Classification Tools for Carotenoid Content Estimation in *Manihot esculenta* via Metabolomics and Machine Learning

Rodolfo Moresco[1(✉)], Telma Afonso[3], Virgílio G. Uarrota[1], Bruno Bachiega Navarro[1], Eduardo da C. Nunes[2], Miguel Rocha[3], and Marcelo Maraschin[1]

[1] Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianopolis, Brazil
rodolfo_moresco@yahoo.com.br
[2] Santa Catarina State Agricultural Research and Rural Extension Agency (EPAGRI), Experimental Station of Urussanga, Urussanga, Brazil
[3] Centre Biological Engineering, School of Engineering, University of Minho, Braga, Portugal

Abstract. Cassava genotypes (*Manihot esculenta* Crantz) with high pro-vitamin A activity have been identified as a strategy to reduce the prevalence of deficiency of this vitamin. The color variability of cassava roots, which can vary from white to red, is related to the presence of several carotenoid pigments. The present study has shown how CIELAB color measurement on cassava roots tissue can be used as a non-destructive and very fast technique to quantify the levels of carotenoids in cassava root samples, avoiding the use of more expensive analytical techniques for compound quantification, such as UV-visible spectrophotometry and the HPLC. For this, we used machine learning techniques, associating the colorimetric data (CIELAB) with the data obtained by UV-vis and HPLC, to obtain models of prediction of carotenoids for this type of biomass. Best values of R^2 (above 90%) were observed for the predictive variable TCC determined by UV-vis spectrophotometry. When we tested the machine learning models using the CIELAB values as inputs, for the total carotenoids contents quantified by HPLC, the Partial Least Squares (PLS), Support Vector Machines, and Elastic Net models presented the best values of R^2 (above 40%) and Root-Mean-Square Error (RMSE). For the carotenoid quantification by UV-vis spectrophotometry, R^2 (around 60%) and RMSE values (around 6.5) are more satisfactory. Ridge regression and Elastic Network showed the best results. It can be concluded that the use colorimetric technique (CIELAB) associated with UV-vis/HPLC and statistical techniques of prognostic analysis through machine learning can predict the content of total carotenoids in these samples, with good precision and accuracy.

Keywords: Chemometrics · Descriptive models · Machine learning · Cassava genotypes · Carotenoids · HPLC · UV-vis

F. Fdez-Riverola et al. (eds.), *11th International Conference on Practical Applications of Computational Biology & Bioinformatics*, Advances in Intelligent Systems and Computing 616, DOI 10.1007/978-3-319-60816-7_34

1 Introduction

Carotenoids refer to the most important natural pigments, being found in all photosynthetic organisms, with colors varying between yellow and dark-red. One of the most important trait of carotenoids is their physiological function as vitamin A precursors to animals [1]. Vitamin A deficiency is a leading cause of morbidity and mortality, especially in young children and pregnant and lactating women. Food-based interventions focused on alleviating vitamin A deficiency in susceptible populations have advantages over supplementation and fortification programs, especially in rural areas, because they can provide a sustainable source of a variety of nutrients and other phytochemicals without the recurring transport and administration costs of these other methods [2]. It is estimated that among all known carotenoids, about 50 can act as precursors of vitamin A in mammals. However, only α-carotene, β-carotene, γ-carotene, and β-cryptoxanthin are common in fruits and vegetables [3]. Cassava genotypes with high contents of pro-vitamin A carotenoids have been identified as a strategy to reduce the prevalence of deficiency of this vitamin [4].

The cassava crops are characterized by the color variability of their roots, which can vary from white to red. The color is related to the presence of several carotenoid pigments, their associations and contents [5]. However, the possibility of adopting the color of roots as an indirect criterion for selection of higher carotene content is questionable, since color is a characteristic of difficult visual evaluation.

In order to standardize color measurements, the CIE (Commission Internationale de L'Eclairage) recommended the use of the CIE L* a* b* or CIELAB color scale. It is currently the most used system for quantitative color description of an object, due to its uniformity, ease of acquisition, and very low cost technique [6].

Chemical extraction followed by the identification and quantification of carotenoid pigments, especially by UV-vis spectrophotometry and high performance liquid chromatography (HPLC) are very accurate, but extremely expensive, also requiring a long time for the analysis. The CIELAB color measurement is a non-destructive and very fast technique, which allows to obtain a series of parameters, in a few seconds. Thereby, it facilitates performing measurement in the field, avoiding the degradation of these compounds in consequence of their chemical extraction, for instance.

