Microchemical Journal 141 (2018) 264–270

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

NMR and multivariate data analysis to assess traceability of argentine citrus

Mario O. Salazar[a,b,*], Pablo L. Pisano[c], Manuel González Sierra[d,1], Ricardo L.E. Furlan[a,b]

[a] Instituto de Investigaciones para el Descubrimiento de Fármacos de Rosario (IIDEFAR-CONICET), Ocampo y Esmeralda, Rosario S2002LRK, Argentina
[b] Farmacognosia, Departamento de Química Orgánica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario S2002LRK, Argentina.
[c] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario S2002LRK, Argentina.
[d] Departamento de Química Orgánica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario S2002LRK, Argentina.

ARTICLE INFO

Keywords:
Orange discrimination
Quality control
Nuclear magnetic resonance
Icoshift algorithm
Principal component analysis
Partial least-squares – discriminant analysis

ABSTRACT

Argentine citrus industry stands out among the top ten largest producers of citrus of the world. Traceability is a key point of the quality control in the international trade context. Discrimination between orange juices produced from different geographical regions in Argentina (San Pedro and Entre Ríos), has been achieved by applying principal components analysis (PCA) and partial least-squares – discriminant analysis (PLS–DA) to ^{1}H NMR spectra of the juices. Different regions of the spectra were analyzed in view of the large range of signal intensities and the chemometric treatment of the NMR data has successfully distinguished the juice of different origins/varieties, as well as the metabolites responsible for their separation. Examination of the PCA loadings showed that citric acid and ethanol levels were the main chemical variables for sample discrimination.

1. Introduction

Citrus is one of the world's major fruit crops, with global availability and popularity contributing to human diets [1]. Citrus production is very important within the Argentine fruit growing, and represents a significant source of revenue, both from domestic and international trade (exports are mainly sent to Europe, Russia and Paraguay). With around 130,000 ha cultivated from the northeast to the northwest, and > 2.6 million tons produced per year, Argentine is among the ten largest citrus producers of the world. This resulted in a large number of jobs (> 100,000) [2].

Quality control is a constant challenge in food industry with respect to contamination and fraud, like wrong labelling of the product type or the type and origin of ingredients. One of the main current issues that are presented to the Argentine official control institution (e.g.: *Servicio Nacional de Sanidad y Calidad Agroalimentaria*, *SENASA*) is determining the geographical origin of citrus. Citrus traceability is vital since the production area is directly related to its organoleptic characteristics and, consequently, to their acceptability by consumers and commercial value [3]. Oranges are produced in the northeastern and northwestern regions of Argentina. The northeastern region, which includes San Pedro and Entre Ríos, is the main supplier for the international market. Characteristic climate and soil conditions in this region facilitate the production of different orange varieties. The Salustiana variety from Entre Ríos and Ombligo from San Pedro are the most required for their organoleptic characteristics (sweet/acid ratio) in the international market. The assurance of origin and variety is important in order to meet the demands that each market requires. The National system of traceability in Argentina involves a rigorous registration methodology of the whole process of citrus from the field to the final destination. Chemical analysis data are not used at all for this purpose.

Several instrumental techniques have been evaluated for determining the geographical origin of food products, including mass spectrometry, spectroscopy, separation techniques, etc. [4–8]. Some techniques such as isotopic composition of orange sugars [9], GC–MS [10], fluorescence spectroscopy [11], HPLC-UV [12], HPLC-MS [13] inductively coupled plasma [14], neutron activation analysis [15] and near-infrared spectroscopy [16] among others have demonstrated to be particularly useful for citrus classification.

Today, NMR is a major tool in a wide range of metabonomics related applications among which food science is included [17]. NMR can detect a wide range of different compounds in one sample run. Furthermore, is non-destructive, rapid, reproducible, and stable over time, requiring only a very simple sample preparation. ^{1}H NMR spectra of mixtures are very rich in information, thus its combination with chemometric analysis can reveal latent patterns in the data, which may

[*] Corresponding author at: Instituto de Investigaciones para el Descubrimiento de Fármacos de Rosario (IIDEFAR-CONICET), Ocampo y Esmeralda, Rosario S2002LRK, Argentina.
E-mail address: salazar@iidefar-conicet.gob.ar (M.O. Salazar).
[1] It has retired.

https://doi.org/10.1016/j.microc.2018.05.037
Received 9 February 2018; Received in revised form 4 April 2018; Accepted 21 May 2018
Available online 22 May 2018

enable classification of the samples in terms of varietal, geographical origin, adulteration, etc. Other analytical techniques (GCMS, HPLC, among other) may be cheaper than NMR, however they usually measure only compounds from one specific chemical class (i.e., volatile compounds, sugars, flavonoids, amino acids, etc.). These methods are appropriate for target analysis. In 2007, the control organisms of European community (one of the leading markets for Argentine citrus) started to use systematically NMR based SGF Profiling for testing semi-finished goods for fruit juiced. Therefore, the method presented in our manuscript would provide a more robust origin certification that would ensure the required quality compliance.

Multivariate or pattern recognition techniques such as principal component analysis (PCA) [18] are important tools for the analysis of the data obtained by NMR. The fingerprint of mixtures generated by ^{1}H NMR spectra in combination with PCA is a useful tool that has been applied for analyzing and sorting of different foods like apple [19], coffee [20], milk [21], mozzarella [22], among others. The potential of NMR spectroscopy in combination with PCA has been demonstrated in the detection of the adulteration of orange juices [23,24] and the discrimination between orange juice and pulp wash [25], in a fully automated quality control analysis of citrus juice called Spin Generated Fingerprint Profiling [26], in the discrimination of orange varieties [27], as well as in studies to extend the shelf life of fruit juices [28].

