10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.1, June 4-6, 2007, Cancún, Mexico

Determination of protein and fat content in fermentation raw materials with NIR reflectance spectroscopy

Ana Patrícia Ferreira, José Cardoso de Menezes

IBB-Institute for Biotechnology and Bioengineering, Centre for Biological and Chemical Engineering, Instituto Superior Técnico Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Av. Rovisco Pais, P-1049-001 Lisbon, Portugal, Tel. (+351) 21 841 9838;
Fax. (+351) 21 841 9197; bsel@ist.utl.pt

Abstract: NIR spectra of sixteen lots of pharmaceutical-grade defatted soybean flour with tightly controlled granulometry were collected and their protein and fat content analysed by standard reference methods. Partial least squares calibration models were developed for both parameters. A new strategy for the use of genetic algorithms for variable selection, weighted genetic algorithm, was tested. This strategy enabled improvements on the predictive ability of the calibration model around 50 %. The spectroscopic method developed will enable faster assessment of raw material quality and will be used to speed up the process of acceptance of new lots. Copyright © 2007 IFAC

Keywords: near-infrared spectroscopy, fermentation, raw material qualification, reference methods, multivariate calibration

1. INTRODUCTION

Soybeans contain approximately 35% protein, 19% lipid and 9 – 12% total sugar, thus constituting an excellent source of nutrients for microbial growth. They are economical, complex nitrogen and carbon sources and, in addition, they also provide minerals, vitamins and other growth factors (Jones and Porter, 1998). Soybean derived products like soybean flour are key medium ingredients for many industrial fermentation processes, e.g., the production of active pharmaceutical ingredients (API).

Upon arrival to the fermentation plant, each new lot of soybean flour must be analysed to assess its quality, prior to the use in fermentation media formulations. The most relevant parameters are total protein and total fat content. Moisture and ash content are also usually determined. Standard reference methods for protein content determination involve the determination of nitrogen, by Kjeldahl method (Kirk and Sawyer, 1991), or combustion analysis (Kirk and Sawyer, 1991), and conversion to crude protein content by a multiplicative factor based on the estimate of the nitrogen percentage in the protein. Fat content is typically determined by Soxhlet extraction (Kirk and Sawyer, 1991), with the aid of an organic solvent (e.g., petroleum ether).

There are several published papers related to protein and fat content determination with NIRS in different materials (Hong et al, 1994; Cozzolino et al, 1996; Kurowski, 1998; Kays et al, 2000; Tarkosova and Copikova, 2000). Near infrared spectroscopy (NIRS) is a widely accepted and well-established technique in food and agricultural industries, which enables the analysis of complex samples in a rapid, non-destructive way, without complex sample pre-treatment (Workman, 1993; Kumagal et al, 2002). The main drawbacks of this spectroscopic technique are the high detection limit and the time and effort required to produce a robust calibration (this is an indirect analytical method).

A near infrared (NIR) spectrum consists of a number of overlapped bands enclosing a great amount of information. This prevents the use of a classical variable selection approach, based on the knowledge about the spectroscopic properties of the sample. The extraction of quantitative and qualitative information from the spectra requires the application of statistical methods and mathematical techniques (Bokobza, 1998).

One potential drawback of using mathematical techniques instead of spectroscopic knowledge for variable selection is the possibility of overfitting (i.e., removing an excessively high number of variables will cause the model to perform well on calibration but not on external validation). When overfitting occurs, less robust models will be obtained (Swierenga et al, 2000).

The aim of this work is to investigate the possibility of applying NIRS and chemometrics for quantitative analysis of soybean flour, with the purpose of developing a new method for qualification of fermentation raw material. A new strategy for the use of genetic algorithms (GA) for variable selection is proposed. The main goal of this new technique is to deal with the stochastic nature of the GA and enable the achievement of more robust results.

2. MATERIALS AND METHODS

2.1 Samples and Reference Analysis

The type of soybean flour used in this study was pharmaceutical grade defatted soybean flour, which is an important raw material of some industrial fermentation processes. The granulometry of the flour is tightly controlled to be around 200 µm. Samples from a total of 16 soybean flour lots, all manufactured by the same supplier, were collected and analysed in respect to their total protein and fat content. A combustion method was used for protein content determination (TruSpec CN, Leco, St. Joseph (MI), USA) and fat content was determined by Soxhlet extraction using a Soxtec analyser (Tecator, Hillerød, Denmark).