The aim of this work is to validate a quantification method for carotenoid contents in roots of *M. esculenta* from colorimetric data using the CIE L* a* b* system, assuming that the statistical techniques of prognostic analysis, as well as machine learning, can correlate colorimetric data easily obtained in the field, with the contents obtained through traditional techniques, e.g., UV-vis spectrophotometry and HPLC and, from this, construct prediction models of carotenoids content for cassava roots. This study applies analytical techniques and bioinformatics tools to detect genotypes of *M. esculenta* with high levels of carotenoids. In addition, it provides tools that can support the plant-breeding program at Epagri (Agricultural Research Company and Rural Extension of the State of Santa Catarina- http://www.epagri.sc.gov.br/) that aims to obtain genotypes with high levels of pro-vitamin A carotenoids and superior nutritional traits.

2 Materials and Methods

Roots of fifty genotypes of *M. esculenta* (2015/2016 season) from the EPAGRI's germplasm bank (Urussanga Experimental Station, 28º31′18″S, 49º19′03″W, Santa Catarina, southern Brazil) were used in this study due to their economic and social importance.

Carotenoids were extracted from fresh roots as described by Rodriguez-Amaya & Kimura (2004) [7]. The absorbances of the organosolvent extracts were recorded on an UV-vis spectrophotometer (Gold Spectrum lab 53 UV-Vis spectrophotometer, BEL photonics, Brazil) over a spectral window from 200 to 700 ηm. Aliquots (10 μl) of the extracts were also injected into a liquid chromatograph (LC-10A Shimadzu) system equipped with a C18 reversed-phase column (Vydac 201TP54, 250 mm × 4.6 mm, 5 μm Ø, 35°C) coupled to a pre-column (C18 Vydac 201TP54, 30 mm × 4.6 mm, 5 μm Ø) and a spectrophotometric detector (450 nm). Methanol: acetonitrile (90: 10, v/v) was used for elution at a rate of 1 ml/min.

The color attributes of the roots samples were measured by a colorimeter (CR-400, Minolta, Japan) immediately after harvest and the results were expressed according to the CIELAB color space scale [4]. Three readings were performed at different sites in fifty samples. Data were collected, summarized, and submitted to analysis of variance (ANOVA) followed by the *post-hoc* Tukey's test ($p < 0.05$) for mean comparison. Spectrophotometric data and the amounts of the target carotenoids determined by HPLC were treated using multivariate statistical analysis and chemometrics techniques, supported by scripts written in R language (v. 3.3.1) [8]. Additionally, we used prognostic tools through machine learning techniques, associating the colorimetric data (CIELAB) with the data obtained by UV-vis and HPLC, to obtain models of prediction of carotenoids for this type of biomass and technique.

The data analysis was supported and structured using the R *specmine* package [9] developed by our research team for metabolomics studies that includes a number of machine learning methods implemented through the package *caret* [10]. In supplementary material, provided in http://darwin.di.uminho.pt/pacbb2017/cassava-carotenoids, we include the data analysis reports automatically generated from the R scripts using the features provided by R Markdown, as well as the respective data and metadata files. This allows fully understanding and reproducing the computational experiments.

3 Results and Discussion

The values of the carotenoid quantification through UV-vis spectrophotometry and HPLC are given in the metadata of the dataset. The roots white-colored pulp presented the lowest concentrations of total carotenoids (values from 0.57 $\mu g.g^{-1}$), while highest concentrations were observed in genotypes with pigmented pulp (yellow and red) roots, i.e., 54.93 $\mu g.g^{-1}$. These results are consistent with data reported in the literature that observe a positive relation between the color of the root pulp and the total content of those pigments [11, 12]. The contents of the major carotenoid compounds, *trans*-β-carotene and *cis*-β-carotene, ranged from 1.82 to 42.82 $\mu g.g^{-1}$ for *trans*-β-carotene and 1.19 to 28.86 $\mu g.g^{-1}$ for *cis*-β-carotene.

The visual interpretation of the sample's location in the CIELAB' space is enough to verify which samples have higher levels of carotenoids [13]. Figure 1 shows the samples location according to the color of roots, in the CIE L* a* b* plane. Samples 105, 119 and 125 (Fig. 1 - ellipse I) contain the highest levels of total carotenoid. The sample 74, due to its reddish color, was represented in the CIELAB space on the positive axis (Fig. 1 - ellipse IV), mostly due to its lycopene contents, which confer reddish coloration to the roots [14]. Samples with lower amounts of carotenoids (123, 27, 05, AO47) shown values of b* closer to zero (ellipse III), while those with medium contents were grouped in a* negative and b* positive (ellipse II).

