Application of Near Infrared Reflectance Spectroscopy to Determine Saccharification Efficiency of Corn Biomass

Sonia Pereira-Crespo
Universidad de Vigo: Universidade de Vigo

Noemi Gesteiro
CSIC: Consejo Superior de Investigaciones Cientificas

Ana López-Malvar
University of Vigo: Universidade de Vigo

Leonardo Gómez
University of York

Rogelio Santiago (rsantiago@mbg.csic.es)
CSIC: Consejo Superior de Investigaciones Cientificas

Research Article

Keywords: Zea mays, NIRS, prediction models, bioethanol

Posted Date: November 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3495770/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Nowadays, in the bioethanol production process, improving the simplicity and yield of cell wall saccharification procedure represent the main technical hurdles to overcome. This work evaluated the application of a rapid and cost-effective technology such as Near Infrared Reflectance Spectroscopy (NIRS) for predicting saccharification efficiency from corn stover biomass. For the development of the calibration model, 1150 dried and ground samples of corn biomass were included, while for external validation, 350 samples were utilized. Modified Partial Least Squares Regression (MPLS) and Multiple Linear Regression (MLR) were used to develop prediction models. The predictive capacity of the NIRS models was assessed by the coefficient of determination value in external validation ($r^2$) and the ratio between the standard deviation and the error of prediction in external validation (RPDev). Overall, MPLS regression provided better prediction statistics for saccharification efficiency ($r^2 = 0.80; \text{RPDev} = 2.21$) compared to MLR regression ($r^2 = 0.68; \text{RPDev} = 1.75$). These results indicated that predictive ability of MPLS model is more appropriate for screening purposes and NIR spectroscopy can be implemented as a large-phenotyping tool to test the saccharification yield of the corn biomass.

Introduction

Dependence on fossil biofuels has led to a major energy crisis with environmental and economic consequences of global concern. This situation has led to the development of new methods to find sustainable energy alternatives to meet the environmental requirements [1–3]. The production of ethanol from starch represents the most technically advanced option but gives rise to strong competition between energy and food supply [4]. Second generation biofuels, such as corn lignocellulose, derived from plant residues has become one of the main sustainable alternatives, not only for its high availability and wide adaptability but also for not interrupting energy demand and food supply [5].

Corn is an important food and feed crop, used as processed food, oil, feed and by-products. In addition, it can be used as a bioenergy crop in two ways, (i) the starch in the seeds can be used to produce ethanol, and (ii) crop residues could potentially be used to produce lignocellulosic ethanol [5, 6]. The conversion of lignocellulosic biomass to ethanol is a three-step process: (i) a pretreatment step, followed by (ii) hydrolytic degradation of the carbohydrates to the constituent sugar monomers (saccharification) and (iii) final fermentation of the free sugars to ethanol [2, 7]. Nevertheless, the key obstacle to the production of second-generation biofuels is the complicated structure of the cell wall, which is naturally resistant to decomposition and sugars release [8, 9].

Besides cell wall recalcitrance, evaluating and selecting the optimal feedstock from large-scale germplasm for saccharification efficiency is an indispensable strategy to improve lignocellulosic biofuel production [10, 11]. The analysis of large plant populations in breeding studies for cell wall digestibility is time-consuming, labor-intensive, and economically expensive [12–14]. At this time, improving the simplicity and yield of cell wall saccharification procedure represents the main technical hurdle to overcome.

The evaluation of corn biomass to ethanol suitability has been carried out using an automated saccharification assay that allows the analysis of a large, but limited, number of samples, and it is still restricted to various physical and chemical pretreatments [15]. In this regard, near infrared spectroscopy (NIRS) is a versatile, low-cost and non-destructive indirect analytical technology. It allows the high-throughput screening of populations at both qualitative and semi-quantitative levels. In recent years, this technology has also been applied to evaluate biomass digestibility in several species, such as wheat [16], eucalyptus [17], miscanthus [11], sweet sorghum [18], rice [19, 20], Jerusalem artichoke [13] and sugarcane [21]. However, so far, little has been explored on the NIRS potential for the determination of corn stover saccharification.

The main requirement for this technology is the need to develop calibration procedures combining laboratory data and spectral information to predict the sample properties and its composition. Hence, the main objective of the present work is to evaluate the capability of NIRS as a tool to predict the saccharification efficiency of lignocellulosic biomass of corn; in order to implement this technology in large scale phenotyping studies.

