15 variables and polynomial order 2) and SNV. Finally, the first derivative with SG smoothing was selected as the pretreatment method. Figure 3 showed the spectra pretreated by the first derivative with SG smoothing based on different pathlength. It was obvious that O–H stretching vibration and bending vibration around 5400 to 5200 cm^{-1} has the strongest absorption in both 1 and 4 mm pathlength. However, O–H first overtone vibration around 7300 to 7000 cm^{-1} of 4 mm pathlength was much stronger than that in 1 mm pathlength. During the ethanol precipitation process, the range from 4545 to 4230 cm^{-1} and 6024 to 5617 cm^{-1} were regarded as the regions for absorption bands of C–H.[19] There were lots of spectral information under the 1 mm optical pathlength from 4545 to 4230 cm^{-1}, but only a small peak can be identified around 4540 to 442 cm^{-1} under 4 mm pathlength.



(a) (b)

Fig. 2. Raw near infrared spectra of supernatants during CS ethanol precipitation process (A. 1 mm optimal pathlength, B. 4 mm optimal pathlength).



(a) (b)

Fig. 3. Near infrared spectra of supernatants during CS ethanol precipitation process pretreated by first derivative (A. 1 mm optimal pathlength, B. 4 mm optimal pathlength).

3.3. *Outlier detection*

An important step in building a PLS model is the identification of outliers because PLS calibration method is strongly influenced by the presence of outliers.[20,21] Therefore, some outliers detection methods[22–25] such as student residence and leverage methods were used to remove outliers. According to this method, no samples in both the 1 and 4 mm pathlength model were considered anomalous in the full spectral regions evaluated, thereby remaining the dataset with 90 samples.

3.4. *Variable selection method*

The NIR spectrum often contains hundreds of wavelengths (variables). So it is important to identify variables that contribute the most useful information. Variable selection is a critical step in data analysis for NIRS. Two different methods were used for variable selection including manual approach and iPLS. However, when using manual approach, the relationship between absorption in the NIRS and the target analytical parameter was always nonlinear and thus difficult to identify. In order to resolve this problem, the original spectra were pretreated by first derivative with SG smoothing, and variables that had no use were removed manually. Also iPLS methods which can identify information rich regions of the spectra were used to construct a more robust multivariate model.[26,27] And the region with smallest prediction error was selected.

The results of the two models were shown in Table 2. According to the table, the results of 1 mm pathlength model using iPLS method was the best (with every 50 cm^{-1} intervals automatically, Fig. 4). It was C–H first overtone vibration, and the absorbance was highly correlated with the content of CS. The results of the value R^2 and of the root mean square error of calibration (RMSEC) of this region were much better than any others. In terms of 4 mm

Table 2. Results of PLS models of different variables based on different pathlength.

Variable selection method	No. of latent variables	R_c^2	RMSEC (g/L)
Full region[a]	6	0.939	2.979
iPLS[a]	9	0.973	1.584
Manual selection[a]	10	0.958	1.663
Full region[b]	5	0.965	1.818
iPLS[b]	8	0.936	2.209
Manual selection[b]	6	0.929	2.982

[a]Variable selection is based on 1 mm pathlength.
[b]Variable selection is based on 4 mm pathlength.

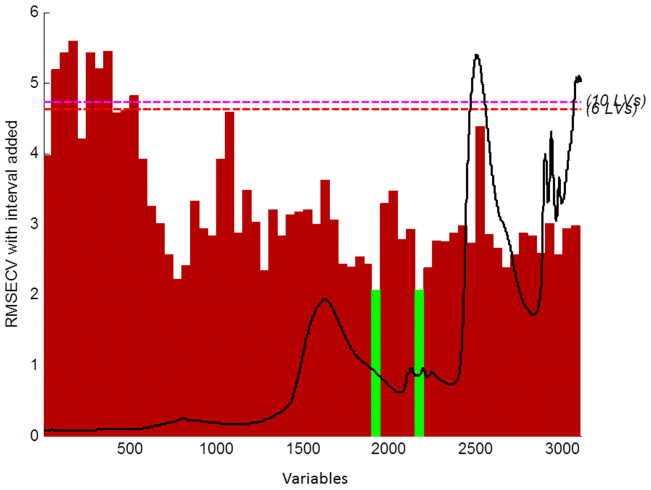


Fig. 4. Variable selection results based on 1 mm optimal pathlength.