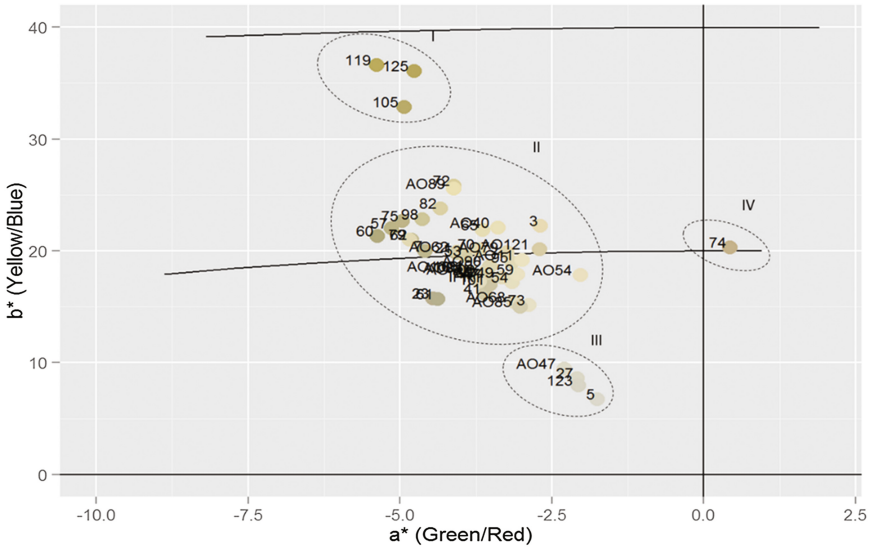

Fig. 1. Location of the cassava samples in the CIE L* a* b* plane according to their root pulp colors. The a* value characterizes the coloration in the regions of red (+a*) to green (−a*). The b*, value b* indicates coloring in the range of yellow (+b*) to blue (−b*). The L indicates the luminosity, varying from white (L = 100) to black (L = 0).

The next step of this work was to correlate the colorimetric data obtainable in the field (CIELAB) with the contents found by traditional techniques, e.g., UV-vis spectrophotometry and HPLC, through statistical techniques of prognostic analysis such as machine learning. From this, we constructed a set of carotenoid concentration predictive regression models for this type of biomass using the information from the samples' color values and the UV-vis spectra.

The *specmine* package provides a number of functions to train, use, and evaluate machine learning methods, being mostly based in the R package *caret* [10], covering both classification and regression methods. In addition, there are functions to evaluate the importance of each variable in the models. A list of possible models and tunable parameters can be seen in https://topepo.github.io/caret/available-models.html.

The implemented functions enable executing model training and can be used to predict new data posteriorly. Also, it is possible to optimize a set of model parameters testing a set of possible values and evaluating those according to the selected validation method and metric errors. The CIELAB data were considered as continuous variables. In this way, regression-derived statistical data mining models (5-fold cross-validation repeated 10 times, testing all models with feature selection with 80, 60, and 40% data filtering) were used, such as Least Absolute Shrinkage and Selection Operator (Lasso) [15], Ridge Regression [16], Elastic Net Regression (Enet) [17], Decision Trees/Randon Forest (RF) [18], Partial Least Squares (PLS), Artificial Neural Net (NNs), and Support Vector Machines (SVMs). These validation methods are available to estimate the metric errors, and usually the decision is based on simple criteria based on the residual values. The chosen evaluation metrics to compare model performance were the Root-Mean-Square Error (RMSE) and the coefficient of determination (R^2), since they explicitly show how much the model predictions deviate, on average, from the actual values in the dataset.

Table 1 shows the performance values of a set of machine learning regression models (RMSE and R^2) associating UV-vis scanning spectrophotometry in the typical region of fingerprint for carotenoids (400–500 nm) as inputs, with the total carotenoids contents determined by HPLC (TCC HPLC), total carotenoids contents determined by UV-vis spectrophotometry (Lambert-Beer formula), and the majoritarian carotenoid found in cassava roots (*trans*-β-carotene), each predicted as an output in distinct experiments using the different methods (details are given in the reports in supplementary materials).

It can be verified that the best R^2 values (>90%) were observed for the predictive variable TCC, determined by UV-vis spectrophotometry. These values were higher than the predictive variables *trans*-β-carotene (best model with R^2 47%) and total carotenoids contents determined by HPLC (with values of R^2 around 60%). This is expected, since they are methodologies that employ the same physical phenomenon of detection of compounds (absorbance). When observed the values of variable importance in this analysis (supplementary material), it can be detected that the wavelength at 450 nm (precisely the wavelength that is used for the quantification of β-carotene through the Lambert-Beer formula) was the most prevalent. This result is important because it attests to the robustness of the models in predicting the contents of these compounds in these samples.

