2014

Classification of the waxy condition of durum wheat by near infrared reflectance spectroscopy using wavelets and a genetic algorithm

Barry K. Lavine
Department of Chemistry, Oklahoma State University

Nikhil Mirjankar
Department of Chemistry, Oklahoma State University

Stephen Delwiche
USDA-ARS, Beltsville Agricultural Research Center, Food Quality Laboratory

Microchemical Journal 117 (2014) 178–182

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Classification of the waxy condition of durum wheat by near infrared reflectance spectroscopy using wavelets and a genetic algorithm

Barry K. Lavine [a,*], Nikhil Mirjankar [a], Stephen Delwiche [b]

[a] Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, United States
[b] USDA-ARS, Beltsville Agricultural Research Center, Food Quality Laboratory, Building 303, BARC-East, Beltsville, MD 20705-2350, United States

ARTICLE INFO

Article history:
Received 24 June 2014
Accepted 27 June 2014
Available online 4 July 2014

Keywords:
'Waxy' wheat
Wavelets
Near infrared spectroscopy
Genetic algorithm
Pattern recognition
Feature selection

ABSTRACT

Near infrared (NIR) reflectance spectroscopy has been applied to the problem of differentiating four genotypes of durum wheat: 'waxy', *Wx A1 null* null, *wx-B1* null and wild type. The test data consisted of 95 NIR reflectance spectra of wheat samples obtained from a USDA-ARS wheat breeding program. A two-step procedure for pattern recognition analysis of NIR spectral data was employed. First, the wavelet packet transform [14,15] was applied to the NIR reflectance data using wavelet filters at different scales to extract and separate low-frequency signal components from high frequency noise components. By applying these filters, each reflectance spectrum was decomposed into wavelet coefficients that represented the sample's constituent frequencies. Second, wavelet coefficients characteristic of the waxy condition of the wheat samples were identified using a genetic algorithm for pattern recognition. The pattern recognition GA employed both supervised and unsupervised learning to identify wavelet coefficients that optimized clustering of the spectra by genotype in a plot of the two largest principal components of the data. By sampling key feature subsets, scoring their PC plots, and tracking those genotypes and samples that were difficult to classify, the pattern recognition GA was able to identify a set of wavelet coefficients whose PC plot showed clustering of the wheat samples on the basis of their 'waxy' condition. Object validation was also performed to assess the predictive ability of the proposed NIR method to identify the 'waxy' condition of the wheat. An overall classification success rate of 78% was achieved for the spectral data.

1. Introduction

Since the mid-1990s there has been renewed interest in the breeding of 'waxy' and 'partial waxy' wheat in the United States and elsewhere [1] due to the special properties of its starches. The 'waxy' condition of wheat is related to its amylose content. An enzyme called granule bound starch synthase (GBSS), also known as 'waxy' protein, is primarily responsible for amylose in wheat [2]. The absence of GBSS gives rise to low (near zero) levels of amylose and high levels of the starch's complementary component, amylopectin, often referred to as the 'waxy' condition. Active isoforms of GBSS under natural conditions are encoded at two genetic loci, *Wx A1 null* and *Wx-B1*, for tetraploid (i.e., durum) wheat and by three loci, *Wx-A1*, *Wx-B1*, and *Wx-D1*, for hexaploid (common) wheat. By comparison, wheat in the native wild-type state possesses all GBSS isoforms. The 'partial waxy' condition in a wheat line occurs through natural mutation or conventional breeding practices, when at least one (but not all) of the waxy genes is a null allele. Applications of 'waxy' and 'partial waxy' wheat include the development of stock flour material for blending by millers, flour for Asian noodle-making, and the use of 'waxy' and 'partial waxy' wheat as a substitute for waxy maize starch in the paper and adhesive industries [3–6].

