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

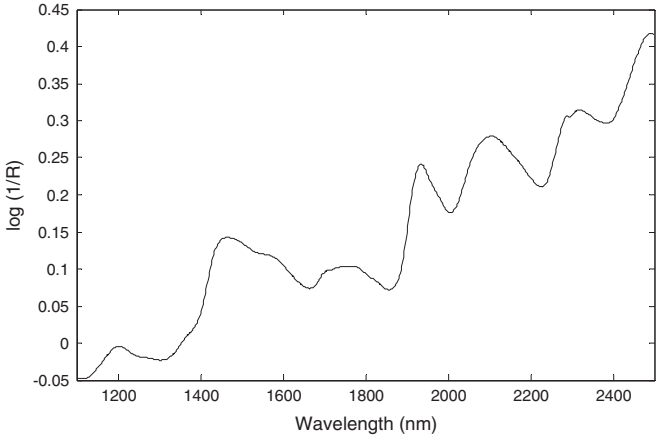

Fig. 1. Typical spectrum of waxy wheat obtained by NIR diffused reflectance spectroscopy.

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.