Accurate Prediction of Algal Biomass Lipid, Protein, and Carbohydrate Composition with Machine Learning Regression Modelling of near-IR Spectra

Steichen, S.[1], Pinowska, A.[2], Brown Clay, J.[2], Hazlebeck, D. [2], Laurens, L.M.L.[1]
[1]National Renewable Energy Laboratory, Golden, CO
[2]Global Algae; 4473 Pahee St. Suite T, Lihue, HI 96766

U.S. DEPARTMENT OF ENERGY | Energy Efficiency & Renewable Energy

Introduction

Algae Biomass Composition

- Cultivated algae can be harvested and used as a renewable energy source
- Lipids (FAME), proteins, and carbohydrates can be extracted from cells and refined to make fuel and other products in a carbon neutral cycle
- Optimal strains for production depend on local environmental factors and output goals. Relative biomass composition can vary significantly between and during individual growth cycles
- Current methods for the compositional analysis of each sample are time consuming, labor intensive, and use volatile chemicals[1]
- Near Infrared (NIR) absorption spectra contain features that can be quantitatively related to chemical compounds underlying composition[2,3]
- Lower cost spectrometers and rapid turnaround analyses are requisite for productions sites
- Combination of robotically assisted sample processing and machine learning model-based predictions can improve times and predictive accuracy from relatively noisy input data.

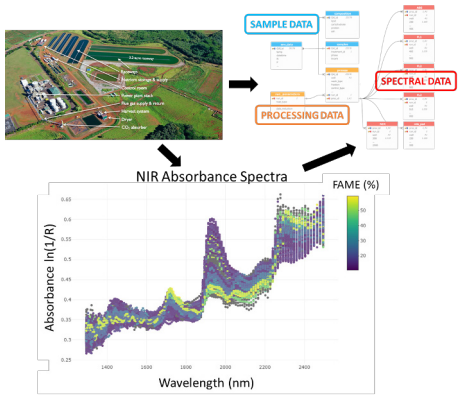

Figure 1. Overview of algal biomass sample and data collection process.

Research Goal

Develop and optimize methods for rapid, cost-effective, and accurate quantification of algal composition based on NIR spectral data

Results and Discussion

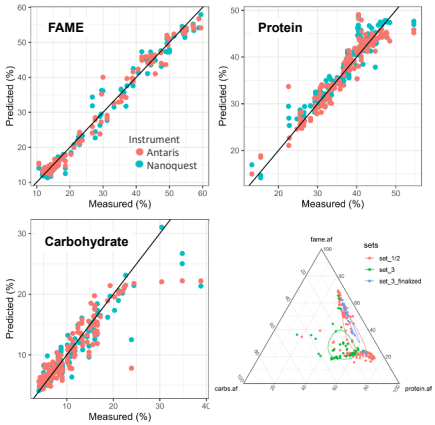

Figure 3. Prediction accuracy of optimized models trained on NIR data collected from a high performance Thermo Antaris II instrument, and low-cost Ocean Insights NanoQuest spectrometer. Separate models were developed for each instrument and component combination using PLSR and ANN methods (ANN-based predictions displayed). Ternary plot displays relative biomass component concentrations on an ash free dry weight basis, colored according to sequentially designed and collected sample sets.

Table 1. Compositional prediction performance metrics

Component	Model spec. (Antaris spectra)	Preprocessing	RSQ	RPIQ	MAPE (relative deviation)	% of samples < 15% APE
FAME	PLS	MinMax	0.977	6.74	15.4	93.4
FAME	ANN	MinMax	0.976	9.95	6.68	95.4
FAME	ANN	SNV-SG	0.981	11.03	5.99	93.4
Protein	PLS	MinMax	0.857	3.76	5.31	89.3
Protein	ANN	MinMax	0.896	4.30	5.02	94.0
Protein	ANN	SNV-SG	0.901	4.38	5.09	96.0
Carbohydrate	PLS	MinMax	0.631	1.84	29.3	53.3
Carbohydrate	ANN	MinMax	0.712	2.50	17.61	55.3
Carbohydrate	ANN	SNV-SG	0.871	3.48	12.54	66.2