Materials and Methods


Dataset/Sample collection


The subset of 408 lines from the MAGIC population was evaluated following a single augmented design with 10 blocks, 42 non-replicate lines were included in each block, along with the eight inbred founders. Each plot consisted of a single row, 2.4 m long and 13 plants per row, with the spacing between consecutive hills in a row being 0.18 m and 0.8 m between rows. Whereas, the subset of 836 lines from the Ames Panel was evaluated following an augmented 17-block design, each block consisting of 50 lines and the six controls. Each plot consisted of a single row, 2.4 m long and 13 plants per row, with the spacing between consecutive hills in a row being 0.21 m and 0.8 m between rows. From the 836 lines evaluated, 300 lines well adapted to the conditions of Pontevedra and with sufficient material for the saccharification analyses were chosen. Local agronomical practices were followed.

The stover samples from both subsets were composed of tissues from 2 to 10 plants collected at grain harvest starting from 55 days after flowering. The samples, once dried (60°C, 7 days) were ground in a Restch SM100 mill with a 0.75 mm mesh for subsequent saccharification determination.

Saccharification efficiency measurements

Saccharification efficiency was determined following the method described by Gómez and coauthors [15]. Ground material was weighed into 96-well plates, each well contained 4 mg of each sample and processed using a high-throughput automated system. Samples were pre-treated with 0.5 M NaOH at 90°C for 30 min, washed four times with 500 µl sodium acetate buffer and finally subjected to enzymatic digestion (Celluclast 2, 7FPU/g) at 50°C for 9 h. The amount of released sugars was assessed against a glucose standard curve using the 3-methyl-2-benzothiazolinone hydrozone method [23].

NIR spectra acquisition

NIR spectra were collected on an instrument FOSS NIRS D2500 spectrometer (FOSS, Hillerød, Denmark) in the visible and near infrared region (400–2498 nm) at 0.5 nm intervals, in reflectance mode. Every sample was allowed to stabilize at room temperature prior to spectral data acquisition. The determinations were carried out in duplicate, with the dry and ground samples (50 g) placed in a circular quartz cuvette for solids at a temperature-controlled room (~24°C). The average spectrum of each sample was used for calibration and validation procedures. The acquired spectra were processed with WinISI software (version 4.12, Infrasoft International, PA, USA).

Chemometric analysis

Chemometric analysis was performed by both Modified Partial Least Square Regression (MPLS) and Multiple Linear Regression (MLR) methods. The MLR models were built with Stepwise selection of wavelength applies an F-test to identify the best-fitted model.

The dataset was randomly divided into two subsets using the SELECT algorithm included in the WinISI IV software. The prediction models were developed using a subset as the calibration set using ~ 75% of the samples (n = 1150) evaluated by leave-one-out cross-validation, and then tested on the remaining ~ 25% of the samples performing an external validation (n = 350).

A principal component analysis (PCA) on the first derivative of the absorbance was used to calculate the global Mahalanobis distance (GH) of each sample to the centre of the population in an n-dimensional space [24] using the CENTER algorithm included in the WinISI IV software. The samples with GH > 3 were identified as spectral outliers and removed, repeating the operation until all samples had a GH value lower than the recommended maximum [24]. During calibration process, three elimination passes of chemical outliers were applied, considering the critical T-statistic value set for chemical outliers detection was 2.5 [25]. Calibrations were developed after removing all outliers.
In order to develop the most accurate calibration models, different combinations of scatter corrections (NONE, no correction; D, detrending; SNV, standard normal variate; SNV + D, standard normal variate and detrending; WMSC, weighted multiplicative scatter correction; and SMSC, standard multiplicative scatter correction) and mathematical treatments (0,0,1,1; 1,4,4,1; 1,5,5,1; 1,6,4,1; 1,8,4,1; 1,10,5,1; 1,10,10,1; 2,4,4,1; 2,5,5,1; 2,6,4,1; 2,8,4,1; 2,10,5,1; 2,10,10,1, where the first digit is the derivative order, the second is the gap over which the derivative is calculated, the third is the number of data points in the first smoothing, and the fourth is the second smoothing) were tested [26].

The best models were assessed based on the standard error of cross-validation (SECV), the standard error of external validation (SEP), the coefficient of determination calculated in internal cross-validation ($1 - VR$) and external validation ($r^2$). In addition, the range error index (RER, defined as the ratio between of the range in the reference data to SECV/SEP) and the index of prediction to deviation (RPD, defined as the ratio between of the standard deviation for reference data to SECV/SEP [27, 28], was applied to evaluate the prediction ability of the models.

Results and Discussion

The average raw spectra from the whole set of samples of corn stover biomass is shown in Fig. 1. Five main absorption peaks at 1456, 1912, 2100, 2252 and 2310 nm, are related to the main components of corn stover spectra, in accordance with the spectral fingerprint showed by Guimarães and coauthors [29] for prediction of theoretical ethanol yield in sorghum biomass.