This methodology has been used in the citrus analysis worldwide. Nevertheless, to our knowledge, there are no reports of orange classification conducted with products in our territory. In this study, we report an NMR spectroscopic method, coupled to PCA, for the metabolomic analysis of Argentine citrus. In addition, six discrimination models based on partial least-squares – discriminant analysis (PLS–DA) were done for samples classification. Based on these data, classification and discrimination were performed for the six Citrus species, five botanical varieties of oranges and oranges of two different Argentine production regions. This leads to a clear differentiation based on a variety of metabolites. The results indicate that chemometric treatment of the ^{1}H NMR data represent a useful methodology to differentiate the geographical origin of oranges from two of the main production areas from Argentina, representing a suitable method for traceability of orange and/or orange juice production in Argentina.

2. Material and methods

2.1. Reagents and standards

All reagents were analytical grade unless stated otherwise. Buffer solution HPCE pH 3.0, sodium azide ≥ 99.5%, deuterium oxide (deuterium degree 99.9%) and 3-(trimethylsilyl) propionic-2,2,3,3-d$_4$ acid sodium salt (deuterium degree 98%) were purchased from Aldrich–Sigma. HCl was purchased from Cicarelli (San Lorenzo, Argentina).

2.2. Citrus samples

The studies were conducted with orange fruits (*Citrus sinensis* (L.) Osbeck) of six different varieties grown in the Estación Experimental Agropecuaria (INTA) from Entre Ríos and San Pedro, Argentina. The varieties selected were: Navelate, Salustiana and Navelina from Entre Ríos, New Hall, Ombligo and Lana Latex and Verano (this variety was used only in studies of species differentiation) from San Pedro. The authenticity of oranges was assessed by INTA inspectors. Fruits were collected at commercial maturity. Other citrus samples were bought in local greengrocers in Rosario, Argentina. These included lemons (*Citrus limon* (L.) Burm.), pink grapefruit (*Citrus paradisi* L.), mandarin (*Citrus reticulata* Blanco), lime (*Citrus aurantifolia* (Christm.) Swingle) and bitter orange (*Citrus aurantium* L.). The results shown in the present work correspond to fruits harvested in two subsequent years (2010/2011).

The fresh juice was extracted using an electric fruit squeezer. Before the NMR experiments, all samples were centrifuged (12,000 × *g*, 20 min), and the supernatant pH was adjusted. Three methodologies were evaluated: 1) addition of phosphate buffer pH = 3 (80% of the supernatant juice with 10% buffer), 2) addition of commercial buffer (80% of the supernatant juice with 10% buffer) or 3) 2200 μL of the supernatant mixed with microliter amounts of 1 M NaOH or 1 M HCl solutions. After adjusting the pH (3.02 ± 0.02) the solution was made up to 2250 μL with MilliQ water (enough amount for four replicates).

A volume of 540 μL of the each sample was mixed with 60 μL of D_2O containing 0.2% w/w of 3-(trimethylsilyl)propionic-2,2,3,3-d$_4$ acid, sodium salt (TMSP) and sodium azide (0.013% w/w), and was transferred to a 5-mm NMR tube. No additional treatment was necessary. The D_2O was used as the field frequency lock signal, TMSP for internal referencing of ^{1}H chemical shifts and sodium azide to suppress microorganism activity. Each sample was acquired in triplicate.

The sample preparation order and the acquisition of spectra were carried out based on a completely random experimental design, so as to avoid clustering generated for the acquisition in set of sample types.

2.3. NMR spectroscopy

All NMR spectra were recorded on a 300 MHz Bruker Avance spectrometer fitted with an autosampler with a 5 mm internal diameter smart probe with ATMA (Automatic Tunning Matching), holding the temperature stable at 300 K and the same acquisition parameters. The suppression of H_2O signals was carried out by using noesygppr-1d (Bruker standard pulses sequence) applying continuous waves during the relaxation delay (10 s) with a mixing time of 10 ms. Gradients were applied right after the relaxation delay and mixing time. Each spectrum was recorded with 64 scans and 48 K data points. The spectral width was adjusted to 20.0233 ppm (6009.615 Hz) with an acquisition time of 4.0 s per scan. The water signal was suppressed by using a low power irradiation frequency (53.52 dB) during both the relaxation delay and the mixing time. Spectra were Fourier transformed, phased, and apk0.noe Bruker's processing routine was applied (Fourier transform, phasing 0.order, and base line correction). The resulting spectra were aligned using the TMSP signal as reference, saved as .1r Bruker files, and transferred to a personal computer for data analysis.

2.4. Software and data pre-processing

All calculations were made using MATLAB (version 7.10; The Mathworks Inc., Natick, MA). PCA was run using an in-house MATLAB code. PLS-DA analysis was performed with MVC1, a new version of the MATLAB toolbox already reported in the literature [29] and available at http://www.iquir-conicet.gov.ar/eng/div5.php?area=12. Additional statistical parameters of the PLS-DA models were obtained with the Classification Toolbox 5.0 (http://michem.disat.unimib.it/chm/download/classificationinfo.htm) [30]. All programs were run on a personal computer with an Intel Core i5-3330, 3.00GHz (Ivy Bridge Technology) microprocessor and 8 GB of RAM. The MATLAB code ProMetab [31], originally written for metabolomics data analysis, was used to extract raw NMR data between 0.2 and 10.0 ppm from the original Bruker 1r files. Subsequently, all samples were aligned with the *Icoshift* algorithm [32] in the organic acids region (2.70–3.05 ppm). In addition, regions containing the signals from water and TMSP were excluded. Therefore, each sample subjected to analysis consisted of a vector of 7857 data points (10–0.2 ppm taken in steps of 0.00125 ppm; i.e., 0.375 Hz).

