http://dx.doi.org/10.5935/0103-5053.20140068

J. Braz. Chem. Soc., Vol. 25, No. 5, 969-974, 2014.
Printed in Brazil - ©2014 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

Non-Destructive NIR Spectrometric Cultivar Discrimination of Castor Seeds Resulting from Breeding Programs

Maria B. H. Santos,[a] Adriano A. Gomes,[a] Welma T. S. Vilar,[a] Pollyne B. A. Almeida,[b] Máira Milani,[b] Márcia B. M. Nóbrega,[b] Everaldo P. Medeiros,[b] Roberto K. H. Galvão[c] and Mário C. U. Araújo*,[a]

[a]Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Departamento de Química, Universidade Federal da Paraíba, CP 37, 58051-970 João Pessoa-PB, Brazil

[b]Laboratório Avançado de Tecnologia Química (LATECQ), EMBRAPA Algodão (CNPA), 58428-095 Campina Grande-PB, Brazil

[c]Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica (ITA), 12228-900 São José dos Campos-SP, Brazil

Este artigo propõe um método de espectrometria no infravermelho próximo (NIR) de reflectância difusa para a discriminação não-destrutiva de sementes de mamona das cultivares mais comumente empregadas nas plantações brasileiras (BRS Nordestina e BRS Paraguaçu). Para esta finalidade, duas técnicas de classificação são comparadas, o SIMCA (modelagem independente flexível por analogias de classe) e PLS-DA (análise discriminante por mínimos quadrados parciais). Ao aplicar a classificação SIMCA a um conjunto de teste contendo 150 sementes, os modelos para as classes BRS Nordestina e BRS Paraguaçu apresentaram valores de sensibilidade/especificidade de 0,91/0,99 e 0,71/1,00, respectivamente. Melhores resultados foram obtidos usando PLS-DA, que classificou corretamente todas as amostras de teste, proporcionando assim valores de sensibilidade e seletividade de 1,00. Estes resultados sugerem que o método proposto é promissor na identificação de genótipos de sementes de mamona, em lotes de sementes ou para fins de reprodução, antes de serem plantadas.

This article proposes a near-infrared (NIR) diffuse reflectance spectrometric method for non-destructive discrimination of castor seeds from the two cultivars most commonly employed in Brazilian plantations (BRS Nordestina and BRS Paraguaçu). For this purpose, two classification techniques are compared, namely SIMCA (soft independent modelling of class analogies) and PLS-DA (partial least squares discriminant analysis). By applying the SIMCA classifier to a test set comprising 150 seeds, the BRS Nordestina and BRS Paraguaçu class models yielded sensitivity/specificity values of 0.91/0.99 and 0.71/1.00, respectively. Better results were obtained by using PLS-DA, which correctly classified all test samples, i.e., yielded sensitivity and specificity values of 1.00. These findings suggest that the proposed method is a promising approach to identify castor seed genotypes, either in seed lots or for breeding purposes, prior to being planted.

Keywords: castor seeds, discrimination of cultivars, near-infrared diffuse reflectance spectrometry, SIMCA, PLS-DA

Introduction

The castor oil plant (*Ricinus communis* L.) is cultivated in several regions of the world. The oil extracted from castor seeds, which is composed of approximately 90% ricinoleic acid (12-hidroxi *cis*-9-octadecanoic), can be used to manufacture a variety of products such as biodiesel, plastics, synthetic fibres, resins, and lubricants.[1,2] In addition, the castor cake obtained from the crushed seeds after extraction of the oil can be used as natural nitrogen fertilizer. Moreover, it can be employed as animal feed after being treated for inactivation of ricin, which is a toxic protein.[3]

Over the years, breeding programs have been used to develop castor cultivars with reduced ricin content, better

*e-mail: laqa@quimica.ufpb.br

productivity, and resistance or adaptation to biotic and non-biotic factors.[4-7] These cultivars are proprietary products which are regulated through plant variety registration laws. Indeed, licensing to individual growers or companies may constitute a source of revenue (royalties) for the developer of the cultivar. For this reason, the unauthorized use of such genetically improved seeds is an infringement of the developer rights. It is also worth noting that the distribution of mixtures of seeds with different genotypes, either intentional or accidental, may compromise a plantation. In fact, the plants will have different characteristics in terms of size, susceptibility to pests and diseases, cycle, maturation, and dehiscence of fruits, among others, which complicates the handling procedures and may lead to a reduction in the production of grain and oil. However, detecting such a problem might be difficult due to the similarity of the seeds, as can be seen in Figure 1.