For the calibration and external validation sample sets the range of the saccharification efficiency of the data collection obtained by laboratory analysis at CNAP are shown in Fig. 2. The samples of calibration set are reported after the removal of all outliers (spectral and chemical), where the means (and ranges) expressed as nmol/mg material hour were: 153.3 (77.6 to 204.5) and 150.5 (77.6 to 204.5) for MPLS and MLR model, respectively. The external validation sample collection also had similar range and mean values, with 153.6 (77.6 to 204.5) nmol/mg material hour. Both in calibration and in external validation, the descriptive statistics showed similar behavior. These ranges were wider and higher than previous works with other crop species such as rice, barley, wheat, triticale, sorghum, miscanthus or brachypodium [30–33]. This great range of enzyme-released glucose was expected due to the broad origins of the samples and the environmental variability tested (two large panels tested from 2016 to 2019) [1, 34] and may suggest that many shifts will be represented in new germplasm phenotyping screenings.

The wavelengths selected in MLR calibration were: 824, 880, 1656, 1660, 1692, 1780 and 2332 nm (Fig. 1). Stepwise MLR coefficients for model developed were ranged from 814.658 (824 nm) to 748.711 (880 nm). The results showed that two wavelengths were the most relevant characteristic absorption peaks, particular at 824 and 880 nm, which are associated to the third overtone of C–H bond, related to sugars [35, 36]. The wavelength region from 1600 to 1800 nm is associated to the absorption band of a C–H stretching first overtone corresponding to fiber components of cell wall [37–39], peaks around 1780 nm being associated to the absorption band of a C–H stretching first overtone corresponding to carbohydrates, such as cellulose and hemicellulose[37, 38, 40, 41]. Other relevant coefficient appears in the region ~ 2332 nm, which assigned to cellulose and lignin absorption (C–H stretching/C–H deformation combination)[38, 42]. The two methods used for calibration showed similar trends in wavelength (or regions) associated/related with the chemical components of corn stover biomass. The cell wall structure governs bioethanol production by saccharification [12]; hence, in the present work, the wavelengths selected provide key information about the relevant chemical components involved in saccharification efficiency.

Based on the calibration mathematical treatments tested; the prediction models resulting from the second derivative (2,4,4,1), and a combination of standard normal variate and detrend as scatter correction method, provided a more accurate and precise estimate for saccharification efficiency. During calibration procedures, the number of samples removed as chemical T outliers, expressed as a percentage of the total initial samples in the set, ranged from 7.2 to 8.3% for both prediction models obtained. These values are lower than the maximum value (20%) recommended by Shenk and Westerhaus [26].

As can be observed in Table 1, calibration and cross-validation statistics revealed a better prediction model for the MPLS in comparison to the MLR. The coefficients of determination ($1 - VR$) and the standard errors of prediction in cross-validation (SECV) were 0.84 and 10.80 nmol/mg material hour for MPLS model, and 0.68 and 14.85 nmol/mg material hour for MLR model, respectively. According to Shenk and coauthors [43], NIRS prediction MPLS model with 1-VR value higher than 0.70 indicated a
good predictive ability, whilst MLR model (1-VR < 0.70) could only be used to qualitatively estimation (separate samples with higher and lower analytical values).

<table>
<thead>
<tr>
<th>Calibration</th>
<th>Cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>n</td>
</tr>
<tr>
<td>MPLS</td>
<td>1150</td>
</tr>
<tr>
<td>MLR</td>
<td>1150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>External validation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>n</td>
</tr>
<tr>
<td>MPLS</td>
<td>350</td>
</tr>
<tr>
<td>MLR</td>
<td>350</td>
</tr>
</tbody>
</table>

On the other hand, RPDcv and RERcv achieved values of 2.55 and 11.75 for MPLS model, and 1.78 and 8.54 for MLR model, respectively, thus indicating a reliable prediction power for MPLS model. By contrast, MLR model would only allow us to classify samples into high and low values of saccharification efficiency [27].

After the calibration process both models were validated with an external (independent) set of samples (Table 1). The values of the coefficient of determination $R^2$ and the RPDev were 0.80 and 2.21 for MPLS model and 0.68 and 1.75 for MLR model, respectively. Considering the criteria defined by Williams [27], the predictive quality of the calibration models based on RPDev values were considered unsuitable for MLR model (with an RPDev below 1.90 is not recommended to be used), and suitable for qualitative predictions for MLPS model (an RPDev between 2.50 and 2.90 could be applied for screening). However, we have to note that the RERev values for both models, shortly exceed the minimum value suggested by Williams [27] for a reliable quantitative model (RER > 10), with 10.03 and 12.64 for MLR and MPLS, respectively.