3. Results and discussion

3.1. General considerations

Visual inspection of a ^{1}H NMR spectrum of an orange juice shows

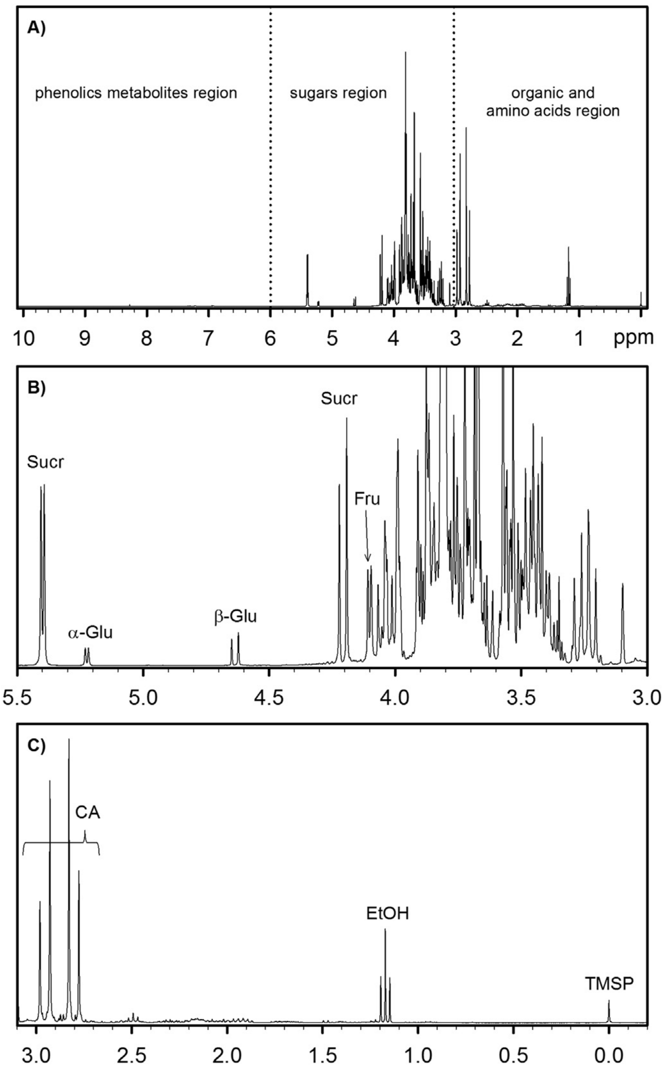

Fig. 1. (A) Typical noesypr1d spectrum of *Citrus sinensis* (Salustiana variety, Entre Rios origin) fresh orange juice. (B) Expanded region 3.0–5.5 ppm. (C) Expanded region −0.2–3.1 ppm. TMSP: 3-(Trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt, Sucr: sucrose, α-Glu: α-glucose, β-Glu: β-glucose, Fru: fructose, CA: citric acid and EtOH: ethanol.

three defined regions (Fig. 1a): the region of hydrogen atoms belonging to organic- and amino acids appear between 0.7 and 3.0 ppm; the region of carbohydrates such as sucrose, α-glucose, β-glucose, and fructose between 3.0 and 6.0 ppm, wherein hydrogens of anomeric carbons are clearly separated from the remaining sugar signals (Fig. 1b), and the phenolic metabolites region between 6.0 and 10.0 ppm. When the first region is analyzed (Fig. 1c), the intensity of citric acid signals (dd, 2.87 ppm) reveals that this compound overcomes by far the amount of other acids present in the orange juice sample. Moreover, another signal of particular interest in this region is the triplet assigned to the CH_3 group of the ethanol molecule (t, 1.17 ppm). This compound is generated in by unwanted alcoholic fermentation that naturally occurs when cutting the orange to make the juice. Although all sample spectra were recorded immediately after the oranges were squeezed and the pH adjusted, an amount of ethanol is always produced by the microorganisms present in the fruit [27].

3.2. Sources of variation of chemical shifts

The position in the frequency axis (chemical shift) of many NMR signals is pH dependent. Molecules whose ionization state changes with changing pH (organic acid, amino acids, etc.) show different shifts at different pH values. These spectral variations can be produced even by small pH differences and can be detrimental for the analysis, in particular because the multivariate analysis procedures require the corresponding signals in different samples to have the same chemical shift. In order to minimize this variation, we evaluated different options: phosphate buffer (pH = 3), a commercial buffer [26] and basic or acid solution for regulate the pH [25,33]. The best results were obtained with the last alternative (pH = 3.02 ± 0.02). In this case, the most problematic signal (citric acid) shows a very narrow variation in chemical shifts in comparison with the methodologies in which a buffer was used for pH regulation.

3.3. Multivariate data analysis

3.3.1. Data pre-processing

Before applying PCA to the NMR data, the effects of various data pre-processing techniques such as baseline correction, alignment, and/or normalization were analyzed. The data matrix was mean-centered and normalized by adjusting the total intensity of each spectrum (row) to total spectral area. Despite the careful pH adjustment commented above (Section 3.2), the organic acids signals (2.70–3.05 ppm, mostly citric acid) were not aligned across samples. The differences in peak positions of the same analyte signal among samples did not conform to the bilinearity assumptions for the application of most chemometric models, and thus they deteriorate the multivariate analysis. Many different algorithms are available for spectral matrix alignment, such as DMW [34], COW [35], Coshift [34], Icoshift [36], PAFFT [37], etc. In our hands, *Icoshift* algorithm yielded the best results. This algorithm, namely interval-correlation-shifting, independently aligns each NMR signal to a target by maximizing the cross-correlation between user-defined intervals, by using the FFT (Fast Fourier Transform) to boost the simultaneous alignment of all spectra in a dataset. Fig. 2 shows a zoomed selection (2.70–3.02 ppm) of the ^{1}H NMR spectra of all samples before (Fig. 2a) and after (Fig. 2b) the alignment. In this region, seven intervals were selected for the application of *Icoshift* to the NMR data (one signal per interval). Sample N° 9 (Lana Latex variety, San Pedro

