Development of protein quantification models for açaí (*Euterpe oleracea*) using fourier-transform mid-infrared spectroscopy (FT-MIR) for applications in biorefinery and circular economy

John W. S. Queiroz[1], Carlos A. R. Barros[1], Salomão F. Silva[1], Nelson R. Ferreira[2], Neirivaldo C. Silva[3], Herve Hogez[3] & Fábio G. Moura[3*]

[1]*Postgraduate Program in Biotechnology, Institute of Biological Sciences (ICB), Federal University of Pará (UFPA), Belém, PA, Brazil*
[2]*Food Engineering Course, Institute of Technology (ITEC), Federal University of Pará (UFPA), Belém, PA, Brazil.*
[3]*Centre for Valorization of Amazonian Bioactive Compounds (CVACBA), Federal University of Pará (UFPA), Belém, PA, Brazil.*
**Corresponding author's email address: famoura@ufpa.br*

ABSTRACT

The study explored Fourier-Transform Mid-Infrared spectroscopy (FT-MIR) potential in protein determination in defatted, freeze-dried açai (*Euterpe oleracea*). Using the full spectral range (4000-650 cm-1), preprocessing, particularly with SNV, yielded promising results in amide I and II regions (1750-1400 cm-1). VIP outperformed iPLS and sRatio, achieving superior prediction (R2 = 0.66) with SNV and smoothing. FT-MIR showed significant potential for açai protein prediction. VIP highlighted relevant spectral ranges (650-830 cm-1, 1000-1200 cm-1, 1500-1780 cm-1, 2840-2970 cm-1), showing promise as an effective tool for protein composition evaluation in açai. Efficient process design can further enhance these findings by maximizing resource utilization while minimizing water, energy, and chemical inputs.

Keywords: MIR. Protein. Açaí pulp. Model development.

1 INTRODUCTION

Açaí (*Euterpe oleracea*) has gained prominence in the global food industry. Originating from the Amazon region, its popularity is attributed to bioactive compounds such anthocyanins (1,365.2 mg/kg fruit) and α-tocopherol (45 mg/100g total solids - TS). Nutritionally, açaí boasts lipids (32.5 to 50.5% TS), total dietary fiber (20.9 to 21.8% TS), and proteins (8.1 to 12% TS) as its main constituents. The booming açaí industry, driven by global demand, highlights the crucial need for quality control to maintain product authenticity and integrity. Advanced analytical methods are essential for consistent quality and nutritional assurance in açaí-based products.[1]

Traditional methods for determining protein content, like Kjeldahl, UV absorption, and HPLC, are laborious and time-consuming. Mid-infrared spectroscopy (MIR) is becoming popular as a faster and more affordable option for food analysis, offering insights into molecular vibrations.[2] However, there's currently no specific MIR method for assessing protein content in açai pulp. This research aims to create quantification models for protein in açai pulp, especially in freeze-dried and defatted forms. Using chemometrics, the study utilizes analytical measurements from reference methods to establish mathematical relationships via multivariate classification and calibration. These relationships enable the deduction of protein content values through indirect observations using infrared spectroscopy.

2 MATERIAL & METHODS

2.1 Samples and Spectral Acquisition

Açaí fruits (9 kg) were sourced from local farmers in twenty-six municipalities in the state of Pará, Brazil, and transported under refrigeration (4°C) to the Centre for Valorization of Amazonian Bioactive Compounds (CVACBA). The processing involved washing, softening, pulping, and freeze-drying for 48 hours, followed by Kjeldahl analysis for protein determination. Additionally, all samples underwent oven drying at 105°C for 16 hours to minimize water content and avoid interference in spectroscopic analyses. The total number of samples was sixty-one (61).

Instrumentation and Spectral Acquisition utilized the Agilent Cary 630 FTIR-ATR equipment with a ZnSe crystal. Spectral acquisition employed an optical resolution of 4 cm^{-1} with a 10-second acquisition time, and 32 scans were conducted for each spectrum within a spectral range of 4000 to 650 cm^{-1}, generating 900 variables. The standard data acquisition software (MicroLab) facilitated crystal cleaning configuration and validation of cleaning efficiency. After placing the sample on the crystal, a reading was performed, presenting the spectrum result.

2.2 Pre-treatments and spectral analysis

Pre-treatments and spectral analysis involved exporting raw protein reference values and spectral data to MATLAB (R2023a, The Mathworks® Software, USA), and processing them with the PLS-Toolbox software (V. 9.2.1, Eigenvector Research Inc., USA). Raw spectra underwent 12 pre-processing algorithms, including detrending, noise reduction smoothing, MSC, SNV normalization,

FD, SD, and combinations like SNV-Detrending, FD-Smoothing, SD-Smoothing, SNV-FD, SNV-SD, and SNV-FD-Smoothing. Principal Component Analysis (PCA) was applied to both pre-processed and raw spectra to detect outliers and identify spectral patterns using Hotelling's T^2 and F-residual plots. Outliers exceeding the 5% limit were removed, and the remaining samples were split into calibration (70%) and prediction (30%) sets.[3-4] Sample categorization utilized the Kennard-Stone (KS) algorithm to assess cumulative probability distributions.