** RSQ = R-squared; MAPE = mean absolute error of prediction; RPIQ = Ratio of Performance to Interquartile range*

- 213 unique samples were collected from conditions representing a wide distribution of production-relevant compositions
- Screening of spectral data types, sample treatment, and data normalization methods indicated that NIR spectra of water rinsed samples was ideal
- All model performance metrics were optimized using Antaris II with ANN
  - Nanoquest inputs produced comparable results (Fig 3)
- Lipid and protein prediction performance exceeded project-defined thresholds (Table 1)

Materials and Methods

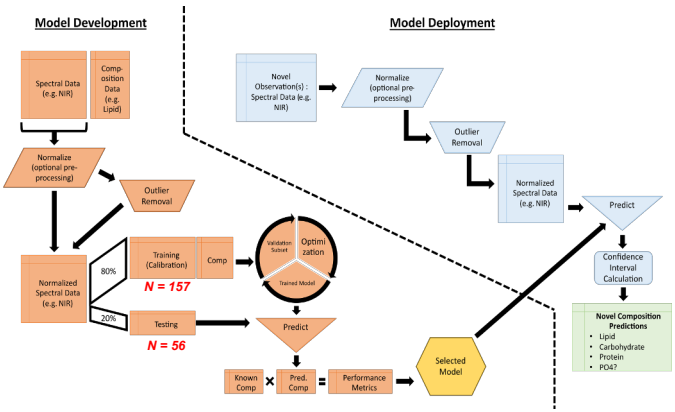

Figure 2. Flowchart demonstrating spectral data analysis method

1. Algae samples were grown, harvested, rinsed, and dried on filters at Global Algae (HI, USA)
2. Standardized chemical analyses were used to quantify each biomass component[4]
3. Replicate scans of filters samples were collected on two NIR spectrometers. Nanoquest (Ocean Insights) and Antaris II (Thermo)
   i. Five additional spectral and fluorescent scan types were collected and screened
4. Spectra were pre-processed in R scripts using standard normal variate (SNV) to normalize data
5. Machine learning algorithms were used to optimize and train models capable of predicting component quantities from NIR inputs (Fig. 2)
   i. Models were produced using partial least squares regression (PLSR) and an artificial neural networks (ANN) implemented in R and Python
   ii. Metrics comparing the accuracy and error of each model against measured values compiled

Conclusions

- NIR spectroscopy can reduce the time needed for compositional analysis to minutes instead of multiple days, using a labor and chemicals-intensive process
- Machine learning algorithms, particularly ANN, produced high accuracy predictive models using inputs from both spectrometers. The significantly lower cost of the Nanoquest instrument presents an opportunity for widespread implementation at field sites
- Possible future work includes building and extending models for broader deployment
  - Collecting additional samples sets of varied algal species under different conditions to expand model predictive space
  - Implement automated spectral outlier detection to ensure highly reliable predictions
  - Developing a user interface on top of existing data processing code to make model predictions available to end users on nascent data at production sites

Acknowledgements/Citations

I would like to thank our Leah Kistner for contributing to spectral data collection, and the algae biomass compositional analysis team at NREL for their efforts in determining biomass composition values.

1. Laurens LML, Dempster TA, Jones HDT, Wolfrum EJ, Van Wychen S, McAllister JSP, et al. Algal Biomass Constituent Analysis: Method Uncertainties and Investigation of the Underlying Measuring Chemistries. Anal Chem. 2012 Feb 21;84(4):1879–87.
2. Burns DA, Ciurczak EW. Handbook of Near-Infrared Analysis. CRC Press; 2007. 836 p.
3. Laurens LML, Wolfrum EJ. High-Throughput Quantitative Biochemical Characterization of Algal Biomass by NIR Spectroscopy; Multiple Linear Regression and Multivariate Linear Regression Analysis. J Agric Food Chem. 2013 Dec 18;61(50):12307–14.
4. Microalgae Compositional Analysis Laboratory Procedures [Internet]. [cited 2022 Oct 3]. Available from: https://www.nrel.gov/bioenergy/microalgae-analysis.html

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Bioenergy Technologies Office (BETO) under grant award number DE-EE0008903. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

International Conference on Algal Biomass, Biofuels and Bioproducts
Waikoloa Beach, Hawaii, USA
(June 12th, 2023)
NREL/PO-2700-86488