Due to the growing demand for waxy and partial waxy wheat, there is a need to develop a reliable and rapid test to authenticate the waxy condition. Identification of waxy seeds is currently restricted to the use of wet chemical techniques such as iodine-binding blue complex colorimetry to measure amylose content. However, this procedure is time consuming, not suitable for commercial grades of wheat with a narrow range of amylose content, and often does not yield definitive results for the identification of partial waxy lines [7]. Rather than applying chemical methods to determine the amylose content of wheat, characterizing waxy and partial waxy wheat lines according to the number of active GBSS genes by detection of the different GBSS isoforms through protein analysis has been formulated using SDS-PAGE [8], ELISA [9], or multiplex PCR techniques [10]. However, these methods are expensive, complex, time consuming and not amenable to either wheat breeding programs or to the various stages of wheat marketing and production.

Near-infrared (NIR) reflection spectroscopy is a simple, fast, and inexpensive method that is well documented and widely used for the determination of protein, moisture content, and other properties of cereals at grain production facilities. Previous efforts to characterize wheat genotypes of ground meal and whole kernel samples using NIR reflectance spectroscopy were not successful [11,12]. These studies were performed using either principal component analysis or linear

* Corresponding author.

http://dx.doi.org/10.1016/j.microc.2014.06.030
0026-265X/

discriminant analysis to analyze the NIR reflectance data. Although classification models developed from the NIR data were able to recognize the waxy genotype, the three other genotypes (*Wx A1 null* null, *wx-B1* null, and wild type) could not be identified. The low classification success rate for wheat varieties obtained in these studies was approximately 50%, which can be directly attributed to the inability of the discriminants to identify partial waxy lines.

In this study, the wavelet packet transform was applied to NIR reflectance spectra, followed by the use of a genetic algorithm for pattern recognition analysis to select informative wavelet coefficients that can be used to characterize NIR spectra according to the four genotypes: waxy, *Wx A1 null* null, *wx-B1* null and wild type. The objective of this study was to evaluate the feasibility of NIR reflectance spectroscopy to genotype wheat samples. The confounding of chemical information with the expression level of the genes was also investigated by analyzing the wavelet coefficients selected by the pattern recognition GA for correlation with both amylose and protein content.

2. Experimental

Ninety-five wheat samples from four genotypes obtained from a USDA-ARS Nebraska wheat breeding program were available for this study (see Table 1). The amylose content of each wheat sample was measured by colorimetry of the iodine-binding complex [13], and the number and type of active GBSS genes in each wheat sample were determined by SDS-PAGE. Each wheat sample was separately ground on a laboratory scale cyclone grinder (Udy Corp, Fort Collins, CO, USA). The ground meal was placed in a standard ring cell loaded into a reflectance NIR spectrometer (Foss-NIR System Model 6500) equipped with a rotating sample attachment. For each wheat sample, an average spectrum was obtained from duplicate successive spectra (wavelength range from 1100 to 2498 nm) run at 32 scans/spectrum and at 2 nm resolution. Fig. 1 shows an average NIR spectrum of a 'waxy' wheat sample. Further details regarding the preparation of the wheat samples and the collection of the NIR data can be found elsewhere [11].

3. Pattern recognition analysis

A two-step procedure for pattern recognition analysis of NIR spectral data was employed. First, the wavelet packet transform [13,14] was applied to the NIR reflectance data using wavelet filters at different scales to extract and separate low-frequency signal components from high frequency noise components. By applying these filters, each reflectance spectrum was decomposed into wavelet coefficients that represented the sample's constituent frequencies. In this study, the Daubechies 4 mother wavelet at the 8th level of decomposition, i.e., 8db4, was used to denoise and to resolve overlapping spectral responses. Wavelet preprocessing of the spectral data produced 9494 nonzero wavelet coefficients for each NIR spectrum.

Second, wavelet coefficients characteristic of the waxy condition of the samples were identified using a genetic algorithm for classification and feature selection [16–21]. The pattern recognition GA utilized both supervised and unsupervised learning to identify coefficients that optimized clustering of the spectra by class (i.e., the waxy condition of the wheat samples) in a plot of the two or three largest principal components of the data. Since principal components maximize variance,

Table 1
Wheat data set.