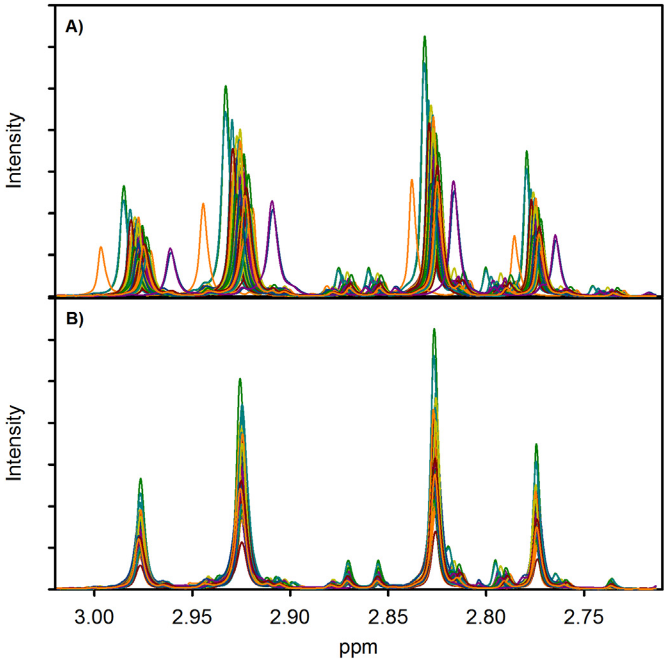

Fig. 2. Expanded region (2.70–3.02 ppm) of noesypr1d spectra of all *C. sinensis* fresh orange juice superimposed samples. (A) Before and (B) after the alignment with *Icoshift* algorithm.

Table 1
Description of the analyzed orange samples according to geographical origin and botanical varieties.

Geographical origin	Botanical variety	Number of samples
San Pedro	Lana Latex	9
	Ombligo	6
	New Hall	10
Entre Ríos	Salustiana	11
	Navel Late	9
	Navelina	11

origin) was used as reference for the alignment of the remaining ones. As shown in the Fig. 2, an almost perfect alignment of the signals was achieved.

3.3.2. Principal component analysis (PCA)

To perform sample discrimination, the obtained data were processed by PCA. This statistical method achieve data reduction by a linear combination of the original variables, highlighting the variance within the original dataset, and retaining most of the relevant information of the variables in the new first components. For the examination of authenticity (geographical origin), we started the study with 56 samples, 25 from San Pedro and 31 from Entre Ríos (Table 1). An initial exploratory study was done by applying PCA to the whole ^{1}H NMR raw spectra (56 × 7857; samples and 0.2–10 ppm data points respectively). Unfortunately, this analysis did not lead to a significant conclusion, because the PCA score plots did not allow for the differentiation between the samples according oranges varieties or geographical origin.

Accordingly, both manual and chemometric (variable selection techniques) searches of regions were done with the original NMR data in order to locate variables that gave the best results. In this case, manual selection achieved the best results by choosing the NMR regions of characteristic organic compounds that most contribute to fruit juice: sugars and acids. Furthermore, the CH_3 ethanol signal was added to the previous regions in order to study the presence of this unexpected but naturally occurring compound. Therefore, regions selected for principal component analysis were the anomeric carbons signals (5.44–5.36, 4.24–4.17, 5.25–5.20, 4.67–4.60, and 4.12–4.08 ppm, from sucrose, α-glucose, β-glucose, and fructose respectively), organic acid region (3.08–2.59 ppm), and CH_3 ethanol signal (1.25–1.06 ppm).

The PCA score plot for the selected regions of the data summarizes the relationships between the 56 samples. The cumulative percentage of explained data variance with the two first PC is above 91% (69.54% and 22.09% for PC1 and PC2 respectively), so data loss is negligible. Fig. 3 shows the PC1 vs PC2 score plot, in which we can observe the successful discrimination between three defined regions: San Pedro; *Salustiana* (Entre Rios); and, *Navelate* and *Navelina* (Entre Rios). Most samples from San Pedro have negative values for PC1 and positives for PC2. On the other hand, *Salustiana* variety from Entre Ríos have positives values both for PC1 and PC2, while other two Entre Ríos varieties (*Navelate* and *Navelina*) have negative values both for PC1 and PC2. Hence, PCA analysis of the selected NMR regions allows discrimination by geographical origin between San Pedro and Entre Rios orange samples. Furthermore, in this figure we can also observe discrimination by botanical origin of the *Salustiana* variety from the rest of Entre Rios samples.

Examination of the loadings resolved by PCA reveals which chemical compounds were decisive for orange discrimination. In Fig. 4 we can observe that the organic acids region (predominantly citric acid) and CH_3 ethanol signal displayed the largest contributions to PC1 (black line), and anomeric carbons signals in absolute values (sucrose, α-glucose, β-glucose, and fructose) and also CH_3 ethanol signal were the largest contributors to PC2 (red slash line). This means that the ethanol variable contributes to both principal components analyzed.

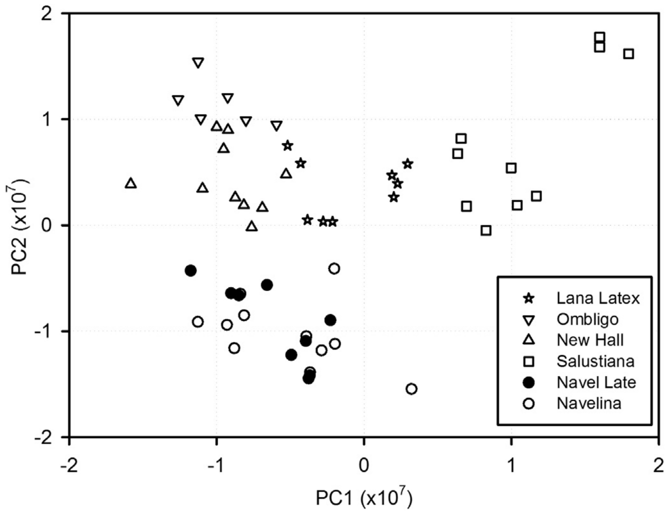

Fig. 3. Score plot of PC1 vs PC2 from selected ^{1}H NMR regions: Navelate (circle bold), Salustiana (square) and Navelina (circle) from Entre Ríos; New Hall (triangle), Ombligo (inverted triangle) and Lana Latex (asterisk) from San Pedro.