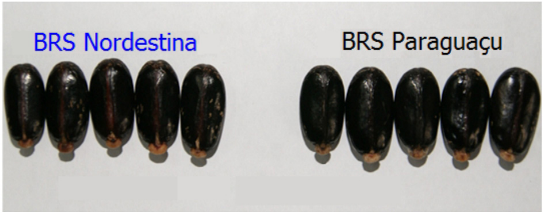

Figure 1. Castor seeds from BRS Nordestina and BRS Paraguaçu cultivars.

As a general rule, the identification of cultivars is carried out by planting the seed and waiting for the germination and development of the plant during at least 30 days until it can be identified on a morphological basis. Techniques based on molecular markers can also be employed,[8-11] but they are destructive, time-consuming and cannot be easily employed for routine identifications. An alternative to circumvent such drawbacks may lie in the use of analytical methods based on near-infrared (NIR) diffuse reflectance spectrometry with appropriate chemometric modelling. Indeed, NIR spectrometry has been successfully used in the classification of a wide variety of agricultural products such as soybean pods,[12] coffee,[13] soybeans[14] and olives.[15] In particular, a recent review[16] pointed out that NIR diffuse reflectance has been applied not only to bulk samples, but also in the quantitative and qualitative analysis of individual seeds. In this context, the potential of this technique for the analysis of single castor seeds has been demonstrated in a study involving the discrimination of two types of seeds with low and high content of oleic acid.[17]

Verification of genetically improved seeds for castor plantations is a problem of special economic and social relevance in Brazil because seed costs are subsidized by the federal government to support small farmers. In this context, the present article proposes a NIR method for non-destructive discrimination of seeds from BRS Nordestina and BRS Paraguaçu cultivars, which are the two most commonly used in Brazil. Despite the visual similarity of the seeds, as seen in Figure 1, the two cultivars have different phenotypic characteristics, which impact the handling of the plantation and the market value of the seeds. For this purpose, two classification techniques are compared, namely SIMCA (soft independent modelling of class analogies)[18,19] and PLS-DA (partial least squares discriminant analysis).[20,21]

SIMCA is a full-spectrum classification method based on the use of Principal component analysis (PCA). A model for each class under consideration is developed on the basis of the principal components (PCs) of a training data set containing only samples from that class. The similarity of an unknown sample with respect to the class is evaluated by projecting the sample onto the PC subspace of the class and comparing the resulting residual variance with the average residual variance of the samples in the training set. For this purpose, two measures are usually employed, namely the leverage (distance to the center of the model within the PC subspace) and the sample-to-model distance (distance to the PC subspace). The classification is carried out by comparing these distances with critical values at a given confidence level.[18,19,22]

PLS-DA is an extension of PLS modelling for use in classification problems. PLS is a multivariate calibration technique which aims to establish a relation between the instrumental response data and some physical or chemical property of the sample. The PLS model is built by using a set of instrumental responses from n samples, recorded over m analytical channels (wavelengths in the case of spectrometric data). These data are usually disposed in the form of a calibration matrix $\mathbf{X}$ ($n \times m$). In addition, the values of the property of interest are arranged in the form of a column vector $\mathbf{y}$ ($n \times 1$). A relation between $\mathbf{X}$ and $\mathbf{y}$ is then derived on the basis of latent variables, which are defined in order to describe the variability in the $\mathbf{X}$ data mostly associated with the variability in the $\mathbf{y}$ data. In the PLS-DA algorithm for binary classification problems, the $\mathbf{y}$ data of the training set correspond to either 0 or 1, depending on the class of the sample.[20,21] After the PLS-DA model is built, the classification of an unknown sample is carried out by calculating the associated y-value and adopting a suitable threshold, usually 0.5, which is the midpoint between 0 and 1.