Genotype	Number of NIR spectra
Waxy	24
Wx-A1	25
Wx-B1	24
Wild type	22
All	95

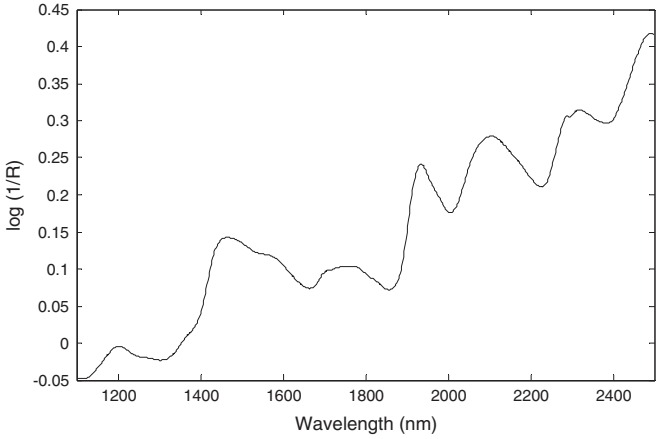

Fig. 1. Typical spectrum of waxy wheat obtained by NIR diffused reflectance spectroscopy.

the bulk of the information encoded by the selected wavelet coefficients was about differences between wheat cultivars in the study. The principal component analysis routine embedded in the fitness function of the pattern recognition GA served as an information filter, significantly reducing the size of the search space as it restricted the search to wavelet coefficients whose principal component plots showed clustering of the wheat samples on the basis of their waxy condition. In addition, the pattern recognition GA focused on wheat samples and or specific GBSS isoforms (i.e., classes) that were difficult to classify as it trained by adjusting (i.e., boosting) both the sample and class weights. Samples or classes that were always correctly classified were not as heavily weighted as samples or classes that were difficult to classify. Over time, the algorithm learned its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrated aspects of artificial intelligence and evolutionary computations to yield a smart one pass procedure for feature selection, classification and prediction.

For pattern recognition analysis, each NIR reflectance spectrum was initially represented as a data vector, $x = (x_1, x_2, x_3 \ldots x_j \ldots x_{700})$ where x_j is the absorbance of the jth point of the NIR reflectance spectrum. Each NIR spectrum was normalized to unit length to correct for differences that exist in the optical path length among the wheat samples. All spectral features (including wavelet coefficients) were autoscaled to remove any inadvertent weighing of the features that otherwise would occur due to differences in magnitude among these variables in the data set.

4. Results and discussion

Fig. 2 shows a plot of the two largest principal components of the 95 wheat samples and 700 points comprising the original spectral data. Each spectrum (i.e., sample) is represented as a data point in the principal component (PC) plot (1 = waxy type, 2 = *wx A1 null*, 3 = *wx-B1* null, and 4 = wild type). The overlap of NIR spectra of the four genotypes in the PC plot of the data is evident. One sample in the plot was identified as an outlier and was deleted from the analysis as its spectrum was very different from the other NIR spectra in the data set.

The next step in this study was feature selection. The pattern recognition GA identified features by sampling key feature subsets, scoring their PC plots, and tracking those classes and samples that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 300 generations, the pattern recognition GA identified 6 spectral features (see Fig. 3). There is some indication from this PC plot that the waxy genotypes can be identified from the other wheat genotypes. As the original spectral features do not contain sufficient information for wheat genotyping, further preprocessing of the original spectral data was necessary.

For this reason, the second derivative was applied to each NIR spectrum using a 7-point quadratic polynomial Savitzky–Golay filter.

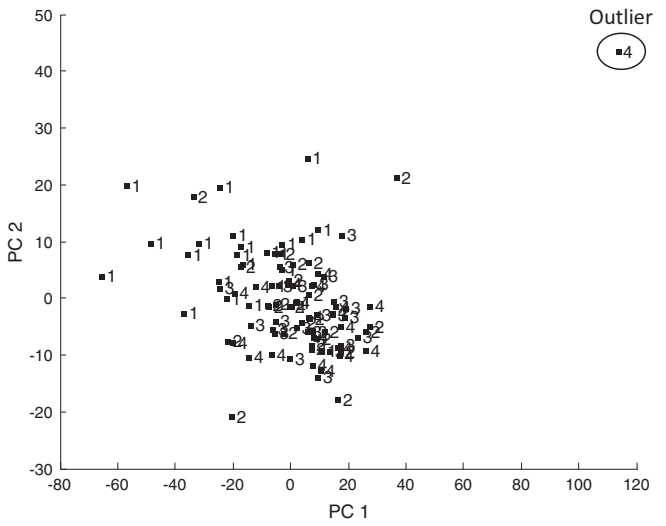

Fig. 2. Plot of the two largest principal components of the 95 NIR spectra and 700 points that comprise the wheat data set. Each NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *Wx A1 null* null, 3 = *wx-B1* null, and 4 = wild type).