2.2 NIR Spectroscopy

The near-infrared diffuse reflectance spectra of the flour samples were acquired using a MB160 (ABB BOMEM, Québec, Canada) spectrometer equipped with an InAs detector and a motorized powder sampling device FTLA – ACC101. Each spectrum was an average of 64 scans, recorded in the wavenumber range of 12000 – 4000 cm^{-1} and with a spectral resolution of 16 cm^{-1}. Spectra were recorded in triplicate for each sample with the aid of the Grams/AI™ software package (Thermo Galactic, Salem, MA, USA).

2.3 Data Analysis

All calculations were carried out using Matlab (MathWorks, Natick, MA, USA) and PLS Toolbox v 3.0 for Matlab (Eigenvector Research, Manson, WA, USA). Multivariate calibration models for protein and fat content were developed with the partial least squares regression (PLS) algorithm. The performance of the calibration models was assessed by computing root mean square errors of internal validation, RMSECV, see Eq. 1 where $\hat{y}_i$ is the predicted concentration for each sample, y_i the measured concentration and N is the total number of samples and the amount of variance in y being predicted by the model, Q$^2_{\text{Y}}$ (Eq. 2).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \qquad \text{Eq. 1}$$

$$Q_Y^2 = \left(1 - \frac{(y-\hat{y})^T(y-\hat{y})}{y^T y}\right) \times 100 \qquad \text{Eq. 2}$$

2.3 Genetic algorithm

Genetic algorithms (GA) are variable selection methods inspired by principles of genetics and natural selection (Abrahamsson et al, 2003, Leardi et al, 2000; Leardi et al, 2002). The main drawback of using GA for variable selection is that due to the stochastic nature of the algorithm, it does not always converges for the same optimal solution, i.e., different results are obtained in different runs of this algorithm (cf. Ferreira et al, 2005). To overcome this, the use of a weighted genetic algorithm (wGA) is proposed.

In this work, the spectral range was divided in segments (the genes), each containing 20 of the original variables. A chromosome contains as many genes as the number of segments created. The GA was initialized by selecting a population with n subsets (chromosomes), each containing a random combination of equally sized spectral intervals. A PLS model is built for each of the chromosomes and its quality evaluated based loss function (Q$^2_{\text{Y}}$, Eq. 2). Table 1 presents the parameters used for the genetic algorithm. The detailed description of the GA can be found in Ferreira et al (2005).

The genes are codified in binary code (e.g., for chromosome *C* with *I* genes, C(i) = 1 means that the i$^{\text{th}}$ interval will be selected for model development; C(i) = 0 signifies that the i$^{\text{th}}$ interval will not be considered for model development).

Table 1 - Parameters of the GA

Parameter	Value
Loss function	Q$^2_{\text{Y}}$ (cross-val)
Cross validation mode	Dynamic split & test (30% for validation set); 20 repetitions
Objective	Maximization of the loss function
Size of each individual	number of spectral regions
Size of population (genetic pool)	30
Number of iterations	100
Selection type	tournament
Probability of cross over	0.8
Probability of mutation	0.08

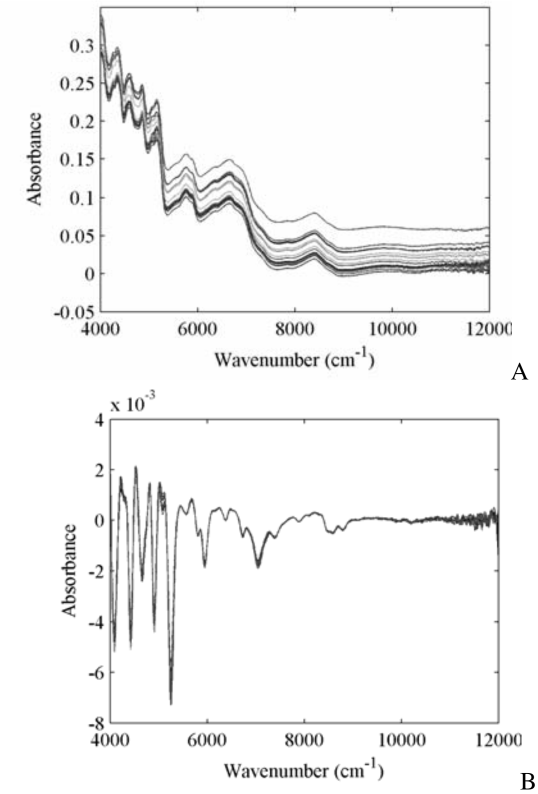

Fig. 1 – Soy flour samples spectra: A) raw spectra; B) first derivative of spectra