Experimental

Samples

Three hundred seeds from each cultivar (BRS Nordestina and BRS Paraguaçu) were provided by Embrapa

Algodão (Campina Grande, Paraíba, Brazil). These seeds were harvested in 2009 at the experimental field of Embrapa Algodão in Patos city (Paraíba, Brazil) and were all subjected to the same plantation, harvesting and storage conditions. The seeds were conditioned at 21 °C and 70% relative humidity for at least 1 h before spectral recording. No chemical treatment was employed.

NIR spectrum acquisition

Diffuse reflectance spectra were obtained by using a XDS Rapid Content™ Analyzer VIS-NIR spectrophotometer (Foss Analytical, Hogans, Sweden), fitted with a circular quartz cell with diameter of 3 cm. Each spectrum was acquired as the result of 32 scans in the range 1100-2500 nm with resolution of 0.5 nm. The measurements were repeated four times for each seed. The average of the four spectra thus obtained was employed throughout the work.

Data analysis and software

A second derivative Savitzky-Golay[23] filter with second-order polynomial and 15-point window was applied to the spectra in order to correct for scattering effects. The derivative spectra were used in all the subsequent calculations.

PCA was employed for a preliminary inspection of the overall spectral data set. The Kennard-Stone[24] (KS) algorithm was then applied to the spectra in order to divide the 600 samples into training, validation and test sets with 300, 150 and 150 samples, respectively. Each of these sets comprised equal proportions of BRS Nordestina and BRS Paraguaçu samples.

Savitzky-Golay derivative filtering, PCA, PLS-DA and SIMCA were carried out by using The Unscrambler® 9.7, whereas KS was implemented in Matlab® R2010.

The number of PCs in SIMCA was selected in order to minimize the total number of type I (sample not included in its own class model) and type II (sample included in a wrong class model) errors in the validation set.

The test samples were used as an external set for the final performance assessment of the classifiers. PLS-DA was employed by assigning y-values of 0 and 1 to the BRS Nordestina and BRS Paraguaçu classes, respectively. A threshold of 0.5 was applied to the predicted y-values in order to discriminate the samples in the test set.

Results and Discussion

NIR spectra

As can be seen in Figure 2, the main bands in the NIR spectra of the castor seeds are in the ranges 1400-1500 and 1900-2000 nm. The former is associated to the first OH and NH overtone, as well as CH combination bands, whereas the latter is associated to the second C=O overtone, as well as OH combination bands.[25,26] In addition, several other bands of smaller magnitude are found throughout the spectra. A clear separation between the two classes (BRS Nordestina and BRS Paraguaçu) is observed, but the spectra also display substantial within-class dispersion associated to baseline variations, which can be ascribed to scattering effects.[27] These baseline features were removed by the second derivative procedure, which resulted in the spectra presented in Figure 3. The differences between the two classes are now more subtle compared to Figure 2. Indeed, the effects of scattering, which may have been associated to physical differences in the seed coats, were largely eliminated. Chemometric tools were then employed to determine whether the remaining differences in the derivative spectra are enough to provide an appropriate discrimination between the two types of seeds.