MIR models were constructed using PLS regression across all wavelength regions, with focus on the amides region (1750 cm-1 – 1400 cm-1), employing sample selection via the KS algorithm. iPLS and VIP algorithms were also assessed for variable selection. Model evaluation included examination of calibration (R_c^2), cross-validation (R_{CV}^2), and prediction (R_P^2) coefficients of determination, as well as root mean square errors for calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP). Additionally, errors for calibration (SEC), cross-validation (SECV), and prediction (SEP) were thoroughly addressed.[3-4]

3 RESULTS & DISCUSSION

3.1 Development of MIR models and performance

During sample preparation and spectral acquisition, interferences like noise and irrelevant background information may occur. Thus, pre-processing procedures are crucial for spectral data. These steps aim to remove noise and unwanted background, enhancing chemical signals.[5] In the MIR spectral region, significant values were found in the ranges of 1720-1650 cm^{-1} and 1635-1550 cm^{-1}, corresponding to the amide I and amide II regions, respectively. These bands, especially in the amide I region, are essential for developing PLS models. However, the amide II region has limited utility due to the complexity associated with various functional groups. The protein prediction data for açaí levels are shown in Figure 1.

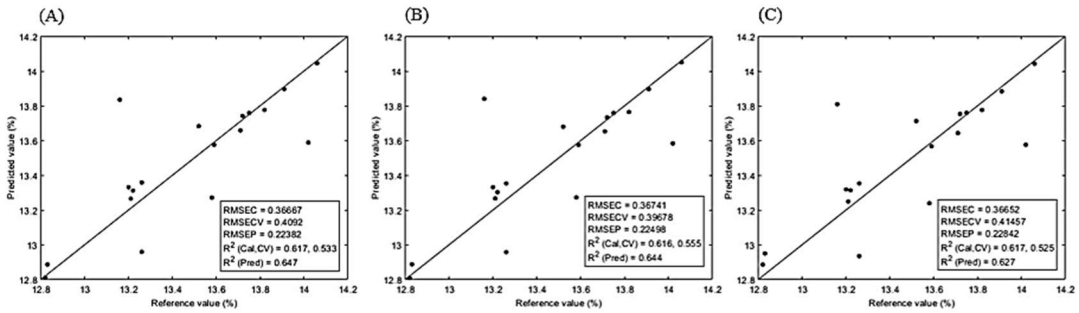

Fig. 1. Predicted and reference values for the most robust MIR prediction models protein in açai pulp using the spectral range of 4000-650 cm^{-1}. (A) SNV; (B) SNV-Detrending; (C) SNV-FD-SM.

The most effective MIR models covered the entire spectral range without specific cuts in the amide region (4000-650 cm^{-1}). Among the top three models (Figure 1) – SNV, SNV-Detrending, and SNV-FD-SM – all showed similar prediction values and RMSE indices for both datasets. An F-test confirmed no statistically significant differences between these models, indicating their consistent and equivalent predictive performance across the datasets.

3.2 Variable selection algorithms

This study employed the iPLS method using 15 intervals, with each interval representing a selected subset of variables. These intervals, totaling 21 variables, were chosen to provide more efficient prediction compared to enhance prediction efficiency compared to using all available variables in the dataset. Additionally, VIP scores were calculated with the removal of 0.1 fraction per iteration. The results of these selections were visualized in Figure 2, providing a graphical representation of the variables chosen by the respective algorithms. These methodological approaches aim to optimize prediction effectiveness by considering the importance of the selected variables within the study's context.

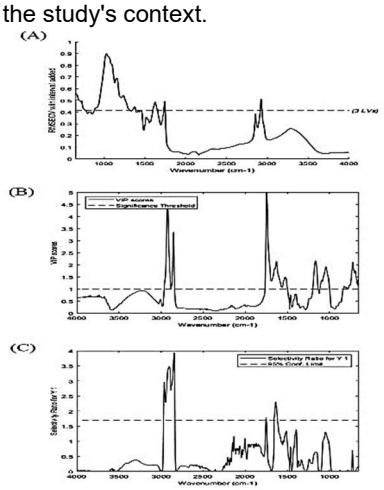

Fig. 2. Variable selection: (A) iPLS; (B) VIP scores; and (C) Selectivity using sRatios.

