JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Article

pubs.acs.org/JAFC

High-Throughput Quantitative Biochemical Characterization of Algal Biomass by NIR Spectroscopy; Multiple Linear Regression and Multivariate Linear Regression Analysis

L. M. L. Laurens* and E. J. Wolfrum

National Bioenergy Center, National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, Colorado 80401, United States

Supporting Information

ABSTRACT: One of the challenges associated with microalgal biomass characterization and the comparison of microalgal strains and conversion processes is the rapid determination of the composition of algae. We have developed and applied a high-throughput screening technology based on near-infrared (NIR) spectroscopy for the rapid and accurate determination of algal biomass composition. We show that NIR spectroscopy can accurately predict the full composition using multivariate linear regression analysis of varying lipid, protein, and carbohydrate content of algal biomass samples from three strains. We also demonstrate a high quality of predictions of an independent validation set. A high-throughput 96-well configuration for spectroscopy gives equally good prediction relative to a ring-cup configuration, and thus, spectra can be obtained from as little as 10−20 mg of material. We found that lipids exhibit a dominant, distinct, and unique fingerprint in the NIR spectrum that allows for the use of single and multiple linear regression of respective wavelengths for the prediction of the biomass lipid content. This is not the case for carbohydrate and protein content, and thus, the use of multivariate statistical modeling approaches remains necessary.

KEYWORDS: *microalgae, biomass, lipids, infrared spectroscopy, multivariate calibration, chemometrics*

INTRODUCTION

Algal biomass compositional analysis data form the basis of a large number of techno-economic analysis models that are used to investigate and compare different processes for algal biofuels and bioproducts.[1] However, the analytical methods used to generate these data can be time-consuming and laborious and are far from standardized.[2] The techno-economic analysis of algal biofuel production has identified that a fast growing, high lipid producing, and easily harvestable algae is critical to make algal biofuel production economically feasible.[1]

Current chemical compositional analysis methods require large amounts of biomass (>1 g) and are thus not applicable for testing large culture collections or finding improved strains from thousands of candidates. As an alternative to the labor intensive chemical analyses, infrared spectroscopy, a non-destructive and high throughput approach, could be used for the simultaneous prediction of lipid, protein, and carbohydrate content in algal biomass. Near-infared (NIR) spectra from dispersive instruments (i.e., containing a polychromatic radiation source in combination with a grating as dispersive elements) consist of complex overtones and combinations of molecular vibrations, broad peaks from solid, opaque, and even liquid samples requiring minimal preparation.[3] Quantitative calibration models can be developed for the accurate prediction of the concentration of biochemical components, based on the correlations between the spectra and the composition of a select sample set. With appropriate calibration models, rapid predictions can be made on the composition of new samples using only spectra of the new samples.[4,5]

A prerequisite for the robustness of the NIR models for predicting composition is that the range in compositional variability needs to be sufficiently large to allow for predictions across species and for regression algorithms to subtract to orthogonal variation from the spectra. With a limited concentration range of the predicted components, the data set will likely not be equally distributed, the quality of the models will be reduced, and it will become more difficult to find a linear correlation in component concentrations.[6] Perhaps more importantly, the quality of the compositional prediction model will only be as good as the primary data. A "good" range for building calibration models depends on the absolute range of values for a given constituent but also on the precision of the primary measurements; the ratio of the range to the precision of a primary measurement is a better metric then either of these parameters alone. In the case of microalgae, obtaining robust primary compositional analysis measurements is not trivial and variation in accuracy of the primary measurements will undoubtedly carry forward to inaccurate prediction models using NIR spectroscopy.[2]

We have previously demonstrated the feasibility of NIR reflectance spectroscopy for quantitative determination of exogenously added lipids to algal biomass.[7] We were able to build calibration models based on NIR spectra solely correlated with the increasing concentration of lipids, indicating that lipids

Received: July 12, 2013
Revised: October 29, 2013
Accepted: November 14, 2013
Published: November 14, 2013

dx.doi.org/10.1021/jf403086f | J. Agric. Food Chem. 2013, 61, 12307−12314

in the presence of algal biomass have a sufficient fingerprint in the IR spectrum and those data formed the basis of this current research. An important early finding was that NIR, in contrast to mid-IR, was able to distinguish between neutral and polar lipids (triglyceride vs phosphatidylcholine lipids) among algal biomass. However, the question remained on whether biomass samples with variation in the endogenous composition can be distinguished on the basis of the NIR spectra.

The use of near and mid-IR on algal biomass has been explored in the literature and included successful demonstration of a relationship between changes in IR spectra with changes in the cells' biochemistry.[6,8,9] Mid-infrared spectroscopy was used for the prediction of the measured composition of algal biomass based on calibration curves from either single wavenumbers or multivariate regression of specific spectral ranges.[9] In addition, a comparison between near- and mid-IR for biomass composition recently showed that total nitrogen and ash content can be quantitatively correlated with infrared spectra; however, the authors in ref 6 were unable to demonstrate significant correlations for lipids or carbohydrates in microalgal biomass, using both NIR and mid-IR. These difficulties may be due to the lack of an adequate calibration sample set. Although the sample set was large, the range in lipid and carbohydrate concentration of the different samples may have been too small for spectral correlations. Because algae are rich in pigments and pigmentation changes during cultivation, it is likely that the visible region of the spectra dominated by these differences may skew quantitative correlations with biomass composition. This effect could compete with spectral information associated with composition and render quantitative multivariate analyses (e.g., partial least-squares regression, PLS) more difficult. For the prediction model development presented here, we have removed the visible region from the spectra. In addition to spectral wavelength selection, we also explored mathematical transformation of NIR spectra prior to building PLS calibration models to help with improving the predictions, and by subtracting scatter and other spectral variation not related to the composition of the biomass.