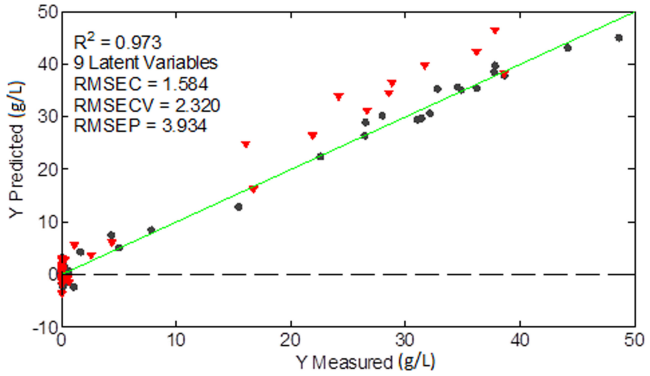


Fig. 5. Prediction results of PLS model based on 1 mm optimal pathlength.

pathlength model, the RMSEC result of iPLS was not so good as the results of 1 mm model. Therefore, the 1 mm model was built after unwanted bands were removed.

3.5. *The PLS model*

After evaluating the possible existence of outliers and selecting the appropriate variables and pathlength, a robust PLS model was established. And the model was validated according to validation set. The RMSEP was used to evaluate the model. The model was established as shown in Fig. 5. It can be seen that the No. of latent variables was 9, the R_p^2 was 0.973 and the RMSEP was 3.934 g·L^{-1}. RMSEP was the most important parameter for evaluating the model's predictive ability.

4. Conclusion

This research studied the relevance of effective information between NIRS and CS during the process of ethanol precipitation. By comprehensive comparison of the R_c^2, RMSEC and RMSEP, it can be concluded that the model built in 1 mm optical pathlength is better than the one built in 4 mm pathlength. Hence, it is better to choose a shorter pathlength when dealing with complicated samples, because it can decrease the error. And it can be concluded that the raw spectrum in 1 mm pathlength is easy to select related variables through iPLS because it includs more information than 4 mm pathlength. Meanwhile, variable selection is simply a way for enhancing the precision of the model. The results show that the selection of variable improved the veracity of the model, and the relevance between the effective information and spectral became more evident, and the model predictability and reliability became much better. In all, throughout this study, it provided a strategy for revealing the effective information of NIRS and a useful PLS model for monitoring the manufacture process.

Acknowledgments

We extend great thanks to National Glycoengineering Research Center of China. Then we appreciate the Chinese National Level College Student Innovation Project (No. 1110422080), the 863 program (Hi-tech research and development program of China) under contract NO.2012AA021505 and the National Training Programs of Innovation and Entrepreneurship for Undergraduates (No.201210422079). At last, we are grateful for Yantai Dongcheng Biochemical Limited Company of Shandong Province in China for providing CS samples.

References

1. F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, "Mutual information for the selection of relevant variables in spectrometric nonlinear modelling," *Chemometr. Intell. Lab.* **80**(2), 215–226 (2006).
2. G. Reich, "Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications," *Adv. Drug. Deliv. Rev.* **57**(8), 1109–1143 (2005).
3. T. Rajalahti, O. M. Kvalheim, "Multivariate data analysis in pharmaceutics: A tutorial review," *Int. J. Pharm.* **417**(1–2), 280–290 (2011).
4. P. S. Jensen, J. Bak, "Near-infrared transmission spectroscopy of aqueous solutions: Influence of optical pathlength on signal-to-noise ratio," *Appl. Spectrosc.* **56**(12), 1600–1606 (2002).
5. A. Candolfi, R. De Maesschalck, D. L. Massart, P. A. Hailey, A. C. E. Harrington, "Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA," *J. Pharm. Biomed. Anal.* **19**(6), 923–935 (1999).
6. R. J. Barnes, M. S. Dhanoa, S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Appl. Spectrosc.* **43**(5), 772–777 (1989).
7. Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, M. Hanpin, "Variables selection methods in near-infrared spectroscopy," *Anal. Chim. Acta.* **667**(1–2), 14–32 (2010).