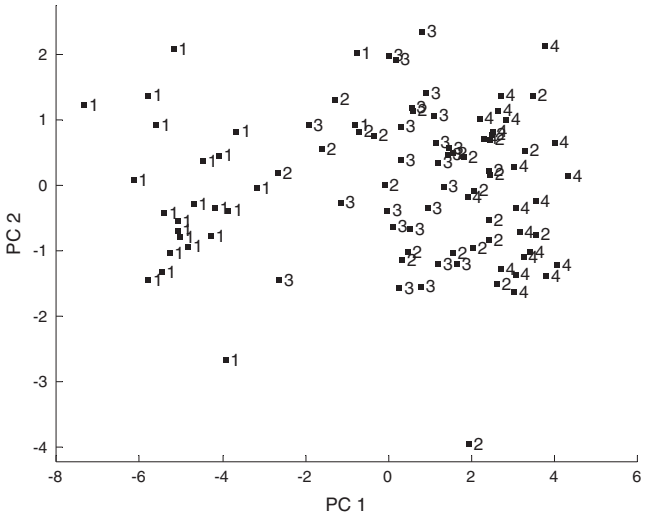

Fig. 4. Plot of the two largest principal components of the 94 s derivative spectra and 17 features identified by the pattern recognition GA. Each second derivative spectrum is represented as a point in the plot (1 = waxy type, 2 = *Wx A1 null* null, 3 = *wx-B1* null, and 4 = wild type).

The second derivative function eliminated offsets and sloping baselines, as well as resolving overlapping spectral bands. Fig. 4 shows a plot of the two largest principal components of the 94 s derivative NIR spectra and the 17 spectral features identified by the pattern recognition GA. Although there is clear separation of waxy wheat from the other wheat cultivars in the PC plot, wild type wheat and the two genotypes corresponding to the partial waxy condition overlapped.

Clearly, even more powerful spectral preprocessing methods were needed to denoise and deconvolve the spectral bands of these samples and to resolve information related to wheat genotype. For this reason, the wavelet packet transform was applied to the spectra. The Daubechies 4 wavelet was selected as other members of the Daubechies family had either sharper or broader features compared to bands in the NIR spectra. The Daubechies 4 mother wavelet at the 8th level of decomposition was used to denoise and deconvolute each NIR spectrum into wavelet coefficients. Wavelet decomposition at the 4th or 6th level was unable to provide sufficient resolution of the signal in the data with respect to information about genotype.

To identify the informative wavelet coefficients, the pattern recognition GA was applied to the spectral data. The pattern recognition GA identified wavelet coefficients correlated to genotype by sampling key feature subsets, scoring their PC plots, and tracking wheat samples and/or genotypes difficult to classify. The boosting routine used this information to steer the population to an optimal solution. Fig. 5 shows a plot of the two largest principal components of the 94 NIR spectra and the 55 wavelet coefficients identified by the pattern recognition GA. Separation of NIR spectra by wheat genotype is evident in the PC plot of these 55 wavelet coefficients.