Although NIR in itself can increase the throughput of compositional analysis, in the case of microalgal biomass, often the amount of material is not sufficient to use in existing, more traditional, spectroscopy configurations. For example, a ring-cup configuration with a 1 in. diameter needs approximately 100−200 mg of biomass to cover the glass surface. In a high-throughput cultivation environment, this amount may be hard to obtain, particularly early in the growth cycle of an algal culture. For example, if the culture volumes are <500 mL, about 20−50 mg of material can be expected from a typical harvest. To allow for spectroscopy on much smaller biomass quantities (~10 mg), we have developed a 96-well plate configuration for NIR spectroscopy and we report here on the results of a comparison of the quality of the spectra and the quantitative predictions of biomass with a ring-cup configuration.

To address these potential challenges, in this work, we have collected spectra (350−2500 nm) of a selected set of algal biomass samples from three different strains, fully characterized in composition with respect to lipids, carbohydrates, protein, and ash composition. This sample set allows us to address the following questions arising from previous work and the literature. Can we distinguish differences in biochemical composition of algae based on the NIR spectra, and does the quality of the spectra and resulting prediction models vary with spectra collected in a ring-cup and a 96-well plate format? Can we build quantitative prediction models for important algal biomass constituents (e.g., lipids, carbohydrates, protein, and ash), and how accurately can we predict the composition of new, independent samples? To our knowledge, this is the first report of the use of NIR for the full biochemical composition of microalgal biomass using a combined species prediction model and the demonstration of predictions using different sample presentation configurations (including a 96-well configuration) and including an independent validation test set of the predicted composition.

MATERIALS AND METHODS

Biomass Samples. Algal biomass was grown in outdoor photobioreactors at Arizona State University as part of the Sustainable Algal Biofuels Consortium (SABC) collaboration and shipped to NREL for full compositional analysis. The biomass was collected over the course of nutrient depletion for three different strains—*Chlorella* sp., *Scenedesmus* sp., and *Nannochloropsis* sp.—and represents strain-specific patterns of lipid, protein, and carbohydrate content, based on the timing of harvest. A total of 38 biomass samples were selected for this work: 10 samples for *Chlorella* sp., 15 for *Scenedesmus* sp., and 9 for *Nannochloropsis* sp.

Biomass Compositional Analysis. The biomass composition was determined using the current best methods selected for compositional analysis after a thorough comparison of method uncertainty and measurement chemistry that was reported in 2012.[2] In brief, lipids were determined as total fatty acid methyl ester content via a direct, whole biomass transesterification reaction.[10] The procedure consisted of dissolving 10 mg of lyophyilized algal biomass sample in 0.2 mL of chloroform:methanol (2:1, v/v) and subsequent transesterification of the lipids *in situ* with 0.3 mL of HCl:methanol (5%, w/v) for 1 h at 80 °C in the presence of 250 μg of tridecanoic acid (C13) methyl ester as an internal standard. The resulting FAMEs were extracted with hexane at room temperature for 1 h and analyzed by gas chromatography with flame ionization detection (GC-FID) (Agilent 6890N; HP5 30 m 0.25 mm i.d. and 0.25 μm film thickness; temperature program 70−300 °C over 23 min at 10 °C min^{-1}).

Carbohydrates were determined via acid hydrolysis as follows: 100 mg of algal biomass was subjected to a two-stage sulfuric acid hydrolysis (1 h at 30 °C in 72% (w/v) sulfuric acid, followed by 1 h at 121 °C in 4% (w/v) sulfuric acid in an autoclave). After hydrolysis, the acid insoluble residue was separated from the hydrolysate using ceramic filtering crucibles. Soluble neutral carbohydrates (glucose, xylose, rhamnose, fucose, galactose, arabinose, and mannose) were determined by high-performance liquid chromatography; HPLC analytical conditions were as described in ref 11.

Protein was determined through an elemental nitrogen-to-protein conversion factor of 4.78 specific for microalgae, derived from the literature.[12] Ash and moisture were determined on 100 mg of algal biomass, weighed into ceramic crucibles, and dried overnight in a drying oven (105 °C), followed by precombustion of oven-dried algal material over a Bunsen burner followed by placement in a muffle furnace (575 °C) until constant weight.

NIR Spectroscopy and Data Analysis. NIR spectra were collected on freeze-dried biomass using either the Foss NIR Systems model XDS Forage Analyzer (Foss, Silver Spring, MD, USA) or the ASD LabSpec Pro (ASD inc., Boulder, CO, USA). Two sample presentation configurations were compared. Spectra for the ring-cup sample presentation were collected using a reflectance module using the Foss XDS spectrometer (as described for terrestrial biomass in ref 13). For each sample, prepared in a circular sample cell with a 1-in. insert, a total of four spectra were collected and averaged (three scans per spectrum of four replicate prepared samples). The spectra were collected in the range 400−2500 nm. WinISI software (Foss) was used for collection, standardization, and export of the spectra. Four replicate spectra were collected for each sample, with each replicate representing a different angle of cup presentation to the spectrometer. NIR spectra in the 96-well format were collected in opaque white 96-

well plates using an ASD LabSpec Pro spectrometer and data collected in solid white microtiter plates, where empty wells were used for collecting reference spectra (baselining). Spectra were transformed from reflectance to absorbance spectra ($\ln(1/R)$) prior to any mathematical and spectral transformations.