The wGA approach is aimed at reducing the stochastic component of this variable selection technique, thus increasing confidence on the results obtained. Ten similar runs of the genetic algorithm were performed using the same data. The overall weight of the ith spectral region over the ten runs, w_i, was computed using Eq. 3, where $Q^2_{y_r}$ and C_r are, respectively, the (best) loss function and (best) chromossome obtained for run r.

$$w_i = \sum_{r=1}^{R} Q^2_{y_r} \cdot C_r(i) \qquad \text{Eq. 3}$$

To decide which spectral regions to include in the final model, the average of these weights ($\overline{w}$) was determined and it was found appropriate to select only those regions with weight superior to the average weight.

3. RESULTS

Fig. 1 displays both the raw flour spectra and its Savitzky-Golay first derivative (considering a filter width of 25 data points and a 3rd order polynomial). Spectral pre-processing techniques, such as derivatives and multiplicative scatter correction, are used to reduce spectral interferences, which affect the linear relationships between the spectral and chemical data upon which the model equations are based (Heise and Winzen, 2002).

Interferences may arise from variations in spectral data not directly related with the chemical composition of the sample, e.g., light scattering, noisy measurements, temperature changes. For solid samples, the most significant interferences arise from light scattering in the solid particles. By applying the first derivative to the spectral data, the superimposed absorption peaks in the original spectrum appear as clearly separated downward peaks (see Fig. 1B). In addition, the additive and multiplicative baseline in original spectra becomes constant (Ozaki and Amari, 2000). The lot-to-lot variations in the spectral region from 7400 to 12000 cm^{-1} is very small (see Fig. 1B) and the exclusion of this region from calibration model development was found to improve the model's predictive capability.

For the protein content calibration models, the preprocessing technique used was multiplicative scatter correction (MSC) while for fat content calibration models the Savitzky-Golay 1st derivative was applied to the MSC-corrected spectra prior to calibration model development. The results obtained with PLS regression for three different spectral variable subsets (entire spectral region, 4000 – 7400 cm^{-1} and after variable selection with wGA considering 10 runs of the GA on the 4000 – 7400 cm^{-1} region) are displayed in Table 2 and 3. Fig. 2 displays the spectral regions considered in the development of the two latter models.

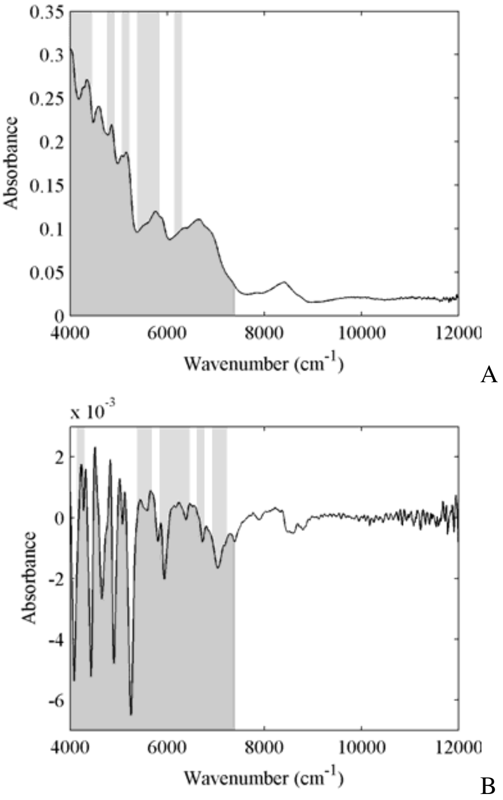

Fig. 2 – Spectral regions used to develop the calibration models; the black line represents the spectra of the soybean flour samples after mean scatter correction; bellow the line: 4000 – 7400 cm^{-1}; above the line: regions selected after application of the wGA. A) Protein content, B) Fat content

Table 2 – Results obtained after developing PLS calibration models for protein content determination