To assess the information content of these 55 wavelet coefficients, partial least squares (PLS) regression was performed for both amylose and protein contents using the wavelet coefficients as independent variables in the regression analysis [22,23]. Table 2 summarizes the results of the PLS analysis using a three-component model. The standard error of calibration (SEC), the range of amylose and protein content

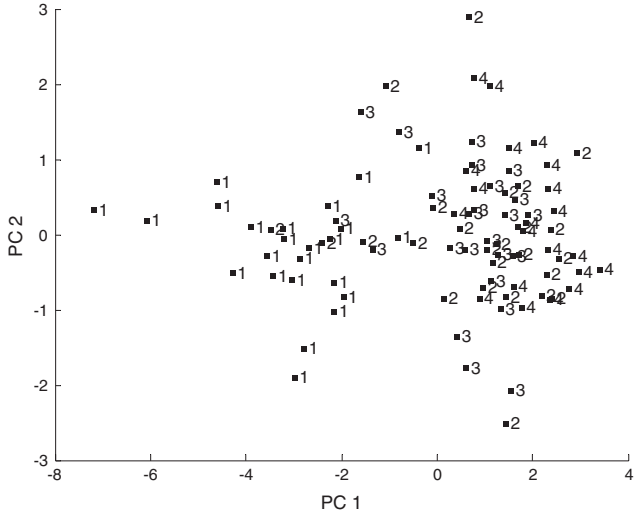

Fig. 3. Plot of the two largest principal components of the 94 NIR spectra and 6 wavelengths identified by the pattern recognition GA. Each NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *Wx A1 null* null, 3 = *wx-B1* null, and 4 = wild type).

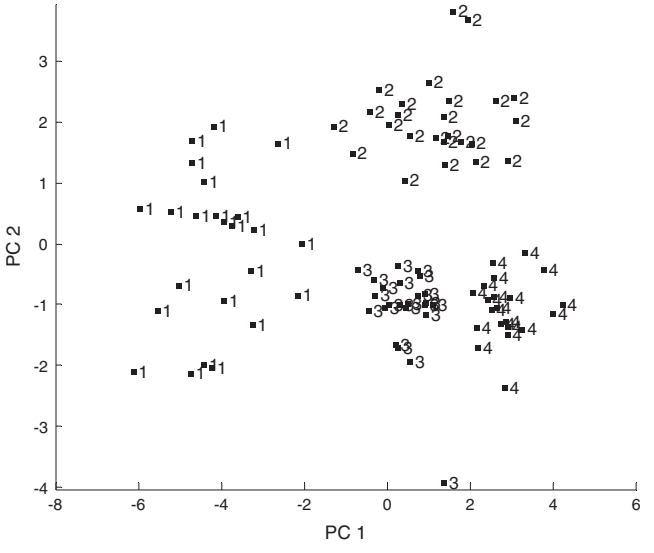

Fig. 5. Plot of the two largest principal components of the 94 wavelet transformed NIR spectra and 55 wavelet coefficients identified by the pattern recognition GA. Each wavelet transformed NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *Wx A1 null* null, 3 = *wx-B1* null, and 4 = wild type).

Table 2
PLS results.

Y-block	PLS components	SEC (%)	Correlation
Amylose (14.6%–21.2%)	3	3.47	0.94
Protein (1.8%–31.9%)	3	0.38	0.94

spanned by the 94 wheat samples and the correlation coefficients for both the amylose and protein calibrations are listed in Table 2. From this table, it is evident that information about amylose content and protein content is also present in the 55 wavelet coefficients identified by the pattern recognition GA. Bar graphs of the mean amylose and mean protein content of each genotype with the standard deviation in parentheses are shown in Fig. 6. From Table 2, and Figs. 5 and 6, one can conclude that classification of the wheat samples by genotype is not influenced by either the protein or amylose content of the wheat.

The predictive ability of the NIR method for genotyping wheat was assessed using a procedure known as object validation. The 94 NIR spectra were initially divided into a training set of 86 wheat samples and a prediction set of 8 samples. Table 3 summarizes the genotypes of the samples comprising the prediction set. Spectra that comprised the prediction set were chosen by random lot. During the course of the GA runs, it was discovered that two training set samples were outliers. Their removal from the training set allowed the pattern recognition