As shown in Figure 2A, the intervals that contributed to the increase in RMSECV values were identified in the iPLS graph and subsequently excluded during the algorithm cycles. On the other hand, the spectral distances represented the variables that remained in the last cycle of the iPLS algorithm solution. The importance of the distances concentrated in the spectral region above the pre-established limit was highlighted, considering the use of 3 LVs for the effective performance of iPLS. VIP scores were computed to assess the relevance of each variable in the projection employed in a PLS model and were compared with the iPLS and sRatio algorithms. As illustrated in Figure 2B, variables with VIP scores close to or higher than a certain threshold value, and those with VIP scores significantly lower than one, were classified as less relevant, suggesting their exclusion from the model.[6] In the VIP method, spectral ranges from 650 to 830 cm^{-1}, 1000 to 1200 cm^{-1}, 1500 to 1780 cm^{-1}, and 2840 to 2970 cm^{-1} were identified as relevant. The results obtained using the sRatio algorithms were not satisfactory, as shown in Table 1.

Table 1. PLS parameters used for protein models in açai pulp were built using variables selected by algorithms.

Pre-processing	Selection algorithm	Calibration		Cross-validation		Prediction	
		R_c^2	RMSEC	R_{cv}^2	RMSECV	R_p^2	RMSEP
SNV	iPLS	0.42	0.36	0.47	0.41	0.61	0.30
	VIP	0.63	0.33	0.55	0.39	0.64	0.22
	sRatio	0.25	0.51	0.17	0.53	0.26	0.32
SNV-Detrending	iPLS	0.53	0.40	0.34	0.49	0.57	0.36
	VIP	0.60	0.35	0.55	0.39	0.61	0.30
	sRatio	0.26	0.50	0.19	0.53	0.29	0.46
SNV-FD-SM	iPLS	0.57	0.38	0.50	0.41	0.58	0.35
	VIP	0.65	0.28	0.60	0.34	0.67	0.19
	sRatio	0.33	0.47	0.26	0.50	0.34	0.43

Abbreviation: R_c^2, coefficient of determination for calibration; RMSEC, root mean square error of calibration; R_{cv}^2, coefficient of determination for cross-validation; RMSECV, root mean square error for cross-validation; R_p^2 , coefficient of determination for prediction; RMSEP, root mean square error for prediction; SM, noise smoothing; MSC, multiplicative scatter correction; SNV, standard normal variate normalization; FD, first derivative; SD, second derivative; SNV-Detrending, SNV + Detrending; FD-SM, first derivative + smoothing; SD-SM, second derivative + smoothing; FD-SNV, first derivative + SNV; SD-SNV, second derivative + SNV; SNV-FD-SM, SNV + first derivative + smoothing.

The algorithms were tested following the same preprocessing protocol used in the previous models. However, the preprocessing trend that prevailed earlier also remained with the best results within these new tests. It was observed that when using iPLS, the results were not superior to the use of the complete spectrum; however, when employing the variables selected by VIP models, similar results were obtained. By using SNV-FD-SM, we achieved the highest R_c^2, value (0.65), which was also repeated when evaluating the prediction value (0.67). Additionally, there was an improvement in the results of RMSEC, RMSECV, and RMSEP, respectively. These results are in line with the study by the ability of VIP to highlight more informative variables provides a more precise selection of relevant spectral features, resulting in more robust and accurate models. The emphasis on crucial variables not only enhances the predictive ability of the model but also contributes to its simplification, avoiding overfitting and promoting better generalization to unseen data, thus substantiating its relevance in optimizing chemometric models

4 CONCLUSION

The study focused on characterizing protein in defatted, freeze-dried açai, revealing moderate protein content variations through statistical analysis. MIR investigation emphasized the importance of preprocessing techniques like SNV and detrending for improving spectral data quality. Development of MIR models via PLS regression highlighted the significance of meticulous preprocessing, with the VIP algorithm yielding optimal results by selecting crucial spectral features. Comparison of techniques underscored the impact of preprocessing on model performance, with SNV, SNV-Detrending, and SNV-FD-SM models exhibiting equivalent statistical consistency. Future integration of MIR spectroscopy with optimized VIP approaches promises enhanced accuracy in determining protein levels in freeze-dried açai, offering valuable insights for advancing spectroscopic methods in complex sample analysis. Efficient process design, aimed at maximizing resource utilization while minimizing water, energy, and chemical inputs, can further enhance these findings.

REFERENCES

[1] Bichara, C. M. G., Rogez, H. 2011. Açai (*Euterpe oleracea* Mart.). In: Postharvest Biology and Technology of Tropical and Subtropical Fruits: Açai to Citrus. Yahia, E. M. (ed). 2nd ed. Woodhead Publishing, Cambridge. 1–27

[2] Sturaro, A., De Marchi, M., Masi, A., Cassandro. M. 2016. J. Dairy Sci., 99 (1), 68-76.

[3] Shi, H., Lei, Y., Prates, L. L., Yu, P. 2019. Food Chem. 272, 507-513.

[4] G.E. Morlock, M. Busso, S. Tomeba, A. Sighicelli. 2021. Journal of Chromatography. A, 1652, 462377.

[5] Hasegawa, K., Funatsu, K. 2010. Curr Comput-Aid Drug. 6 (2), 103-127

[6] Galindo-Prieto, B., Eriksson, L., Trygg, J. 2015. Chemometr Intell Lab Syst, 146, 297-304.