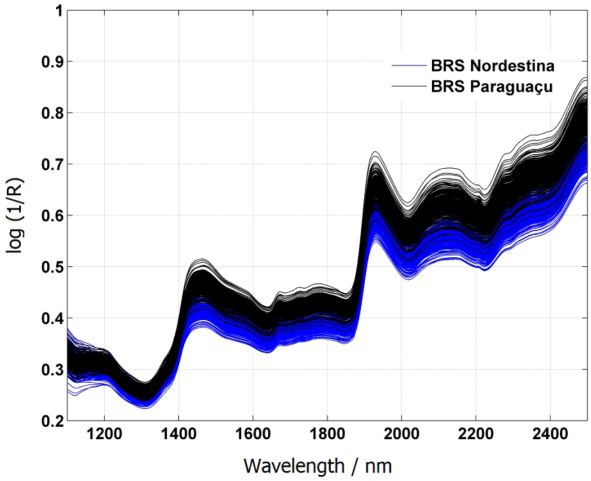

Figure 2. Raw spectra of the 600 castor seeds.

Principal component analysis

Figure 4a presents a score plot of the first *versus* second principal components (PC1 and PC2) obtained by applying PCA to the overall set of derivative spectra. A separation between the two classes can be clearly observed. This finding indicates that the spectral differences between the two seed types, although subtle, are systematic and can be used for discrimination purposes. Moreover, it can be argued that these differences are not restricted to scattering since these physical effects were largely eliminated by the second derivative procedure. It is also worth noting that the first principal component is mainly related to the separation between the classes, which indicates that the

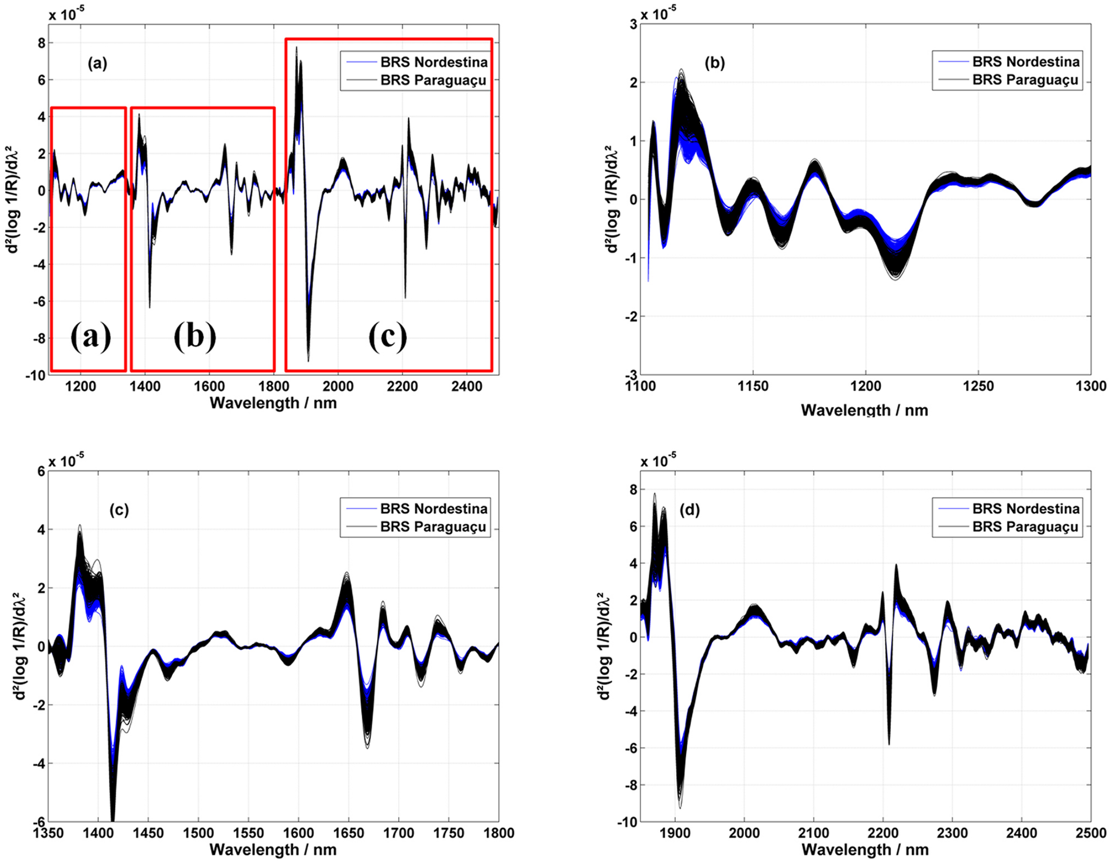

Figure 3. (a) Second derivative spectra of the 600 castor seeds, with expanded views in (b), (c) and (d).