8. N. Volpi, A. Mucci, L. Schenetti, "Stability studies of chondroitin sulfate," *Carbohydr. Res.* **315**(3–4), 345–349 (1999).
9. H. Zang, L. Li, F. Wang, Q. Yi, Q. Dong, C. Sun, J. Wang, "A method for identifying the origin of chondroitin sulfate with near infrared spectroscopy," *J. Pharm. Biomed. Anal.* **61**, 224–229 (2012).
10. K. Meyer, E. Davidson, A. Linker, P. Hoffman, "The acid mucopolysaccharides of connective tissue," *Biochim. Biophys. Acta.* **21**(3), 506–518 (1956).
11. L. C. F. Silva, "Isolation and purification of chondroitin sulfate," Advances in Pharmacology, Nicola V (Ed) 21–31, Academic Press (2006).
12. B. Xu, Z. Wu, Z. Lin, C. Sui, X. Shi, Y. Qiao, "NIR analysis for batch process of ethanol precipitation coupled with a new calibration model updating strategy," *Anal. Chim. Acta.* **720**, 22–28 (2012).
13. Z. Wu, B. Xu, M. Du, C. Sui, X. Shi, Y. Qiao, "Validation of a NIR quantification method for the determination of chlorogenic acid in Lonicera japonica solution in ethanol precipitation process," *J. Pharm. Biomed. Anal.* **62**(0), 1–6 (2012).
14. H. Huang, H. Qu, "In-line monitoring of alcohol precipitation by near-infrared spectroscopy in conjunction with multivariate batch modeling," *Anal. Chim. Acta.* **707**(1–2), 47–56 (2011).
15. B. M. van den Hoogen, P. R. van Weeren, M. Lopes-Cardozo, L. M. G. van Golde, A. Barneveld, C. H. A. van de Lest, "A microtiter plate assay for the determination of uronic acids," *Anal. Biochem.* **257** (1), 107–111 (1998).
16. M. Cesaretti, E. Luppi, F. Maccari, N. Volpi, "A 96-well assay for uronic acid carbazole reaction," *Carbohydr. Polym.* **54**(1), 59–61 (2003).
17. J. Märk, M. Andre, M. Karner, C. W. Huck, "Prospects for multivariate classification of a pharmaceutical intermediate with near-infrared spectroscopy as a process analytical technology (PAT) production control supplement," *Eur. J. Pharm. Biopharm.* **76**(2), 320–327 (2010).
18. J. Workman, L. Weyer, *Practical Guide to Interpretive Near-Infrared Spectroscopy*, CRC press (2007).
19. H. Morita, T. Hasunuma, M. Vassileva, R. Tsenkova, A. Kondo, "Near infrared spectroscopy as high-throughput technology for screening of xylose-fermenting recombinant saccharomyces cerevisiae strains," *Anal. Chem.* **83**(11), 4023–4029 (2011).
20. X. Bao, L. Dai, "Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties," *Fuel* **88**(7), 1216–1222 (2009).
21. E. Furusjö, A. Svenson, M. Rahmberg, M. Andersson, "The importance of outlier detection and training set selection for reliable environmental QSAR predictions," *Chemosphere* **63**(1), 99–108 (2006).
22. M. A. M. Silva, M. H. Ferreira, J. W. B. Braga, M. M. Sena, "Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy," *Talanta* **89**, 342–351 (2012).
23. P. Valderrama, J. W. B. Braga, R. J. Poppi, "Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy," *J. Agric. Food. Chem.* **55**(21), 8331–8338 (2007).
24. B. Walczak, D. L. Massart, "Multiple outlier detection revisited," *Chemometr. Intell. Lab. Syst.* **41**(1), 1–15 (1998).
25. M. H. Zhang, J. Luypaert, J. A. Fernández Pierna, Q. S. Xu, D. L. Massart, "Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration," *Talanta* **62**(1), 25–35 (2004).
26. T. Mehmood, K. H. Liland, L. Snipen, S. Sæbø, "A review of variable selection methods in Partial Least Squares Regression," *Chemometr. Intell. Lab. Syst.* **118**, 62–69 (2012).
27. S. Cho, K. Kwon, H. Chung, "Varied performance of PLS calibration using different overtone and combination bands in a near-infrared region," *Chemometr. Intell. Lab. Syst.* **82**(1–2), 104–108 (2006).