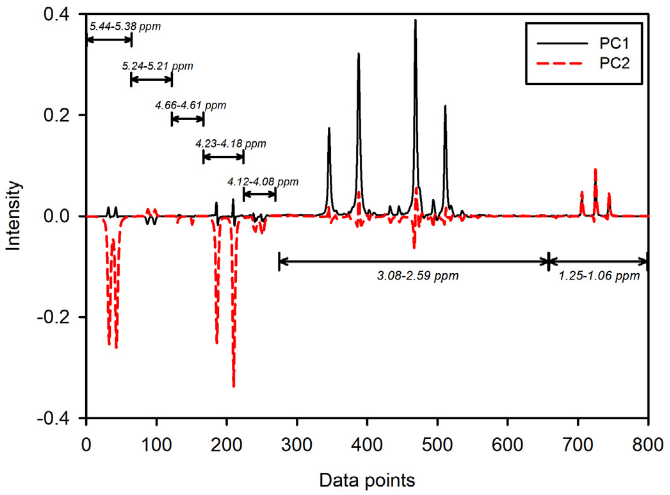

Fig. 4. Loadings plot of PC1 (black line) and PC2 (red slash line). Displayed range 5.44–4.08 ppm correspond to the selected signals of anomeric carbons (5.44–5.36, 4.24–4.17, 5.25–5.20, 4.67–4.60, and 4.12–4.08 ppm, from sucrose (H_{glu}-1), α-glucose, β-glucose, sucrose (H_{fru}-3) and fructose respectively). Range from 3.08–2.59 ppm correspond to organic acid region, and 1.25–1.06 ppm region correspond to CH_3 ethanol signal. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Comparison of areas for the signals corresponding to organic acid region of the orange NMR spectra showed a significant difference for all the samples (Wilcoxon test, $p < 0.05$). On average, areas of the signals found in the range $\delta = 3.08$–2.59 ppm account for a 68.7% of the total area of the spectrum for spectra of orange from Entre Ríos (SE = 19.3), whereas the integration of the signals in the same range for the samples of orange from San Pedro represents on average 50.9% of the total area of the spectrum (SE = 17.7). Similarly, significant difference for the signals corresponding to ethanol region was observed (Wilcoxon test, $p < 0.05$). In this case, the integration of the signals in the range $\delta = 1.14$–1.20 ppm account on average 6.5% (SE = 5.3) and 2.1% (SE = 1.2) of the total area of the spectrum the samples of orange from Entre Ríos and San Pedro respectively. Finally, the differences were not significant for sucrose, α-glucose, β-glucose, and fructose (Wilcoxon test, $p > 0.05$).

In summary, alcoholic fermentation is the source of ethanol, and organic acids in media with high sugar concentrations have an enormous impact on metabolism of microorganisms such as *S. cerevisiae* [38]. Therefore, classification of orange juices both for geographical and botanical origin achieved with the contribution of ethanol variable imply that microorganisms naturally present in orange fruit play a key role in the classification, but further studies are needed to validate this finding.

Apart from the origin, with this methodology we evaluated the ability of differentiating citrus species. In the score plots (matt supp Fig. 1) it is possible to differentiate orange sweet, orange bitter, tangerine, grapefruit, lemon and lime. Of course, the macroscopic differentiation of different species is rarely a reason for complaints. If well the colors of extracts obtained from citrus fruits have clear differences in some cases in others the difference is not clear (orange sweet, orange bitter and tangerine). Even the adulteration of orange juice with tangerine juice (lower cost), pulp wash, preservatives, sugar, or other ingredients added the difference respect to pure orange juice may not be distinguishable. Preliminary results indicate that this methodology can also be extended to discriminate oranges from different harvest time and/or inadequate conservation processes.

3.3.3. Partial least-squares – discriminant analysis (PLS-DA)

Partial least-squares regression is a versatile technique for multivariate data analysis that finds a regression model by projecting the predicted variables (Y) and the observable variables (X) to a new space [39]. In classical PLS analysis, the values of Y are analyte concentrations or sample properties. When classification is the objective of the PLS modeling, the values of Y are coded so that they reflect the different sample categories. This can be done using digital values such as 0 and 1, as in the present case. Employed in this manner the model is called PLS-DA, for partial least-squares – discriminant analysis [40].

Although PCA showed good differentiation between the geographical origin of the samples (San Pedro vs Entre Ríos), all oranges varieties were not satisfactory discriminated within each other with PCA method. Therefore, we select different PLS-DA models in order to discriminate between oranges varieties. The data selected for this analysis was the same NMR regions selected for the PCA analysis (i.e., anomeric carbons signals, organic acid region, and CH_3 ethanol signal). Due to the high variability in the NMR data for both botanical and geographical origin, selection of PLS-DA models was directed to form well-defined groups for each classification model. Therefore, we design five different PLS-DA models in order to discriminate the orange samples by botanical origin, and one PLS-DA model to confirm the classification by geographical origin achieve previously with PCA analysis. Consequently, classification models were designed as follows with the intention of discriminating orange samples according to: (1) geographical origin (PLS-DA1), (2) Lana Latex variety (PLS-DA2), (3) Ombligo variety (PLS-DA3), (4) New Hall variety (PLS-DA4), (5) Salustiana variety (PLS-DA5), and (6) Nave Late plus Navelina varieties (PLS-DA6). In the PLS-DA1 model, two classes were considered: 1 (Entre Ríos) and 0 (San Pedro), whereas in the varieties discrimination models (PLS-DA2 to PLS-DA6), the two considered classes were: 1 (variety) and 0 (remaining samples not belonging to the selected variety). In PLS-DA6 model, two orange varieties define together a unique class. Initially, two separate classification models were designed for Nave Late and Navelina varieties but unfortunately neither of these two models could discriminate any of the varieties under study. Therefore, PLS-DA6 model was proposed in order to differentiate at least these two varieties from the remaining samples.