differences between the two seed types are the main source of variability in the spectral data. As seen in Figure 4b, the PC1 loadings have a profile similar to the derivative spectra in Figure 3, i.e., all spectral peaks are present in the PC1 loadings. It can thus be concluded that the entire spectrum contributes to the separation of the classes. This result stands to reason since the major chemical components are the same in both types of seeds, with minor changes in composition.

Assuming that the spectral differences between the two classes are indeed related to chemical features, and not only to scattering effects, a question remains as to whether these features are associated to the interior of the seed or only to the seed coat. In fact, in reflectance mode, the penetration of

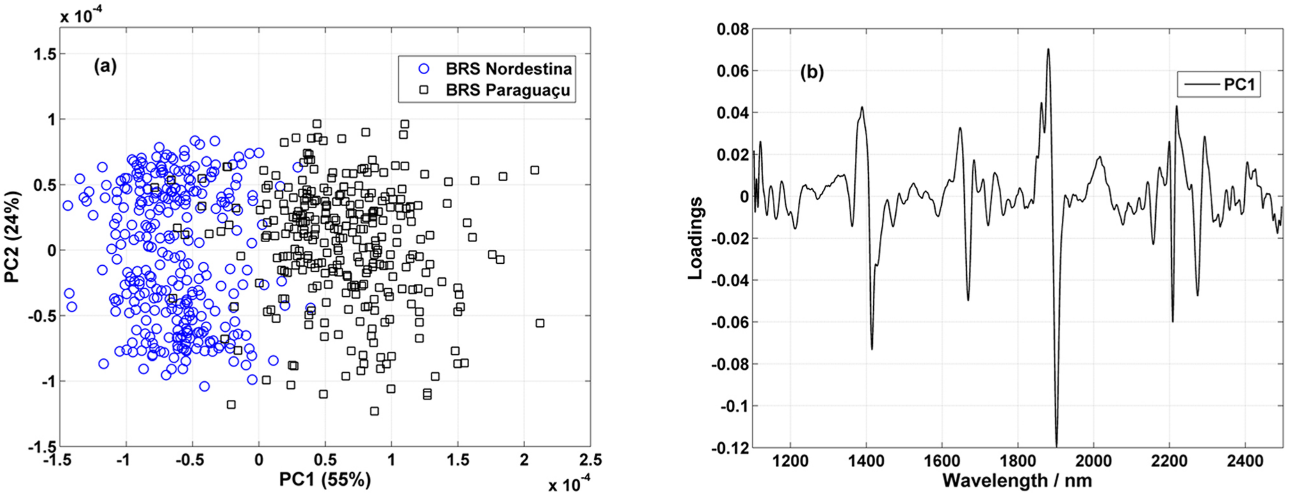

Figure 4. (a) PC1 × PC2 score plot for the 600 samples and (b) plot of PC1 loadings.

the NIR radiation is restricted to a few millimeters into the sample.[16] However, as reported elsewhere,[28] the thickness of the coat in castor seeds is smaller than 0.3 mm, on average. Therefore, it can be argued that the penetration of the NIR radiation is enough to probe the interior of the seed.

SIMCA and PLS-DA classifications

A SIMCA model was built for each of the two classes: BRS Nordestina (one principal component) and BRS Paraguaçu (two principal components). Figures 5a and 5b present the resulting boundaries of the BRS Nordestina and BRS Paraguaçu models, respectively, at the default significance level (5%) of the software package. These boundaries were used in the classification of the 150 test samples, which are also shown in each plot. In the analysis of these results, it is worth noting that SIMCA errors can be of type I (sample not included in its own class model) or II (sample included in a wrong class model). The BRS Nordestina model (Figure 5a) yielded a single type-I error and seven type-II errors. The BRS Paraguaçu model (Figure 5b) yielded 22 type-II errors and no type-I error. These errors can be translated into percent sensitivity/specificity[29] values of 0.91/0.99 for the BRS Nordestina model and 0.71/1.00 for the BRS Paraguaçu model.