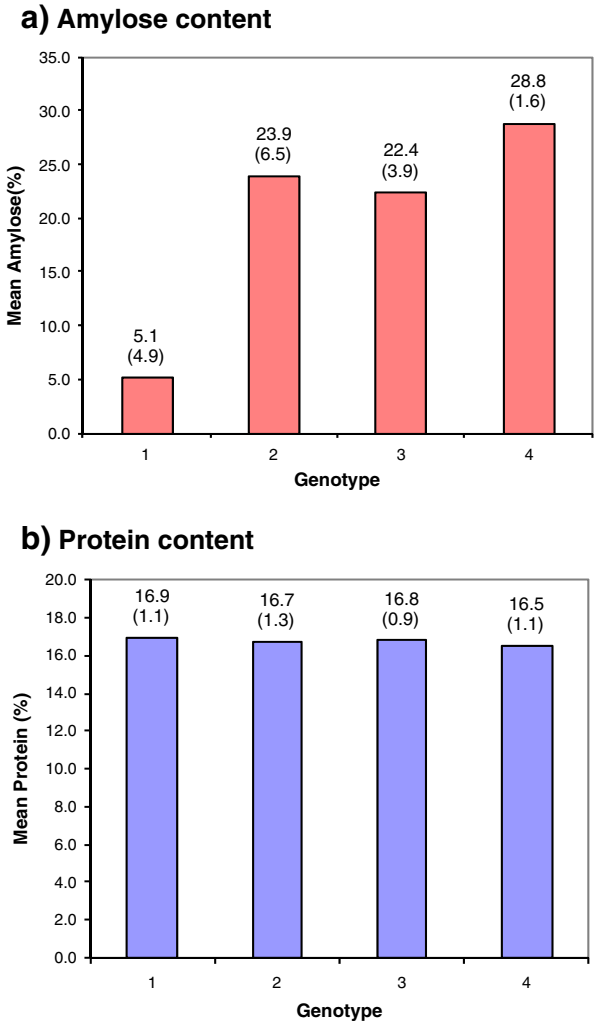

Fig. 6. Plot of (a) amylose and (b) protein contents (mean %, standard deviation) for each genotype in the wheat data set (1 = waxy type, 2 = *Wx A1 null* null, 3 = *wx-B1* null, and 4 = wild type).

Table 3
Prediction set.

Genotype	Number of NIR spectra
Waxy (1)	2
Wx A1 null (2)	2
Wx-B1 (3)	2
Wild type (4)	2

GA to converge towards a solution. These two training set samples were waxy but their amylose content was approximately 13% which was found to be substantially higher than the other waxy wheat samples whose amylose content was less than 5%. For this reason, these two samples were deleted from the study.

The pattern recognition GA was applied to the 84 wavelet preprocessed spectra that comprised the training set. By sampling key feature subsets, scoring their PC plots, and tracking those genotypes and samples that were difficult to classify, the pattern recognition GA was able to identify 32 wavelet coefficients whose PC plot (see Fig. 7) showed clustering of the training set samples on the basis of genotype. To assess the predictive ability of the 32 wavelet coefficients identified by the pattern recognition GA, the 8 spectra from the prediction set were employed. These 8 NIR spectra were directly mapped onto the principal component plot defined by the 84 spectra of the training set and the 32 wavelet coefficients identified by the pattern recognition GA. Fig. 7 shows the prediction set samples projected onto the principal component map developed from the training set data. All but two wheat samples (one *wx A1* null and one *wx-B1* null) were projected in a region of the plot near wheat samples with the same class label (i.e., genotype).

To assess the performance of the NIR method, additional object validation studies were undertaken. 23 training set/prediction set pairs were generated by random selection where each training set consisted of 88 samples (i.e., wavelet transformed NIR spectra) and the prediction set consisting of 4 samples. Each wheat sample was present in only one of the 23 prediction sets generated. For each training set, wavelet coefficients whose PC plot showed clustering on the basis of wheat genotype were identified by the pattern recognition GA. The ability of these wavelet coefficients to classify wheat samples (i.e., NIR spectra) by genotype