In the PLS-DA1 model, the number of samples belonging to each class was equivalent between San Pedro and Entre Ríos. On the other hand, discriminations models PLS-DA2 to PLS-DA6 presented unbalanced datasets; i.e., there is a greater number of samples that do not belong to the variety that is intended to discriminate. Moreover, all PLS-DA models were built with few samples in the datasets. Therefore, the samples partitioning between calibration and validation datasets were done by the application of the Kennard–Stone algorithm [41] to each class separately. Accordingly, for each PLS-DA model, one-half of the orange samples was used for calibration, and the other half was used for validation.

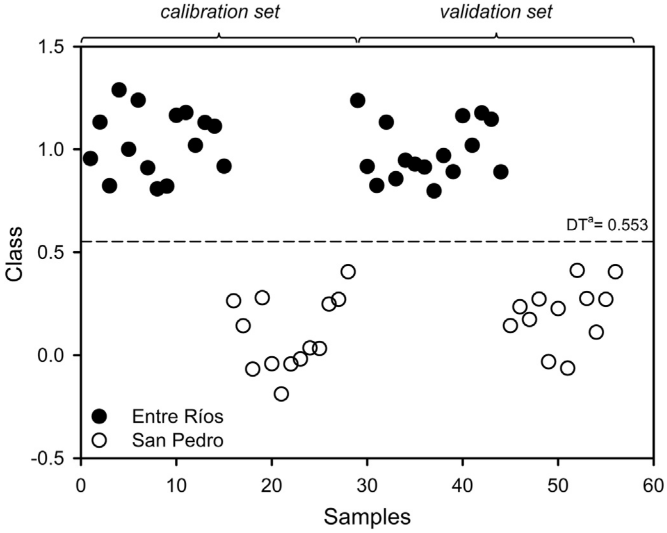

Fig. 5. Class values obtained after the application of PLS-DA1 model to discriminate the samples by geographical origin. Code one stand for Entre Ríos oranges and code zero for San Pedro oranges. [a]DT: discrimination threshold.

Fig. 5 shows the discrimination of the orange samples by geographical origin achieved with the PLS-DA1 model. Moreover, Fig. 6 shows the PLS-DA models designed for discrimination of the oranges samples by botanical origin (PLS-DA2 to PLS-DA6). Discrimination threshold (DT) for each PLS-DA model were defined with the analysis of the Receiver operating characteristic (ROC) curves obtained from the Classification Toolbox 5.0 [30]. Except for PLS-DA5 (four latent variables), for all the PLS-DA models two PLS latent variables, estimated from leave-one-out cross-validation, accounted for most of the variability of the data, capturing between 88 and 92% and 85–88% of the X and Y variables, respectively.

Table 2 resumes the classification parameters obtained for each PLS-DA model such as precision (Pr), sensitivity (Sn), specificity (Sp) along with the mean prediction errors related to the application of the designed models to the NMR data. Analysis of the designed models shown that PLS-DA3 (Ombligo variety) was the less accurate model with misclassified samples, as can be observed in the lower values of Pr (60%) and Sp (92%). The remaining PLS-DA discrimination models both for botanical and geographical origin yielded excellent classification parameters with perfect separation among the orange samples. Obviously, as can be seen both from Table 2 as Figs. 5 and 6, the mean square errors (RMSEC and RMSEP) did not equal for all the models. PLS-DA5 (Salustiana variety), PLS-DA1 (geographical origin), and PLS-DA2 (Lana Latex variety) account for the most accurate classification models. Furthermore, these three models had the closest DT to the mean value (0.5); i.e., PLS-DA5, PLS-DA1, and PLS-DA2 were the models with lower differences in the dispersions of the estimated values for each discriminated class.

These results were consistent with those obtained previously by PCA analysis; i.e., regarding geographical origin discrimination, San Pedro oranges were correctly differentiated from Entre Ríos samples. Furthermore, Salustiana variety was clearly separated from the rest of oranges varieties; and Nave Late and Navelina varieties could not be differentiated from each other.