A PLS-DA model with five latent variables was chosen as the best compromise between explained variance and number of errors by using the validation set (150 samples), as shown in Figure 6a. As can be seen in Figure 6b, all the 150 test samples (external set) were correctly classified, i.e., the sensitivity and specificity values were both 1.00.

It may be argued that such a better discrimination is achieved because PLS-DA employs samples from both classes in the model-building process. In contrast, the

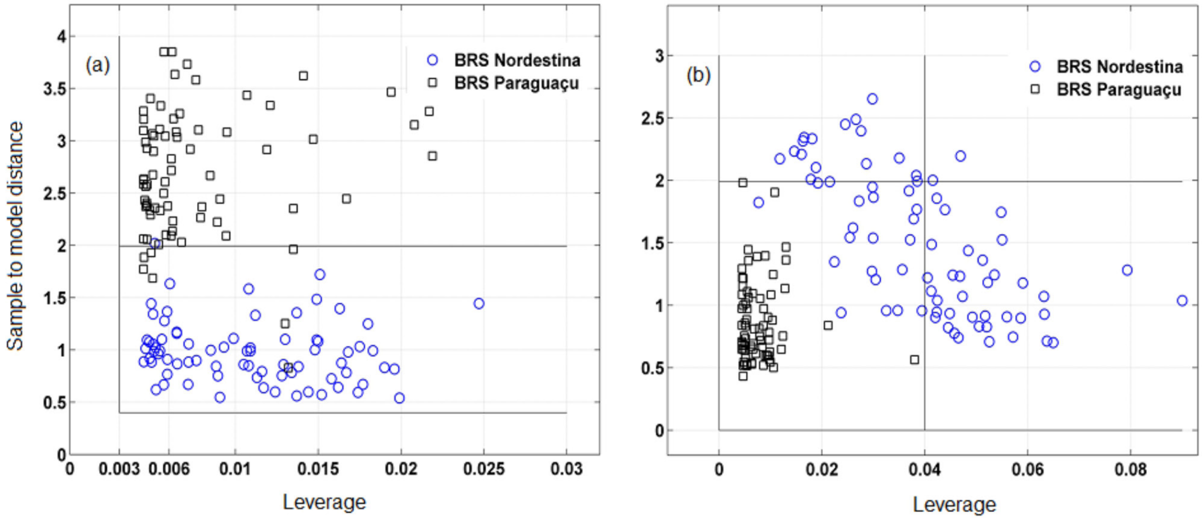

Figure 5. SIMCA boundaries of the (a) BRS Nordestina and (b) BRS Paraguaçu class models.