All transformed NIR spectra were processed in R version 3.0.1,[14] and statistical analyses were carried out using the following packages: "*chemometrics*" version 1.3.8,[15] "*signal*" version 0.7-1,[16] and "*pls*" version 2.3-0[17] along with functions present in base R. Principal Component Analyses (PCA) were calculated using the singular value decomposition (SVD) algorithm. Partial least squares (PLS) regression analysis was used for quantitative correlation. For all models, PLS regression was performed using the NIPALS algorithm, using full, leave-one-out cross validation on a centered data set. The optimum number of principal components used for the PLS regression is shown in the text accompanying the figures and was selected on the basis of an apparent minimum in root-mean-square error of the prediction (RMSEP) of the cross-validation of the models. The effect on the statistics of the calibration models of eliminating part of the visible spectrum was investigated by recalculating the models excluding the visible region of the spectrum (wavelengths 400−1100 nm). In order to find the best calibration model, we investigated the effect of mathematical spectral pretreatment and spectral derivatives on the quality of the prediction model for NIR spectra including or excluding the visible region of the spectra. The algorithms we used were multiplicative scatter correction (MSC), standard normal variate (SNV), and Savitsky−Golay smoothing/derivatization of the spectra.

Single and multiple linear regression models were included and used the following lipid-specific wavelengths: 1215, 1725, and 2305 nm. The models and predictions of the independent validation set were performed using functions available in base R.

All raw data and scripts used to generate the results presented here are available as Supporting Information.

RESULTS AND DISCUSSION

Compositional Analysis of Algal Biomass Samples. We selected 38 biomass samples from three different species: *Chlorella* sp. (CZ), *Scenedesmus* sp. (SD), and *Nannochloropsis* sp. (NC). These species were grown under conditions to maximize lipid, carbohydrate, or protein content, as part of a larger project that aims to investigate the trade-offs between total product value and production costs, which emphasized the need for full biomass compositional analysis.

A summary of the composition and the range for each of the components is shown in Figure 1. This data set represents compositional variation spanning the concentration ranges necessary to build prediction models. The compositional and spectral variation that is found in these samples shows an even distribution of lipids (6.8−53.0% DW), protein (7.4−42.5% DW), carbohydrate (9.5−52.3% DW), and ash (1.1−10.1% DW) content in the algal biomass. The compositional variability in this data set makes it possible to develop robust models for characterization of a wide range of new algal biomass samples and allows for the statistical reduction and subtraction of spectral information not correlating to the biochemical components of biomass.

Spectroscopy in Two Different Configurations. We collected four replicate spectra from each of the biomass samples, in both a ring-cup and a 96-well-type sample presentation configuration with 200 and 10 mg per sample, respectively. We found that for the small quantities of biomass good quality spectra could be obtained in the 96-well plate format; however, a reduction of the absorbance of the 2300−2500 nm region (and concomitant increase in the noise levels) was observed due to light absorption by the fiberoptic probe. Visual differences in the biomass from the different strains are

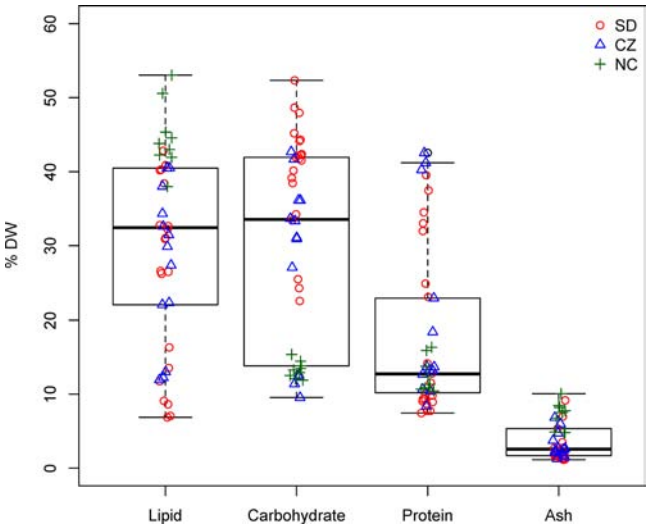

Figure 1. Summary of the compositional data used for multivariate calibration. Lipid, carbohydrate, protein, and ash content for three strains: *Chlorella* sp., CZ (n = 10); *Nannochloropsis* sp., NC (n = 9); and *Scenedesmus* sp., SD (n = 15).

reflected in large spectral variation in the visible region, as shown in Figure 2, where typical spectra of a high and low lipid content biomass sample for each of the three strains are shown. The spectra illustrate significant interspecies differences in the visible region of the spectrum. When comparing the respective high- and low-lipid spectra, it is clear that the same regions of the NIR spectrum are increasing with increased lipid content for all three algal strains, with the largest changes found at 1215, 1725, and 2305 nm, respectively. These observations are consistent with the spectral absorption bands associated with lipids and found in the literature[18] and supported by the major absorbance from a triglyceride standard included in Figure 2. The characteristic absorption bands of lipids in the NIR spectrum are (i) the first overtones of C−H stretching vibrations (1600−1900 nm), (ii) the region of second overtones of C−H stretching vibrations (1100−1250 nm), and (iii) two regions (2000−2350 and 1350−1500 nm) which contain bands due to combinations of C−H stretching vibrations and other vibrational modes.[18]

Principal Component Analysis (PCA). To investigate structure in the data set and identify the major variation contributions, we performed principal component analysis (PCA) on the ring-cup full spectra. PCA indicates grouping mainly based on species (along principal component 1, PC1, explaining 84.1% of the variation) and based on the compositional differences (along PC2, explaining 9.1% variation). The contribution of the spectral variation after spectral normalization (MSC and SnvDF) follows a different pattern, with PC1 indicating a higher contribution of the compositional information, indicated by the measured lipid content for each sample (explaining 56.5% variation), and species-specific information is less pronounced. This illustrates the advantage of performing mathematical pretreatment prior to multivariate analysis of spectra, in particular when large spectral variation due to the use of different strains is present and not desirable for a species-agnostic prediction model (data shown in Figure 3, with lipid content indicated for each sample). The effect of the visible region was not noticeable in that the principal-component-based groupings observed were conserved with or without the visible region, indicating that the