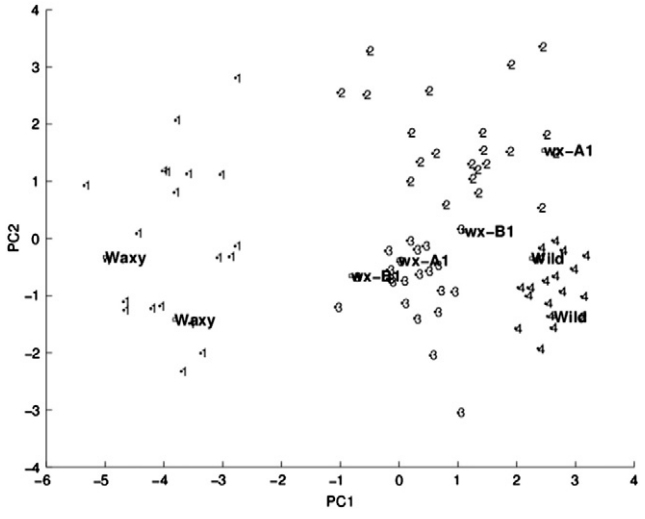

Fig. 7. Projection of the 8 prediction set samples onto the PC plot of the 84 wavelet transformed NIR spectra and the 32 wavelet coefficients identified by the pattern recognition GA. Each wavelet transformed NIR spectrum in the training set (1 = waxy type, 2 = *Wx A1 null* null, 3 = *wx-B1* null, and 4 = wild type) and prediction set (Waxy = waxy type, *Wx A1 null* = *Wx A1 null* null, wx-B1 = *wx-B1* null, and Wild = wild type) is represented as a point in the PC plot.

Table 4
Summary of object validation results.

Genotype	Number of samples	Number of correct classifications	Classification success rate (%)
Waxy (1)	22	22	100.0
Wx A1 null (2)	25	12	48.0
Wx-B1 (3)	24	20	83.3
Wild type (4)	21	18	85.7
Total	92	72	78.2

was assessed using the corresponding prediction set. A summary of the validation set results tabulated for all 23 training set/prediction set pairs is given in Table 4. Misclassified wild types were assigned as either *wx A1 null* or *wx-B1* null. Misclassified *wx-B1* null samples were assigned to *wx A1 null*, and misclassified *wx A1 null* samples were assigned to *wx-B1* null.

In a previous study [12] a leave out one object validation method was used to determine the overall recognition rate of discriminants developed from NIR spectral data to recognize the waxy condition of wheat. The fully waxy genotype was reliably classified, whereas the classification success rate among the other three genotypes was typically very low, less than 50%.

The object validation procedure used in this study differed from the procedure used in other studies. For each training set, a different set of wavelet coefficients is selected by the pattern recognition GA. In a typical object validation study, the same features are used for each training set and are identified using the entire data set prior to dividing the data into training set/prediction set pairs. For this reason, the technique of object validation generally provides overly optimistic estimates of the classification success rate. In this study, the validation set samples do not influence the wavelet coefficients selected for each training set by the pattern recognition GA. Hence, the classification success rate reported for this data is an accurate assessment of the overall recognition rate of the proposed NIR method.

5. Conclusions

These results suggest that NIR reflectance spectroscopy has the potential to predict the waxy condition of wheat for single tetraploid (durum) wheat kernels. Using wavelets to preprocess the NIR spectra and the pattern recognition GA to select informative wavelet coefficients, it has been demonstrated that wheat samples could be differentiated by their waxy condition with a recognition rate of 78% versus that of 50% which was reported in a previous study [11]. Amylose and protein contents of the wheat samples were not shown to be significant covariates that would confound the classification of the NIR spectra by genotype.

Acknowledgments

The authors thank R.A. Graybosch (USDA-ARS at University of Nebraska at Lincoln) for providing the wheat samples and B. Stetzler (formerly USDA-ARS, Beltsville) for collecting the NIR reflectance spectral data.