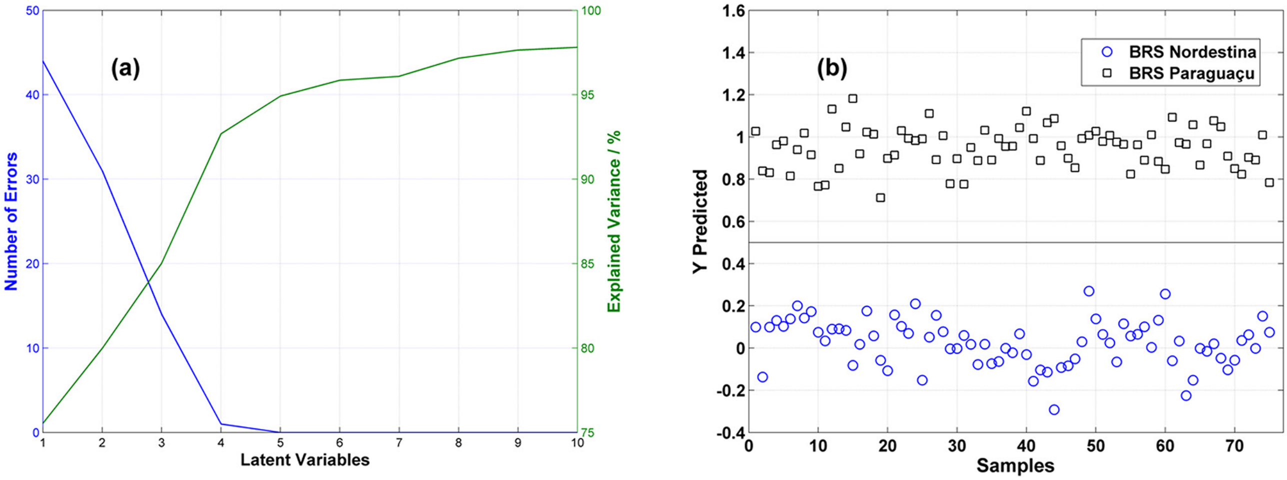

Figure 6. PLS-DA results: a) explained **y**-variance (green line) and number of errors (blue line) *versus* latent variables included by using validation set, and b) **Y** predicted for the test set. The classification threshold is indicated by a horizontal line.

two SIMCA models are built separately, i.e., the BRS Nordestina samples are not considered in the construction of the BRS Paraguaçu model and *vice versa*.

Conclusion

This article proposed a method for non-destructive discrimination of castor seeds from the two cultivars most commonly employed in Brazilian plantations (BRS Nordestina and BRS Paraguaçu). For this purpose, NIR diffuse reflectance spectra were employed to develop SIMCA and PLS-DA classification models. An inspection of the NIR spectra, followed by a PCA investigation, indicated that the spectral measurements can indeed be used to discriminate the two types of seeds. By applying the SIMCA classifier to a test set comprising 150 seeds, the BRS Nordestina and BRS Paraguaçu class models yielded sensitivity/specificity values of 0.91/0.99 and 0.71/1.00, respectively. Better results were obtained by using the PLS-DA model, which correctly classified all test samples, i.e., yielded sensitivity and specificity values of 1.00. These findings suggest that the proposed method is a promising approach to identify castor seed genotypes, either in seed lots or for breeding purposes, prior to being planted.

Acknowledgements

The authors acknowledge the financial support of CAPES (MSc and PhD scholarships), CNPq (Grant 552868/2007-8 and Research Fellowships) and Embrapa (0209070050005).