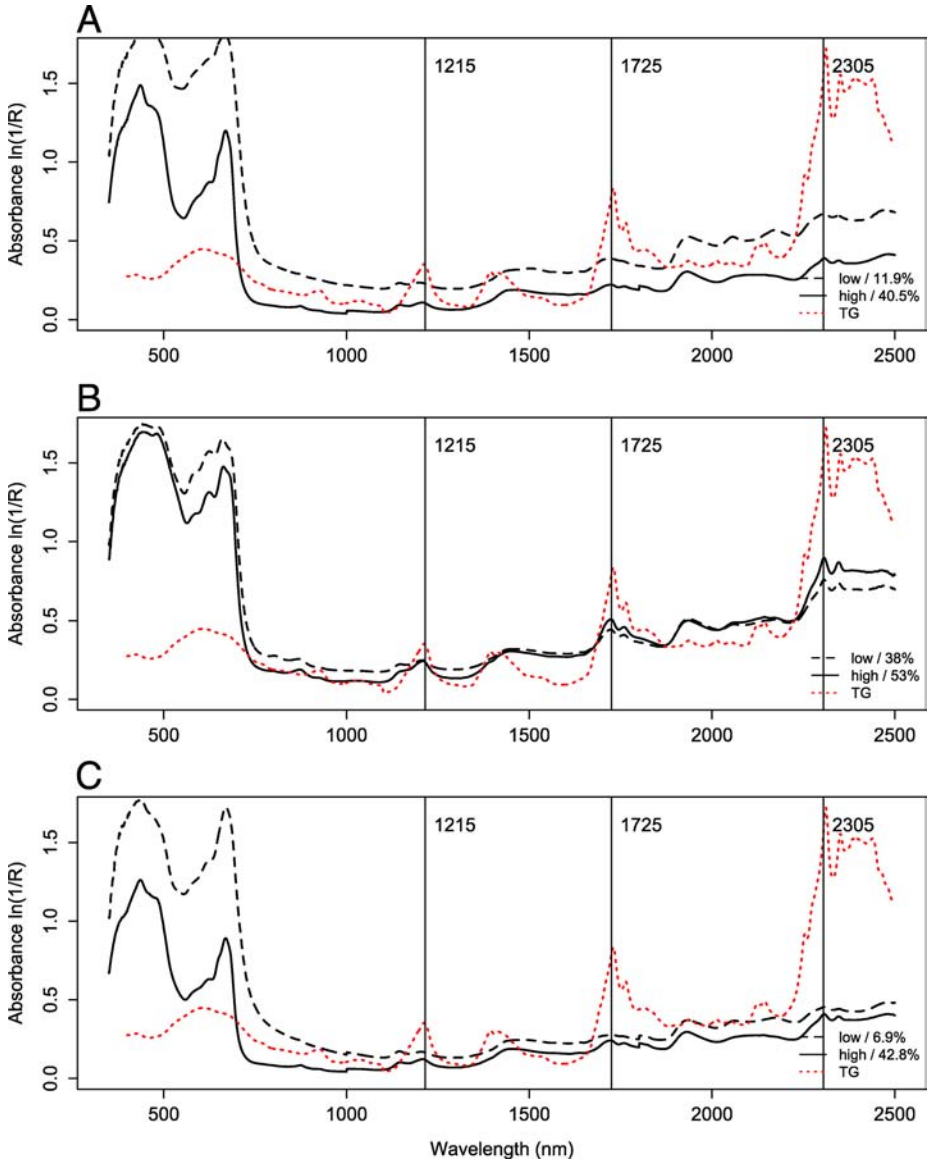

Figure 2. Overlay of spectra of high (solid line) and low (dashed line) lipid containing samples for (A) *Chlorella* sp., (B) *Nannochloropsis* sp., and (C) *Scenedesmus* sp., superimposed on a pure triglyceride spectrum (dotted red line), collected in a 96-well plate format. Selected wavelengths in the spectra are highlighted that are corresponding to the main lipid overtones, 1215, 1725, and 2305 nm.

interspecies differences in the visible region of the spectra do not significantly carry through in the IR region (data not shown).

Partial Least Square (PLS) Regression. We used PLS regression to develop quantitative predictive models of algal biomass composition. The quality of these models is shown in Figure 4, showing predicted-versus-measured plots of the calibration and leave-one-out cross-validation models based on ring-cup-spectra for lipids, carbohydrates, and protein. These models used three principal components (PCs) for prediction.

A plot of the root mean squared error of the cross validation (RMSECV) prediction relative to the number of components or latent variables used in the models is shown in Supplementary Figure A (Supporting Information) and shows a clear minimum at two or three components, which supports our use of three components for the quantitative linear regressions. The effect of different spectral pretreatments to remove scatter due to different particle sizes or strain-specific features is typically scored on the basis of RMSECV and R^2 values[19] and the number of principal components needed to build the regression model. The use of fewer principal components typically gives more robust models, since less noise is being included in the fitting algorithm. We performed multiple mathematical spectral pretreatments and found that a standard normal variate (SNV) correction, where the sum-squared deviation over the spectrum equals unity, gives the best models thanks to the removal of the species-specific spectral fingerprints. This causes a concentration of the spectral variation around the biochemical composition variation.