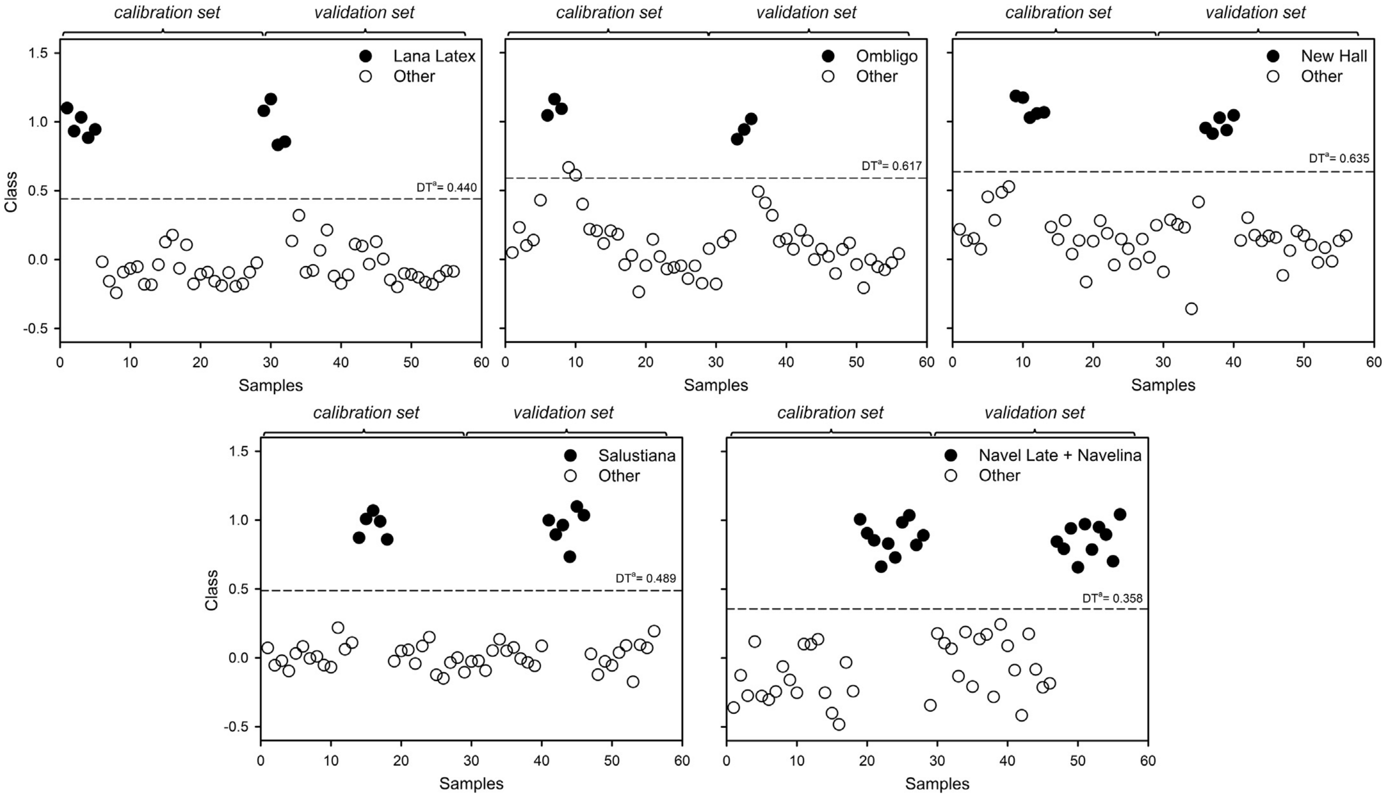

Fig. 6. Class values obtained in the varieties discrimination models (PLS-DA2 to PLS-DA6). Code one stand for variety and code zero for the remaining samples not belonging to the selected variety. PLS-DA2 (Lana Latex), PLS-DA3 (Ombligo), PLS-DA4 (New Hall), PLS-DA5 (Salustiana), and PLS-DA6 (Nave Late plus Navelina). [a]DT: discrimination threshold.

Table 2
Discrimination models for orange samples according to geographical origin (PLS-DA1) and botanical origin (PLS-DA2 to PLS-DA6).

Model	Class	Pr[a] (%)	Sn[b] (%)	Sp[c] (%)	LV[d]	DT[e]	RMSEC[f]	RMSEP[g]
PLS-DA1	Entre Ríos	100	100	100	2	0.553	0.166	0.181
	San Pedro	100	100	100				
PLS-DA2	Lana Latex	100	100	100	2	0.440	0.108	0.145
	Other	100	100	100				
PLS-DA3	Ombligo	60	100	92	2	0.617	0.183	0.133
	Other	100	92	100				
PLS-DA4	New Hall	100	100	100	2	0.635	0.181	0.120
	Other	100	100	100				
PLS-DA5	Salustiana	100	100	100	4	0.489	0.090	0.091
	Other	100	100	100				
PLS-DA6	Navel Late + Navelina	100	100	100	2	0.358	0.201	0.191
	Other	100	100	100				

[a] Pr: precision.
[b] Sn: sensitivity.
[c] Sp: specificity.
[d] LV: latent variable number.
[e] DT: discrimination threshold.
[f] RMSEC: root mean square error of calibration.
[g] RMSEP: root mean square error of prediction.

4. Conclusions

Untargeted metabolic profiling is a valuable exploratory tool capable of providing extensive chemical information for use of fingerprint with classification purposes. In this preliminary study we have been able to distinguish the origin of oranges from two of the main production areas in Argentina, representing a rapid method for quality control of orange in Argentina. Citric acid and ethanol and were the chemical variables which have greater influence on the differentiations. The alcoholic fermentation appears to be particularly sensitive to alterations in the pH of the medium. This could justify the relationship between the variables that have the largest effect on the observed classifications. Besides oranges classification by geographical origin, this methodology was able to differentiate citrus species and orange varieties.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2018.05.037.

Acknowledgements

This work was supported by Universidad Nacional de Rosario, CONICET (PIP 695) and ANPCyT (PICT2015-3574) from Argentina. We thank INTA for providing samples and information about them.

References

[1] Y. Liu, E. Heying, S.A. Tanumihardjo, History, global distribution, and nutritional importance of citrus fruits, Compr. Rev. Food Sci. Food Saf. 11 (2012) 530–545.

[2] Federcitrus: The Argentine Citrus Industry, Annual Report of Argentine Citrus Federation, (2016).

[3] L.M. Reid, C.P. O'Donnell, G. Downey, Recent technological advances for the determination of food authenticity, Trends Food Sci. Technol. 17 (2006) 344–353.

[4] L. Hu, C. Yin, S. Ma, Z. Liu, Tracing the geographical origin of burdock root based on fluorescent components using multi-way chemometrics techniques, Microchem. J. 137 (2018) 456–463.

[5] E. Cubero-Leon, R. Peñalver, A. Maquet, Review on metabolomics for food authentication, Food Res. Int. 60 (2014) 95–107.

[6] G. Xiang, H. Yang, L. Yang, X. Zhang, Q. Cao, Mingming Miao Multivariate statistical analysis of tobacco of different origin, grade and variety according to polyphenols and organic acids, Microchem. J. 116 (2014) 107–117.

[7] P.L. Pisano, M.F. Silva, A.C. Olivieri, Anthocyanins as markers for the classification of Argentinean wines according to botanical and geographical origin. Chemometric modeling of liquid chromatography–mass spectrometry data, Food Chem. 175 (2015) 174–180.

[8] M. Efenberger-Szmechtyk, A. Nowak, D. Kregiel, Implementation of chemometrics in quality evaluation of food and beverages, Crit. Rev. Food Sci. Nutr. (2017) 1–20.

[9] J. Bricout, J. Koziet, Control of the authenticity of orange juice by isotopic analysis, J. Agric. Food Chem. 35 (1987) 758–760.