References

1. Ogunniyi, D. S.; *Bioresour. Technol*. **2006**, *97*, 1086.
2. Scholz, V.; Silva, J. N.; *Biomass Bioenerg.* **2008**, *32*, 95.
3. Godoy, M. G.; Gutarra, M. L. E.; Maciel, F. M.; Felix, S. P.; Bevilaqua, J. V.; Machado, O. L. T.; Freire, D. M. G.; *Enzym. Microb. Technol*. **2009**, *44*, 317.
4. Auld, D. L.; Zanotto, M. D.; Mckeon, T.; Morris, J. B. In *Handbook of Plant Breeding*, 1st ed.; Vollmann J.; Rajcan, I., eds.; Springer: New York, USA, 2009.
5. Anjani, K.; *Ind. Crops Prod.* **2010**, *31*, 139.
6. Pinkerton, S. D.; Rolfe, R. D.; Auld, D. L.; Ghetie, V.; Lauterbach, B. F.; *Crop Sci*. **1999**, *39*, 353.
7. Baldoni, A. B.; Carvalho, M. H.; Sousa, N. L.; Nóbrega, M. B. M.; Milani, M.; Aragão, F. J. L.; *Pesq. Agropec. Bras.* **2011**, *46*, 776.
8. Allan, G.; Williams, A.; Rabinowicz, P. D.; Chan, A. P.; Ravel, J.; Keim, P.; *Genet. Resour. Crop Evol*. **2008**, *55*, 365.
9. Qiu, L. J.; Yang, C.; Tian, B.; Yang, J. B.; Liu, A. Z.; *BMC Plant Biol*. **2010**, *10*, 278.
10. Vasconcelos, S.; Souza, A. A.; Gusmão, C. L. S.; Milani, M.; Benko-Iseppon, A. M.; Brasileiro-Vidal, A. C.; *Micron*. **2010**, *41*, 746.
11. Vasconcelos, S.; Milani, M.; Benko-Iseppon, A. M.; Brasileiro-Vidal, A. C.; *Plant Breeding* **2012**, *9*, 201.
12. Sirisomboon, P.; Hashimoto, Y.; Tanaka, M.; *J. Food Eng*. **2009**, *93*, 502.
13. Esteban-Díez, I.; González-Sáiz, J. M.; Sáenz-González, C.; Pizarro, C.; *Talanta* **2007**, *71*, 221.
14. Lee, J. H.; Choung, M. G.; *Talanta* **2011**, *126*, 368.
15. Casale, M.; Zunin, P.; Cosulich, M. E.; Pistarino, E.; Perego, P.; Lanteri, S.; *Food Chem*. **2010**, *122*, 1261.
16. Agelet, L. A.; Hurburgh Jr., C. R.; *Talanta* **2014**, *121*, 299.
17. Fernández-Cuesta, A.; Fernández-Martínez, J. M.; Velasco, L.; *J. Am. Oil Chem. Soc.* **2012**, *89*, 435.
18. Wold, S.; *Pattern Recognit*. **1976**, *8*, 127.
19. Beebe, K. R.; Pell, R. J.; Seasholtz, B.; *Chemometrics - A Practical Guide*, 4th ed.; Wiley: New York, USA, 1998.
20. Silva, A. C.; Pontes, L. F. B. L.; Pimentel, M. F.; Pontes, M. J. C.; *Talanta* **2012**, *93*, 129.
21. Pomerantsev, A. L.; Rodionova, O. Y.; *J. Chemom*. **2012**, *26*, 310.
22. Almeida, M. R.; Correia, D. N.; Rocha, W. F. C.; Scafi, F. J. O.; Poppi, R. J.; *Microchem. J.* **2013**, *109*, 170.
23. Savitzky, A.; Golay, M. J. E.; *Anal. Chem*. **1964**, *36*, 1627.
24. Kennard, R. W.; Stone, L. A.; *Technometrics* **1969**, *11*, 137.
25. Xiaoboa, Z.; Jiewena, Z.; Poveyb, M. J. W.; Holmesb, M.; Hanpin, M.; *Anal. Chim. Acta*. **2010**, *667*, 14.
26. Rinnan, A.; Berg, F. V.; Engelsen, S.; *Anal. Chem.* **2010**, *28*, 1221.
27. Workman, J. J.; Weyer, L.; *Practical Guide to Interpretive Near-Infrared Spectroscopy*; CRC Press, Taylor & Francis Group: Boca Raton, USA, 2008.
28. Severino, L. S.; Auld, D. L.; Baldanzi, M.; Cândido, M. J. D.; Chen, G.; Crosby, W.; Tan, D.; Lakshmamma, X. H. P.; Lavanya, C.; Machado, O. L. T.; Mielke, T.; Milani, M.; Miller, T. D.; Morris, J. B.; Morse, S. A.; Navas, A. A.; Soares, D. J.; Sofiatti, V.; Wang, M. L.; Zanotto, M. D.; Zieler, H.; *Agron. J.* **2012**, *104*, 880.
29. Galvão, R. K. H.; Araújo, M. C. U. In *Compreehensive Chemometrics*, vol. 3; Walczak, B.; Tauler, R.; Brown, S., eds.; Elsevier: Oxford, 2009, p. 233.

Submitted: November 20, 2013
Published on line: April 4, 2014

FAPESP has sponsored the publication of this article.