	A	B	C
Number of variables	1039	442	180
Number of LV	7	7	7
Variance captured			
X-block	96.4	99.9	100
Y-block	95.5	96	97.4
RMSECV	0.63	0.43	0.32
Cross validation Q^2_Y	23.3	64.4	80.3

A) entire spectral region
B) 4000–7400 cm^{-1}
C) after variable selection with wGA

Table 3 – Results obtained after developing PLS calibration models for fat content determination

	A	B	C
Number of variables	1039	442	200
Number of LV	8	7	8
Variance captured			
X-block	93.8	99.5	99.4
Y-block	99.6	92.7	99.5
RMSECV	0.35	0.34	0.18
Cross validation Q^2_Y	-48.9	-33.7	63.2

A) entire spectral region
B) 4000–7400 cm^{-1}
C) after variable selection with wGA

For protein content determination, the concentration range was 53 to 56 % protein. It is clear that the variable selection strategy adopted lead to a significant increase in model's predictive ability (RMSECV decreases by half from the model built using the entire spectral range and the one built considering only the GA selected variables).

Concerning the determination of fat content (the range of concentration values is 1.6 to 2.8 % fat), sample #3 was determined to be an outlier (squared prediction error above the 95 % confidence limits) and excluded from the calibration data set. Like for protein content determination, the wGA variable selection strategy adopted lead to a significant increase in model's predictive ability (again, RMSECV decreases by half from the model built using the entire spectral range and the one built considering only the GA selected variables). However, the calibration model for fat content determination should be improved further with the inclusion of additional flour samples in the calibration model.

Fig. 3 displays the regression line for the calibration models developed with the entire variable set while Fig. 4 displays the same regression line after variable selection with wGA. The improvements in model performance are very clear from the comparison of Fig. 3A and 4A (protein content) and Fig. 3B and 4B (fat content).

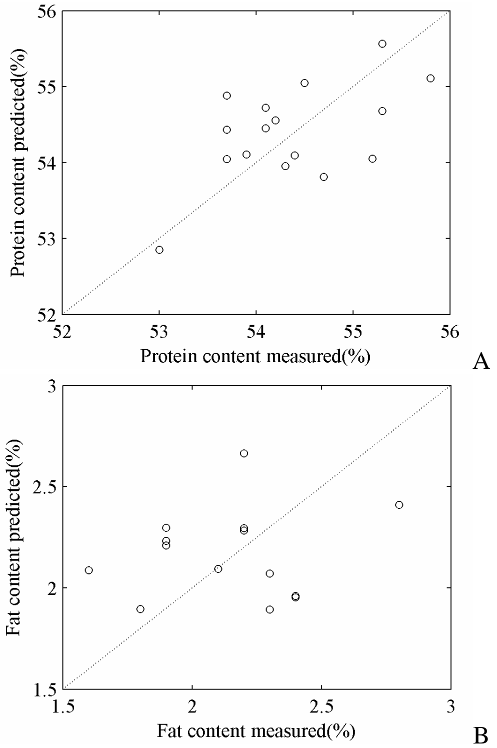

Fig. 3 – Regression line for calibration models developed with the entire variable set. A) Protein content, B) Fat content

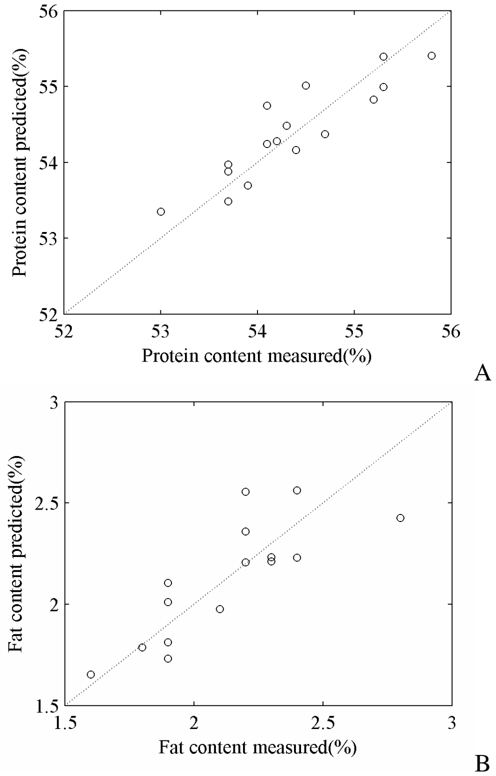

Fig. 4 – Regression line for calibration models developed after variable selection with wGA. A) Protein content, B) Fat content

4. CONCLUSIONS

This work demonstrates the possibility of employing near infrared spectroscopy for the evaluation of fermentation's raw materials quality by characterization of protein and fat content in new lots of soybean flour.