References

[1] R.A. Graybosch, E.J. Souza, W.A. Berzonsky, P.S. Baezinger, D.J. McVey, O.K. Chung, Registration of nineteen waxy spring wheats, Crop Sci. 44 (2004) 1491–1492.
[2] T. Nakamura, M. Yamamori, Production of waxy (amylose-free) wheats, Mol. Gen. Genet. 248 (1995) 253–259.
[3] R.A. Graybosch, Waxy wheats: origin, properties, and prospects, Trends Food Sci. Technol. 9 (1998) 135–142.
[4] J. Reddy, P.A. Seib, Modified waxy wheat starch compared to modified waxy corn starch, J. Cereal Sci. 31 (2000) 25–29.
[5] T. Hoshino, R. Yoshikawa, S. Ito, K. Hatta, T. Nakamura, M. Yamamori, K. Hayakawa, K. Tanaka, H. Akashi, S. Endo, S. Tago, and S. Ishigami, Flour Blends for Breads, Cakes, or Noodles, and Foods Prepared from the Flour Blends, United States Patent, (2000) 6,042,867.
[6] J. Epstein, C.F. Morris, K.C. Huber, Instrumental texture of white salted noodles prepared from recombinant inbred lines of wheat differing in the three granule bound starch synthase (waxy) genes, J. Cereal Sci. 35 (2002) 51–63.
[7] R.A. Graybosch, C.J. Peterson, L.E. Hansen, S. Rahman, A. Hill, J.H. Skerritt, Identification and characterization of U.S. wheats carrying null alleles at the wx loci, Cereal Chem. 75 (1998) 162–165.
[8] X.C. Zhao, P.J. Sharp, An improved 1-D SDS-PAGE method for the identification of three bread wheat 'waxy' proteins, J. Cereal Sci. 23 (1996) 191–193.
[9] S. Rahman, B. Kosarhashemi, M.S. Samuel, A. Hill, D.C. Abbott, J.H. Skerritt, J. Preiss, R. Appels, M.K. Morell, The major proteins of wheat endosperm starch granules, Aust. J. Plant Physiol. 22 (1995) 793–803.
[10] T. Nakamura, P. Vrinten, M. Saito, M. Konda, Rapid classification of partial waxy wheats using PCR-based markers, Genome 45 (2002) 1150–1156.
[11] S.R. Delwiche, R.A. Graybosch, Identification of waxy wheat by near-infrared reflectance spectroscopy, J. Cereal Sci. 35 (2002) 29–38.
[12] S.R. Delwiche, R.A. Graybosch, L.E. Hansen, E. Souza, F.E. Dowell, Single kernel near infrared analysis of tetraploid (durum) wheat for classification of the waxy condition, Cereal Chem. 83 (2006) 287–292.
[13] C.A. Knutson, M.J. Grove, Rapid method for estimation of amylose in maize starches, Cereal Chem. 71 (1994) 469–471.
[14] B.B. Hubbard, The World According to Wavelets, 2nd edition A.K. Peters, Natick, MA, 1998.
[15] F. Chau, Y. Liang, J. Gao, X. Shao, Chemometrics — From Basics to Wavelet Transform, Wiley-Interscience, Hoboken, NJ, 2004.
[16] B.K. Lavine, A. Fasasi, N. Mirjankar, C. White, Search prefilters for library matching of infrared spectra in the PDQ database using the autocorrelation transformation, Microchem. J. 113 (2014) 30–35.
[17] B.K. Lavine, K. Nuguru, N. Mirjankar, J. Workman, Pattern recognition assisted infrared library searching, Appl. Spectrosc. 66 (2012) 917–925.
[18] B.K. Lavine, N. Mirjankar, R. LeBouf, A. Rossner, Prediction of mold contamination from microbial volatile organic compound profiles using head space gas chromatography/mass spectrometry, Microchem. J. 103 (2012) 119–124.
[19] B.K. Lavine, C. White, C. Matthew Sundling, C. Breneman, Odor-structure relationship studies of tetralin and indan musks, Chem. Senses 37 (2012) 723–736.
[20] B.K. Lavine, N. Mirjankar, S. Ryland, M. Sandercock, Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics, Talanta 87 (2011) 46–52.
[21] B.K. Lavine, K. Nuguru, N. Mirjankar, One stop shopping — feature selection, classification, and prediction in a single step, J. Chemometr. 25 (2011) 116–129.
[22] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, NY, 1989.
[23] P. Geladi, B.R. Kowalski, Partial least squares regression — a tutorial, Anal. Chim. Acta. 186 (1986) 1–17.