One important objective of this work was the demonstration of a high-throughput configuration for NIR spectroscopy. The data comparing the prediction model quality between two sample presentation configurations and after different spectral treatments is shown in Table 1. For both configurations, removing the visible region of the spectrum significantly improved the models, presumably due to the removal of the effect of pigments and other visible spectral features distinguishing between the three different strains. For lipids, protein, and

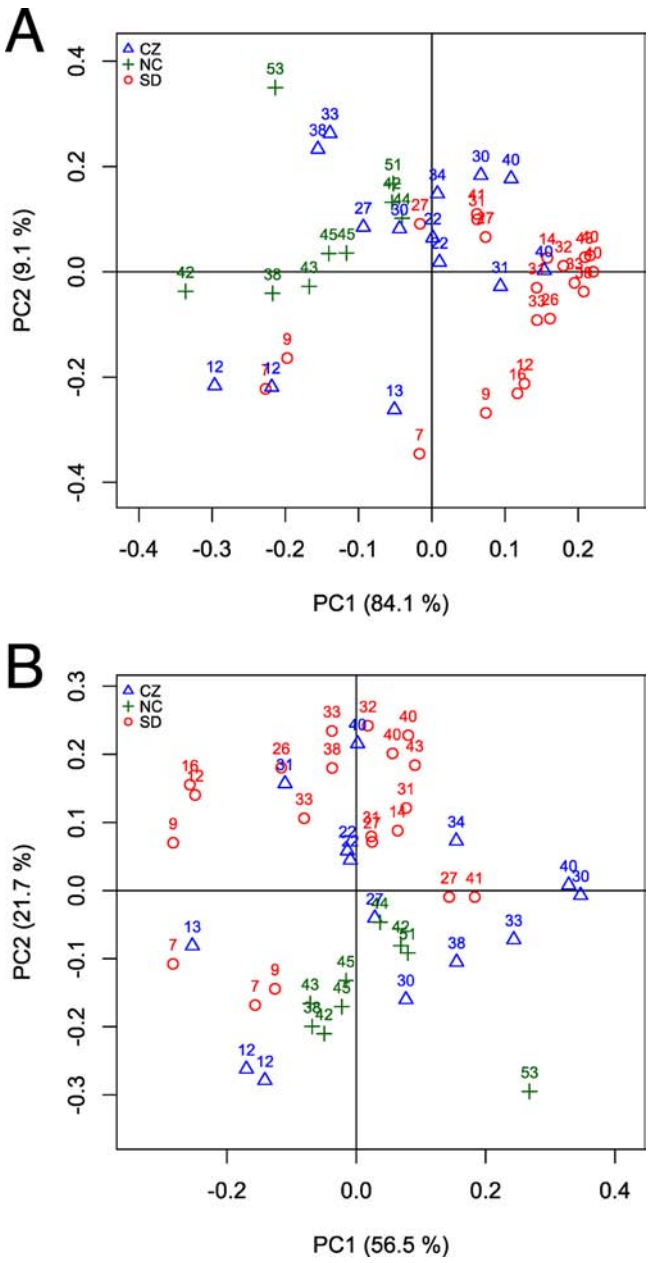

Figure 3. Principal component analysis of ring-cup full spectra, colored by strain, *Chlorella* sp., CZ; *Nannochloropsis* sp., NC; and *Scenedesmus* sp., SD before (A) and after (B) spectral normalization. The lipid content for each spectrum is shown above each symbol.

carbohydrates, the correlation coefficients were in all cases close to 0.9 (except for carbohydrates, where the coefficients are closer to 0.8), indicating strong and unique fingerprints in the NIR spectra. The root-mean-squared error in the cross-validation (RMSECV) values of the SnvDF model with two PCs indicates a less than 4% deviation for lipids, ~2.1% for protein, and less than 6% for carbohydrates (Table 1). The RMSECV was lowest for the protein measurements; however, the RMSECV plot against the components indicates a much less clear minimum, and a larger number of components was needed to obtain the best calibration model.

When comparing and interpreting the regression coefficients (Supplementary Figure B, Supporting Information) as those wavelengths that contribute positively to the calibration prediction model, we notice significant influence from the wavelengths that show resemblance to the pure triglyceride absorbance peak (1215, 1700–1900, and 2305 nm). This supports the specificity of the model for lipid-specific absorbance peaks. The coefficients for the protein and carbohydrate prediction models were distinct from the lipid-specific peaks and from each other and are shown in Supplementary Figure B (Supporting Information).

The prediction models for the ash content are not shown because of their poor performance (R^2 = 0.66), accuracy, and precision. IR models for the prediction of ash are typically based on the inverse correlation against the "absence" of organic material because the inorganic content of biomass does not have a fingerprint in the IR region. The lack of good prediction models can be partly explained by the rather limited range in ash concentration (1.1–10%) of the biomass used in these models.

In another test of the accuracy and precision of the quantitative prediction models, we have used four samples (~10% of the sample set) as independent validation samples, which were not included in the calibration set. We predicted the composition using the multivariate linear regression models built and summarized in Table 1 and compared the accuracy of the predictions against the measured composition (shown as the actual value (A) in Table 2). The results of the quantitative predictions for the independent validation set are shown in Table 2. Of the 36 predicted values, only 6 predictions showed more than 15% relative difference from the actual values. Both the ring-cup and the 96-well format performed well in the independent prediction. The ring-cup predictions were on average within 9.1% (relative) of the actual value and the 96-well predictions within 12.1% of the actual value. This indicates that the 96-well plate configuration for NIR spectroscopy can be used for the high-throughput application that we set out to develop. It is likely that, with the development of prediction models and larger data sets, the accuracy of the prediction of independent test samples will improve and the detection of outliers will be made easier.