Then we tested the machine learning models using the CIELAB values as inputs, with the same outputs as before. For the total carotenoids contents quantified by HPLC, the Partial Least Squares (PLS), Support Vector Machines (kernlab), and Elastic Net models presented the best values of R^2 and lower values of RMSE (Table 2). It can be verified that these values are smaller than when the inputs are the UV-vis (400–500 nm) data (Table 1). This is due to the fact that the colorimetric and chromatographic techniques are different in their physicochemical bases, and the UV-vis data has many more variables measured.

Table 1. Performance values (RMSE and R^2) associating UV-vis scanning spectrophotometry (400–500 nm) with the total carotenoids contents determined by HPLC (TCC HPLC), total carotenoids contents determined by Lambert-Beer formula (TCC Spectrophotometry), and the majoritarian carotenoids of cassava roots samples (*trans*-β-carotene).

	UV-vis. 400–500 nm					
	TCC Spectrophotometry		TCC HPLC		*trans*-β-carotene	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Partial Least Squares (simpls)	**3.492**	**0.920**	5.789	0.572	4.309	0.362
Support Vector Machines (e1071)	**3.709**	**0.931**	5.844	0.597	4.218	0.399
PLS (widekernelpls)	**3.732**	**0.923**	5.779	0.570	4.324	0.453
Random Forest	**3.768**	**0.948**	7.275	0.359	5.753	0.239
Elastic Net	3.793	0.918	**5.934**	**0.634**	4.191	0.412
Partial Least Squares (pls)	3.800	0.952	**5.643**	**0.597**	**4.265**	**0.470**
Ridge Regression (w/FS)	3.855	0.947	**5.880**	**0.603**	4.159	0.356
Ridge Regression	3.877	0.928	7.282	0.616	4.407	0.316
SVM (kernlab)	3.928	0.940	5.907	0.589	4.230	0.466
PLS (kernelpls)	4.096	0.896	5.878	0.566	4.211	0.422
Linear Regression (Stepwise)	4.158	0.919	8.341	0.526	6.135	0.206
Linear Regression (Forward)	4.178	0.888	8.783	0.471	5.142	0.311
Linear Regression (Backwards)	4.392	0.871	6.373	0.522	5.355	0.278
K-Nearest Neighbors	4.732	0.922	6.277	0.445	4.597	0.224
Lasso	5.207	0.817	17.508	0.249	16.145	0.189
Conditional Inference RF	6.713	0.791	6.806	0.558	4.703	0.369
Conditional Inference Tree	7.363	0.711	6.916	0.480	4.894	0.288
Decision Trees	7.582	0.683	6.795	0.473	5.189	0.053

When the CIELAB values were used to predict the values of carotenoid contents by UV-vis spectrophotometry, R^2 and RMSE values were more satisfactory. Ridge regression and Elastic Network showed the best results. Observing the importance of the variables in the prediction (supplementary material), it can be verified that the values of b* were more relevant. In the CIELAB space, the value b^* indicates coloration in the range from yellow ($+b^*$) to blue ($-b^*$), an important finding since most carotenoids confer yellowish pigmentation in foods, associating their pro-vitamin A activity.

Table 2. Performance values (RMSE and R^2) associating CIELAB colorimetric data with the total carotenoids contents determined by Lambert-Beer formula (TCC Spectrophotometry), total carotenoids contents determined by HPLC (TCC HPLC), and the content of the majoritarian carotenoid found in cassava roots samples (*trans*-β-carotene).

	CIELAB Data					
	TCC Spectrophotometry		TCC HPLC		trans-β-carotene	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Partial Least Squares (simpls)	7.043	0.543	6.789	0.414	4.781	0.194
Support Vector Machines (e1071)	7.136	0.500	6.645	0.380	4.800	0.155
PLS (widekernelpls)	6.771	0.541	6.696	0.396	4.857	0.170
Random Forest	7.280	0.448	7.571	0.293	5.393	0.149
Elastic Net	**6.515**	**0.573**	**6.534**	**0.412**	**4.690**	**0.212**
Partial Least Squares (pls)	7.085	0.538	6.622	0.394	4.859	0.164
Ridge Regression (w/FS)	**6.469**	**0.608**	6.653	0.389	4.951	0.238
Ridge Regression	**6.497**	**0.590**	**6.584**	**0.421**	4.848	0.238
SVM (kernlab)	6.919	0.528	**6.534**	**0.366**	**4.745**	**0.201**
Partial Least Squares (kernelpls)	6.865	0.540	6.756	0.431	4.815	0.162
Linear Regression	6.651	0.558	6.749	0.400	4.945	0.220
K-Nearest Neighbors	7.267	0.525	7.278	0.256	4.956	0.153
Lasso	6.757	0.575	6.669	0.411	4.793	0.182
Conditional Inference RF	8.021	0.454	6.930	0.408	4.782	0.223
Conditional Inference Tree	9.636	0.339	7.307	0.384	4.929	0.130
Decision Trees	9.737	0.316	7.641	0.353	5.000	0.297