A new approach to the use of genetic algorithms for variable selection was tested with success. The strategy adopted with weighted genetic algorithm (wGA) enabled a 50 % decrease in RMSECV for both protein and fat content determination.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to CIPAN S.A. for supporting this work and for permission to publish the above results. APF gratefully acknowledge financial support from the Portuguese Foundation for Science and Technology (research grants POCI BD/8807/2002, respectively).

REFERENCES

Abrahamsson C, Johansson J, Sparén A, Lindgren F (2003), Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemometrics and Intelligent Laboratory Systems 69:3-12.

Bokobza L (1998), Near infrared spectroscopy, Journal of Near Infrared Spectroscopy 6:3–17.

Cozzolino D, Murray I, Paterson R, Scaife JR (1996), Visible and near infrared reflectance spectroscopy for the determination of moisture, fat and protein in chicken breast and thigh muscle, Journal of Near Infrared Spectroscopy 4: 213–223

Ferreira AP, Alves TP, Menezes JC (2005), Monitoring complex media fermentations with near-infrared spectroscopy: Comparison of different variable selection methods, Biotechnology and Bioengineering, 91: 474 - 481

Heise HM, Winzen R (2002), Fundamental chemometric methods, In: Siesler H, Ozaki Y, Kawata S, Heise HM (Eds.), Near-infrared spectroscopy Principles, instruments, applications, Weinheim (Germany): Wiley-VCH, pp 125-162.

Hong TL, Tsaib SJ, Tsou SCS (1994), Development of a sample set for soya bean calibration of near infrared reflectance spectroscopy, Journal of Near Infrared Spectroscopy 2: 223–227.

Jones AM, Porter MA (1998), Soy Protein in Fermentation, 2nd edition, Cedar Rapids (IA, USA): Cargill Incorporated.

Kays S, Barton F, Windham W (2000), Predicting protein content by near infrared reflectance spectroscopy in diverse cereal food products, Journal of Near Infrared Spectroscopy 8:35–44.

Kirk RS, Sawyer R (1991), Pearson's Composition and Analysis of Foods, 9th edition, Harlow (UK): Longman Scientific & Technical.

Kumagal M, Karube K, Sato T, Ohisa N, Amano T, Kikuchi R, Ogawa N (2002), A near infrared spectroscopic discrimination of noodle flours using a principal-component analysis coupled with chemical information, Analytical Sciences 18: 1145–1150.

Kurowski C, Timm D, Grummisch U, Meyhack U, Grunewald H (1998), The benefits of near infrared analysis for food product quality, Journal of Near Infrared Spectroscopy 6, 343A–348A.

Leardi R, Seasholtz MB, Pell RJ (2002), Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Analytica Chimica Acta 461:189-200.

Leardi R (2000), Application of genetic algorithm-PLS for feature selection in spectral data sets, Journal of Chemometrics 14:643-655.

Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB (2000), Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Applied Spectroscopy 54: 413-419.

Otto M (1999), Chemometrics, Weinheim (Germany): Wiley-VCH.

Ozaki Y, Amari T (2000), Near-infrared spectroscopy in chemical process analysis, In: Chalmers JM (Ed.), Spectroscopy in Process Analysis, Sheffield (UK): Sheffield Academic Press, pp 53-95.

Swierenga H, Wülfert F, de Noord OE, de Weijer AP, Smilde AK and Buydens LMC. 2000. Development of robust calibration models in near-infrared spectrometric applications. Anal Chim Acta 411:121-135.

Tarkosova J, Copikova J (2000), Determination of carbohydrate content in bananas during ripening and storage by near infrared spectroscopy, Journal of Near Infrared Spectroscopy 8: 251–257.

Wise BM, Gallagher NB (1996), The process chemometrics approach to process monitoring and fault detection, Journal of Process Control 6:329-348.

Workman, J (1993), A review of process near infrared spectroscopy: 1980–1994, Journal of Near Infrared Spectroscopy 1:221–245.