Multiple Linear Regression. Because the spectra shown in Figure 2 illustrate distinct increases in regions of the spectra related directly to lipid-specific overtones in the NIR spectra, we investigated the utilization of single and multiple linear regression for lipid content quantification. This application opens up more possibilities for rapid screening of biomass composition and may eliminate the need for multivariate statistical prediction models for lipid quantification. This application is often dismissed because of the complexity of the IR spectra and the likelihood that the lipid-specific features would be dwarfed by other spectral variation. The correlations shown in Figure 5 illustrate good correlation when spectral data from either 1725 or 2305 nm are used (R^2 of 0.86 and 0.77, respectively, Figure 5A and B), although the correlation between lipid content and absorbance at 1215 nm alone is unsatisfactory (R^2 = 0.66, data not shown). The two wavelengths 1725 and 2305 nm show a significant improvement in the correlation when both wavelengths are used together for multiple linear regression (Figure 5C).

The independent quantitative predictions with the four independent test samples are shown in Table 2 and indicate that the lipid content in all four samples can be predicted to within 4% relative deviation of the actual value. These data suggest that in this particular scenario MLR might outperform the multivariate prediction models. It remains to be seen whether this observation is true with other sample sets.

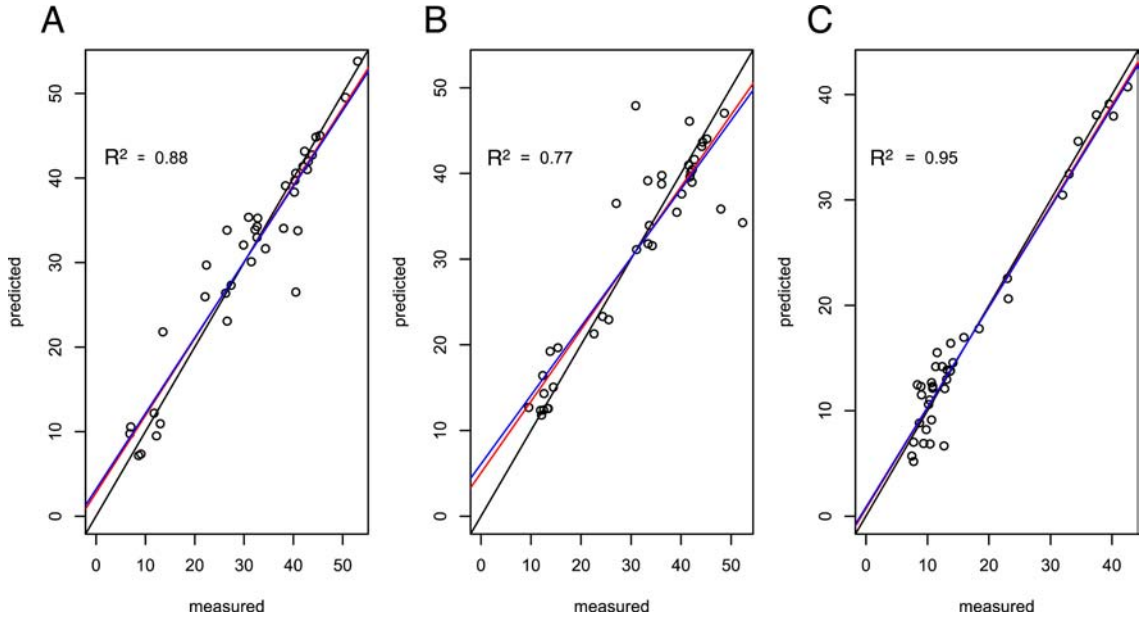

Figure 4. Partial least squares (PLS2) modeling results for lipid (A), carbohydrate (B), and protein (C) content of biomass. Results are shown as predicted vs measured plots. The blue lines represent the calibration prediction, the red line the cross validation prediction, and the black line perfect agreement between measured and predicted values for each constituent, using three principal components. Open circles represent cross-validated results of the calibration set. Spectra were smoothed and normalized using a standard normal variate correction (SnvDF) prior to modeling. Other spectral pretreatments provided similar results (see text and Table 1).

Table 1. Summary Statistics of Correlation Coefficients for Prediction Models[a]

		ring-cup format					96-well format			
	no. of PCs		lipid	carb	protein	no. of PCs		lipid	carb	protein
raw	5	R^2-c	0.86	0.84	0.93	5	R^2-c	0.87	0.86	0.96
		RMSEC	4.78	5.15	2.67		RMSEC	4.56	4.93	2.24
		R^2-cv	0.80	0.79	0.89		R^2-cv	0.79	0.78	0.92
		RMSECV	5.67	6.05	3.38		RMSECV	5.79	6.27	2.93
		RMSEP	4.78	5.15	2.67		RMSEP	4.56	4.93	2.24
SnvDF	3	R^2-c	0.91	0.83	0.96	3	R^2-c	0.92	0.82	0.96
		RMSEC	3.84	5.31	2.11		RMSEC	3.64	5.71	2.23
		R^2-cv	0.88	0.77	0.95		R^2-cv	0.84	0.72	0.92
		RMSECV	4.38	6.32	2.40		RMSECV	5.06	7.07	2.91
		RMSEP	3.84	5.31	2.11		RMSEP	3.64	5.71	2.23
first deriv	2	R^2-c	0.81	0.83	0.89	2	R^2-c	0.83	0.85	0.88
		RMSEC	5.62	5.31	3.37		RMSEC	5.29	5.25	3.58
		R^2-cv	0.76	0.80	0.85		R^2-cv	0.78	0.81	0.85
		RMSECV	6.33	5.87	4.02		RMSECV	5.93	5.88	4.15
		RMSEP	5.62	5.31	3.37		RMSEP	5.29	5.25	3.58

[a]Abbreviations: R^2-c, correlation coefficient of the calibration; RMSEC, root-mean-squared error of calibration model; R^2-cv, correlation of cross validation; RMSECV, root-mean-squared error of cross validation model; RMSEP, root-mean-squared error of predictions of independent test set.