These results are very promising because they enable CIELAB technique as an alternative for measuring carotenoids in cassava roots to the use of more expensive analytical techniques such as UV-vis spectrophotometry and HPLC. Thus, it has been shown that the concomitant use of UV-vis and color (CIELAB) techniques with statistical techniques of prognostic analysis (i.e., machine learning) can predict the content of total carotenoids in cassava roots, with good precision and accuracy and low metrical error.

4 Conclusions

The present study has shown how CIELAB color measurement can be used as a fast and non-destructive method to calibrate for the total carotenoid content of cassava genotypes roots with acceptable prediction error. In addition, the information obtained by coupling the analysis of pro-vitamin A biochemical markers to bioinformatics tools helps supporting the rational design of biochemically-assisted breeding programs of *M. esculenta*, that aims to obtain cultivars with high levels of pro-vitamin A carotenoids and superior nutritional traits.

Acknowledgements. To CNPq (National Counsel of Technological and Scientific Development) for financial support (Process no. 407323/2013-9), to CAPES (Coordination for the Improvement of Higher Education Personnel (CAPES), and EPAGRI (Agricultural Research and Rural Extension Company of Santa Catarina). The research fellowship from CNPq on behalf of M. Maraschin is acknowledged. The work is partially funded by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq.

References

1. Rodriguez-Amaya, D.B.: A Guide to Carotenoid Analysis in Foods (2001)
2. Tanumihardjo, S.A., Palacios, N., Pixley, K.V.: Provitamin a carotenoid bioavailability: what really matters? Int. J. Vitam. Nutr. Res. **80**, 336–350 (2010)
3. Stahl, W., Sies, H.: Antioxidant activity of carotenoids. Mol. Aspects Med. **24**, 345–351 (2003)
4. La Frano, M.R., Woodhouse, L.R., Burnett, D.J., Burri, B.J.: Biofortified cassava increases β-carotene and vitamin A concentrations in the TAG-rich plasma layer of American women. Br. J. Nutr. **110**, 310–320 (2013)
5. Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., Zum Felde, T., Domínguez, M., Davrieux, F.: Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. Food Chem. **151**, 444–451 (2014)
6. CIE: The Evaluation of Whiteness. Color, 3rd edn., vol. 552, p. 24 (2004)
7. Rodriguez-Amaya, D., Kimura, M.: HarvestPlus handbook for carotenoid analysis. Harvest. Tech. Monogr. **59**, 525–528 (2004)
8. R Core Team: R: A Language and Environment for Statistical Computing (2014). http://www.r-project.org/
9. Costa, C., Maraschin, M., Rocha, M.: An R package for the integrated analysis of metabolomics and spectral data. Comput. Methods Programs Biomed. **129**, 117–124 (2015)
10. Max, A., Contributions, K., Weston, S., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C.: Package "caret". Max Kuhn (2016)
11. Champagne, A., Bernillon, S., Moing, A., Rolin, D., Legendre, L., Lebot, V.: Carotenoid profiling of tropical root crop chemotypes from Vanuatu. South Pacific. J. Food Compos. Anal. **23**, 763–771 (2010)
12. Chávez, A.L., Sánchez, T., Jaramillo, G., Bedoya, J.M., Echeverry, J., Bolaños, E., Ceballos, H., Iglesias, C.: Variation of quality traits in cassava roots evaluated in landraces and improved clones. Euphytica **143**, 125–133 (2005)

13. Kljak, K., Grbeša, D., Karolyi, D.: Reflectance colorimetry as a simple method for estimating carotenoid content in maize grain. J. Cereal Sci. **59**, 109–111 (2014)
14. Meléndez-Martínez, A.J., Britton, G., Vicario, I.M., Heredia, F.J.: Relationship between the colour and the chemical structure of carotenoid pigments. Food Chem. **101**, 1145–1150 (2006)
15. Tibshirani, R.: Regression Selection and Shrinkage via the Lasso (1994). http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574
16. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)
17. Zou, H.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Series B **67**, 301–320 (2005)
18. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R.P., Feuston, B.P.: Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. **43**, 1947–1958 (2003)