[10] S. Wibowo, T. Grauwet, B.T. Kebede, M. Hendrickx, A. Van Loey, Study of chemical changes in pasteurised orange juice during shelf-life: a fingerprinting-kinetics evaluation of the volatile fraction, Food Res. Int. 75 (2015) 295–304.

[11] F. Ammari, L. Redjdal, D.N. Rutledge, Detection of orange juice frauds using front-face fluorescence spectroscopy and Independent Components Analysis, Food Chem. 168 (2015) 211–217.

[12] E. Muntean, Simultaneous carbohydrate chromatography and unsuppressed ion chromatography in detecting fruit juices adulteration, Chromatographia 71 (2010) S69–S74.

[13] R. Díaz, O.J. Pozo, J.V. Sancho, F. Hernández, Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry, Food Chem. 157 (2014) 84–93.

[14] J.E. Gaiad, M.J. Hidalgo, R.N. Villafañe, E.J. Marchevsky, R.G. Pellerano, Tracing the geographical origin of Argentinean lemon juices based on trace element profiles using advanced chemometric techniques, Microchem. J. 129 (2016) 243–248.

[15] R.G. Pellerano, S.S. Mazza, R.A. Marigliano, E.J. Marchevsky, Multielement analysis of Argentinean lemon juices by instrumental neutronic activation analysis and their classification according to geographical origin, J. Agric. Food Chem. 56 (2008) 5222–5225.

[16] M. Amenta, S. Fabroni, C. Costa, P. Rapisarda, Traceability of 'Limone di Siracusa PGI' by a multidisciplinary analytical and chemometric approach, Food Chem. 211 (2016) 734–740.

[17] A.P. Minoja, C. Napoli, NMR screening in the quality control of food and nutraceuticals, Food Res. Int. 63 ((2014) 126–131.

[18] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831.

[19] A. Francin, S. Romeo, M. Cifelli, D. Gori, V. Domenici, L. Sebastiani, ^{1}H NMR and PCA-based analysis revealed variety dependent changes in phenolic contents of apple fruit after drying, Food Chem. 221 (2017) 1206–1213.

[20] M.V. de Moura Ribeiro, N. Boralle, H.R. Pezza, L. Pezza, A.T. Toci, Authenticity of roasted coffee using ^{1}H NMR spectroscopy, J. Food Compos. Anal. 57 (2017) 24–30.

[21] R. Lamanna, A. Braca, E. Di Paolo, G. Imparato, Identification of milk mixtures by ^{1}H NMR profiling, Magn. Reson. Chem. 49 (2011) S22–S26.

[22] P. Mazzei, A. Piccolo, ^{1}H HRMAS-NMR metabolomic to assess quality and traceability of mozzarella cheese from Campania buffalo milk, Food Chem. 132 (2012) 1620–1627.

[23] J.T.W.E. Vogels, L. Terwel, A.C. Tas, F. van den Berg, F. Dukel, J. van der Greef, Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques, J. Agric. Food Chem. 44 (1996) 175–180.

[24] E. Vigneau, F. Thomas, Model calibration and feature selection for orange juice authentication by ^{1}H NMR spectroscopy, Chemom. Intell. Lab. Syst. 117 (2012) 22–30.

[25] G. Le Gall, M. Puaud, I.J. Colquhoun, Discrimination between orange juice and pulp wash by ^{1}H nuclear magnetic resonance spectroscopy: identification of marker compounds, J. Agric. Food Chem. 49 (2001) 580–588.

[26] M. Spraul, B. Schütz, P. Rinke, S. Koswig, E. Humpfer, H. Schäfer, M. Mörtter, F. Fang, U.C. Marx, A. Minoja, NMR-based multi parametric quality control of fruit juices: SGF profiling, Nutrients 1 (2009) 148–155.

[27] C.R. de Oliveira, R.L. Carneiro, A.G. Ferreira, Tracking the degradation of fresh orange juice and discrimination of orange varieties: an example of NMR in coordination with chemometrics analyses, Food Chem. 164 (2014) 446–453.

[28] E.G. Alves Filho, F.D.L. Almeida, R.S. Cavalcante, E.S. de Brito, P.J. Cullen, J.M. Frias, P. Bourke, F.A.N. Fernandes, S. Rodrigues, ^{1}H NMR spectroscopy and chemometrics evaluation of non-thermal processing of orange juice, Food Chem. 204 (2016) 102–107.

[29] A.C. Olivieri, H.C. Goicoechea, F.A. Iñón, MVC1: an integrated MatLab toolbox for first-order multivariate calibration, Chemom. Intell. Lab. Syst. 73 (2004) 189–197.

[30] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (2013) 3790–3798.

[31] M.R. Viant, Improved methods for the acquisition and interpretation of NMR metabolomic data, Biochem. Biophys. Res. Commun. 310 (2003) 943–948.

[32] F. Savorani, G. Tomasi, S.B. Engelsen, Icoshift: a versatile tool for the rapid alignment of 1D NMR spectra, J. Magn. Reson. 202 (2010) 190–202.

[33] G. del Campo, I. Berregi, R. Caracena, J.I. Santos, Quantitative analysis of malic and citric acids in fruit juices using proton nuclear magnetic resonance spectroscopy, Anal. Chim. Acta 556 (2006) 462–468.

[34] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (2004) 231–241.

[35] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, J. Chromatogr. A 805 (1998) 17–35.

[36] G. Tomasi, F. Savorani, S.B. Engelsen, Icoshift: an effective tool for the alignment of chromatographic data, J. Chromatogr. A 1218 (2011) 7832–7840.

[37] J.W.H. Wong, C. Durante, H.M. Cartwright, Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets, Anal. Chem. 77 (2005) 5655–5661.

[38] M.J. Torija, G. Beltran, M. Novo, M. Poblet, N. Rozés, A. Mas, J.M. Guillamón, Effect of organic acids and nitrogen source on alcoholic fermentation: study of their buffering capacity, J. Agric. Food Chem. 51 (2003) 916–922.

[39] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193 − 1202.

[40] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding, Anal. Chim. Acta 879 (2015) 10–23.

[41] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.