Table 2. Quantitative Prediction of the Composition of an Independent Validation Set of Four Samples[a]

	lipids				carbohydrates			protein		
	A	ring-cup	96-well	MLR	*A*	ring-cup	96-well	*A*	ring-cup	96-well
Scenedesmus sp.	16.33	16.93	16.16	17.26	38.45	34.54	38.25	24.95	25.58	25.46
Nannochloropsis sp.	37.99	35.68	37.78	38.68	12.92	14.00	10.93	16.35	20.41	19.63
Scenedesmus sp.	31.04	33.13	31.59	30.22	42.38	40.23	38.98	9.32	11.04	13.03
Chlorella sp.	11.95	7.48	11.14	11.28	11.38	13.61	8.79	41.20	41.99	41.55

[a]Values represent predicted concentrations based on either ring-cup or 96-well SNV-corrected spectra using multivariate prediction models or multiple linear regression (MLR) as indicated. *A* = actual measured concentration.

When using those same wavelengths for carbohydrate and protein determination, we found correlations of 0.38 and 0.76, respectively, with large scatter (data not shown). These results suggest that these wavelengths are suitable for lipid content correlation and perhaps for protein content but not for carbohydrate content predictions. However, further investiga-

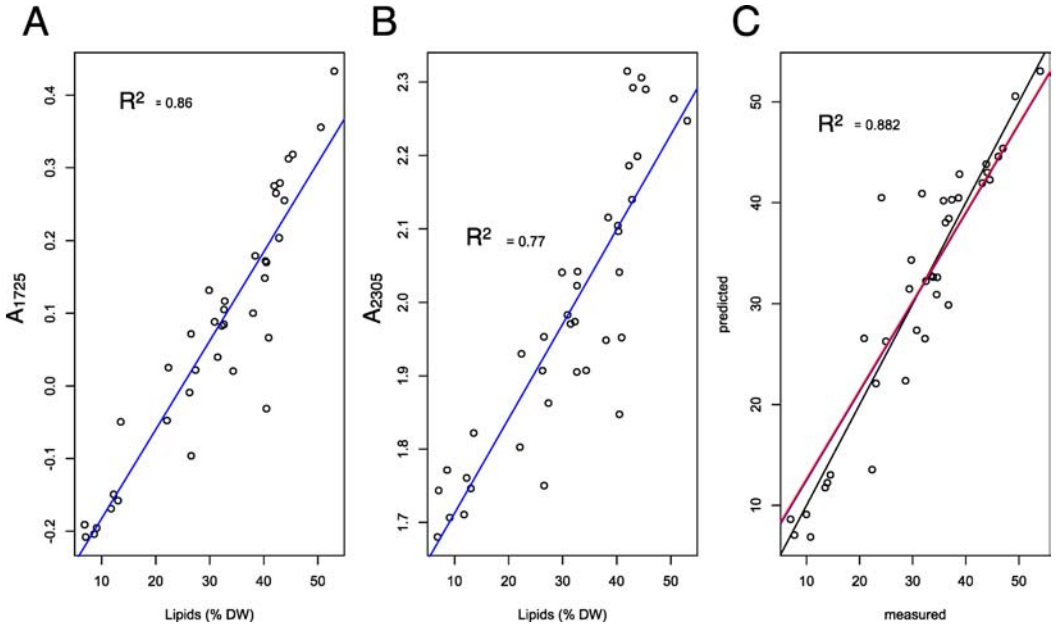

Figure 5. Results of single and multiple linear regression (MLR) for the prediction of lipids from NIR spectra, correlation between the lipid content and absorbance at 1725 nm (A), absorbance at 2305 nm (B), and multiple linear regression (MLR) predicted vs measured plot of lipid content against two wavelengths (1725 nm, 2305 nm) (C). The blue lines represent the calibration prediction, the red line the cross validation prediction, and the black line perfect agreement between measured and predicted lipid concentrations.

tion of the protein correlation is needed to ensure, for example, that the protein correlation model is not an inverse relationship with diminishing lipid content, particularly since protein and lipid content are inversely correlated in algal biomass. Carbohydrates do not seem to correlate with the three wavelengths investigated, although additional wavelengths could be found. It is likely that the carbohydrate NIR fingerprints are much less pronounced such that multivariate full spectrum regression analysis is necessary to obtain good quantitative correlations.

In conclusion, we have demonstrated that there is spectral information in the NIR region that can be quantitatively correlated with compositional differences among algal biomass samples from three different strains. There are large influences of interspecies differences in the visible and NIR portions of the spectrum; spectral transformation functions could partly reduce this effect and aid in further multivariate analyses. Our work suggests that regression models can indeed be used on the basis of the measured lipid content found in algal biomass. However, it remains to be determined whether cross-species prediction across more widely different algal species will be possible to generate robust prediction models, or whether individual groups of organisms (based, for example, on phyologenetic relationships) will require separate calibration models, which would greatly limit the applicability of NIR spectroscopy for microalgal compositional analysis. The 96-well high-throughput NIR approach presented here shows that we can obtain accurate independent predictions from a data set consisting of 38 biomass samples and together with the application of multiple linear regression analysis allows for a much improved and increased throughput of microalgal compositional analysis. A fully integrated high-throughput approach would involve cultivation of biomass in 96-well plate format followed by quantitative NIR spectroscopic prediction of the composition. For the work presented here, we did not use this approach; rather, we added biomass to 96-well plates. We do not anticipate a significant difference in spectroscopy characteristics; however, generating the biomass quantities needed for full biochemical compositional analysis (75−150 mg of dry material) may be a challenge for building the initial and strain-specific calibration models. The technology described here shows promise as a non-invasive, rapid measurement of full biochemical composition of algal biomass.