Comparing the results of the two calibration procedures evaluated, the MPLS calibration method demonstrated to have more predictive ability than the MLR to measure the prediction of saccharification efficiency (Fig. 3). MPLS is known to be a more effective model than MLR for the development of NIRS calibration models, particularly with large data sets, by reducing the dataset into a small number of orthogonal factors and to enabling avoid collinearity and over-fitting [44]. Additionally, MPLS is known to be more reliable than MLR for the calibration of complex chemical parameters [45]. The MPLS technique was better than MLR model in both calibration development and validation on independent set. Therefore, we were recommending constructed calibration model by MPLS technique in preference to MLR technique in this study.

Finally, contrasting the results obtained with other potential species for bioethanol production, Huang and coauthors [19], using MPLS model, reported similar or slightly lower predictive ability for estimating biomass saccharification (expressed as total releases sugar) ($R^2 = 0.75$, RPDev = 2.0), in a rice straws; while van der Weijde and coauthors [46] developed NIRS models to predict of saccharification efficiency of the crop Miscanthus and obtained good correlations (1-VR: 0.82 – 0.92); Li and colleagues [12] developed a calibration model that included different sugarcane genotypes, and they found RPD values of over 2.0 in calibration, internal cross-validation, and external validation. These results are comparable with our current work. However, and related to the
complexity of the parameter, we should note that the performance of our calibration models was more limited than those reported in sugarcane [21], who obtained NIRS models for fermentable hexoses and total sugar that exhibited excellent prediction capability (RPD values greater than 4.0) for predicting biomass digestibility. The usefulness of those last traits to estimate bioethanol potential could be evaluated in future studies with corn biomass.

Overall, the comparison of regression methods for estimating saccharification efficiency showed that MPLS was a better method than MLR based in terms of a larger correlation coefficient between predicted and reference values and smaller standard error of external validation. The results suggest that NIRS prediction by MPLS model can be used in a suitable way for discriminating corn genotypes attending to their saccharification output. Therefore, NIR spectroscopy represents an eco-friendly, fast, and low-cost predictive technique that can be used in the screening of large germplasm corn collections in relation to biomass potential.

**Declarations**

**Data Availability**

The data presented in this study are available on request from the corresponding author.

**Acknowledgements**

N.G. acknowledges an FPI contract (“Contrato predoctoral para la formación de doctores”) funded by project RTI2018-096776-B-C21 (financed by MCIU/AEI/FEDER, UE). A.L-M. acknowledges her postdoctoral scholarship to the Xunta de Galicia.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

This research was funded by subsequent coordinated projects financed by MCIU/AEI/FEDER, UE (RTI2018-096776-B-C21, RTI2018-096776-B-C22, PID2021-122196OB-C21 and PID2021-122196OB-C22).

**Author contributions**

Conceptualization, R.S.; Methodology, S.P.-C, N.G., A.L.-M.; Software, S.P.-C; Formal Analysis, S.P.-C. and N.G.; Investigation, R.S. and L.D.G; Data Curation, N.G. and A.L.-M.; Writing—Original Draft Preparation, S.P.-C, N.G.; Writing—Review and Editing, R.S., A.L.-M., L.D.; Supervision, R.S.; Project Administration, R.S; Funding Acquisition, R.S. All authors have read and agreed to the final version of the manuscript.

**References**


**Figures**

**Figure 1**

Average raw NIR absorption spectrum of a total set (n=1500) of corn biomass samples using near-infrared spectroscopy in reflectance mode. Dotted lines indicate five main absorption peaks related to the main components of corn stover spectra, in accordance with the spectral fingerprint for prediction of theoretical ethanol yield in sorghum biomass. NIR spectral absorbance values [log (1/R)], where R is the reflectance.
Boxplots of the saccharification data obtained in the two subsets included in this study. A: subset of 408 lines from a MAGIC population together with six founders (EP17, EP53, EP86, F473, A509, and EP125), and B: a subset of 836 lines, belonging to the Ames association panel (North Central Regional Plant Introduction Station, USA), together with 6 controls (A619, A632, A662, A665, PH207, EP42). Red dots indicate the mean values.
Figure 3

Scatter plot of reference data vs. predicted values by NIRS of saccharification efficiency (nmol mg-1 material-1 hora-1) for samples of corn stover biomass. A: Modified Partial Least Squares Regression (MPLS) and B: Multiple Linear Regression (MLR).