■ ASSOCIATED CONTENT

Supporting Information

A pdf file containing two figures: Supplementary Figure A: Plots of root-mean-square error of cross validation (RMSEP) for the PLS2 model shown in Figure 4; models for lipid (A), carbohydrate (B), and protein (C) concentrations. Spectra were derivatized prior to modeling. Two principal components (PCs) appear to provide the best results. Supplementary Figure B: Plots of the regression coefficients of the prediction models of the cross validation results shown in Figure 4, models for lipid (A), carbohydrate (B), and protein (C) concentrations. A txt file containing all raw data and scripts used to generate the results presented in the paper. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Lieve.Laurens@nrel.gov. Phone: +1-303-384-6196.

Funding

This work was supported by the U.S. Department of Energy under Contract No. DE-AC36-08-GO28308 with the National Renewable Energy jointly as part of the BioEnergy Technology Office (BETO), under task 9.6.1.8 and the Sustainable Algal Biofuels Consortium project, funded under DOE Award No. DE-EE0003372.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge Stefanie Maletich and Stefanie Van Wychen who provided the compositional analysis data and collected the spectra used for this work and Margaret Robinson who worked on algal biomass NIR spectroscopy and made improvements to sample configurations. We acknowledge Drs. John McGowen and Thomas Dempster (Arizona State University) for providing the biomass and Dr. Philip Pienkos for critical review of the manuscript.

■ REFERENCES

(1) Davis, R.; Aden, A.; Pienkos, P. T. Techno-economic analysis of autotrophic microalgae for fuel production. *Appl. Energy* **2011**, *88*, 3524−3531.

(2) Laurens, L. M. L.; Dempster, T. A.; Jones, H. D. T.; Wolfrum, E. J.; Van Wychen, S.; McAllister, J. S. P.; Rencenberger, M.; Parchert, K. J.; Gloe, L. M. Algal biomass constituent analysis: method uncertainties and investigation of the underlying measuring chemistries. *Anal. Chem.* **2012**, *84*, 1879−87.

(3) Ciurczak, E. W., Burns, D. A., Eds. *Handbook of near-infrared analysis*; Marcel Dekker: New York, 2001.

(4) Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. Selection of samples for calibration. *A user-friendly guide to multivariate calibration and classifications*; NIR publications: Chichester, U.K., 2002.

(5) Martens, H.; Naes, T. *Multivariate calibration*; John Wiley: New York, 1989.

(6) Mulbry, W.; Reeves, J.; Liu, Y.; Ruan, Z.; Liao, W. Near- and mid-infrared spectroscopic determination of algal composition. *J. Appl. Phycol.* **2012**, *24*, 1261−1267.

(7) Laurens, L. M. L.; Wolfrum, E. J. Feasibility of Spectroscopic Characterization of Algal Lipids: Chemometric Correlation of NIR and FTIR Spectra with Exogenous Lipids in Algal Biomass. *BioEnergy Res.* **2010**, *4*, 22−35.

(8) Dean, A. P.; Sigee, D. C.; Estrada, B.; Pittman, J. K. Using FTIR spectroscopy for rapid determination of lipid accumulation in response to nitrogen limitation in freshwater microalgae. *Bioresour. Technol.* **2010**, *101*, 4499−4507.

(9) Wagner, H.; Liu, Z.; Langner, U.; Stehfest, K.; Wilhelm, C. The use of FTIR spectroscopy to assess quantitative changes in the biochemical composition of microalgae. *J. Biophotonics* **2010**, *3*, 557−66.

(10) Laurens, L. M. L.; Quinn, M.; Van Wychen, S.; Templeton, D.; Wolfrum, E. J. Accurate and reliable quantification of total microalgal fuel potential as fatty acid methyl esters by in situ transesterification. *Anal. Bioanal. Chem.* **2012**, *403*, 167−178.

(11) Templeton, D.; Quinn, M.; Van Wychen, S.; Hyman, D.; Laurens, L. M. L. Separation and quantification of microalgal carbohydrates. *J. Chromatogr., A* **2012**, *1270*, 225−234.

(12) Lourenço, S. O.; Barbarino, E.; Lavín, P. L.; Lanfer Marquez, U. M.; Aidar, E. Distribution of intracellular nitrogen in marine microalgae: Calculation of new nitrogen-to-protein conversion factors. *Eur. J. Phycol.* **2004**, *39*, 17−32.

(13) Wolfrum, E. J.; Sluiter, A. D. Improved multivariate calibration models for corn stover feedstocks and dilute-acid pretreated corn stover. *Cellulose* **2009**, *16*, 567.

(14) R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing (http://www.R-project.org): Vienna, Austria, 2013.

(15) Filzmoser, P.; Varmuza, K. Chemometrics: Multivariate Statistical Analysis in Chemometrics. *R Packag. version 1.3.8*, 2012.

(16) SignalDevelopers signal: Signal Processing. 2013.

(17) Bjørn-Helge Mevik, R. W.; Liland, K. H. pls: Partial Least Squares and Principal Component Regression. *R Packag. version 2.3-0*, 2011.

(18) Ismail, A. A.; Nicodemo, A.; Sedman, J.; van de Voort, F. R.; Holzbaur, I. E. Infrared spectroscopy of lipids: principles and applications. In *Spectral properties of lipids*; Hamilton, R. J., Cast, J., Ed.; CRC Press LLC: Boca Raton, FL, 1999.

(19) Esbensen, K. H. *Multivariate Data Analysis - in practice: an introduction to mutlivariate data analysis and experimental design*; CAMO Process AS: Oslo, Norway, 2002.