Coffee 'Identitation' through chromatographic fingerprint: simultaneous classification of geographical origin & post-harvest treatments

Davide Bressanello[a], Maryam Abbasi Tarighat[c], Erica Liberto[a], Gloria Pellegrino[b], Manuela R. Ruosi[b], Carlo Bicchi[a]

[a] Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via Pietro Giuria 9, Torino, Italy
[b] Lavazza spa, Strada Settimo 410, 10156 Torino, Italy
[c] Department of Chemistry, Persian gulf University, Bushehr-Iran

Email: erica.liberto@unito.it

INTRODUCTION

Fingerprinting strategies are widely used in food authentication[1]. Authentication implies to confirm the stated specifications, including quali-quantitative identification of characteristic components, adulterants, contaminants or to verify quality requirements as botanical origin or processing procedures. However, food authentication is based on the evaluation of similarities of an instrumental fingerprint versus a reference representative of sample variability, such as the human fingerprint in forensic science. This step is known as food 'Identitation', i.e. the establishment of an instrumental fingerprint characteristic of the authenticity[2]. The reliability of food authentication depends on a correct 'Identitation'. The fingerprint approaches require an adequate number of samples stated as authentic food to establish a representative data base of the genuine food population[1-3]. The use of chemometric as exploration, classification and prediction tools is fundamental to extract significant and not-evident information to develop pattern recognition models. In this study, HS-SPME-GC–MS was applied to the aroma chemical fingerprinting of a set of coffee samples of different origin to discriminate simultaneously their origin and post-harvest treatments. An untargeted analysis of the pre-processed coffee data, including classical multivariate analysis as principal component analysis (PCA) and partial least square-discriminant analysis (PLS-DA) was performed. The results showed that, PLS-DA provided significant results for coffee classification, in particular in agreement with their geographical origin and processing, with error rates of 0.03 and 0.06 for fitting and prediction samples, respectively. Coffee aroma fingerprint can therefore further be exploited for food 'identition' in view of origin & processing coffee authentication.

References

1.G.P Danezis, A.S. Tsagkaris, V. Brusic & C.A. Georgiou. *Current Opinion in Food Science* 2016, 10:22–31.

2.L. Cuadros-Rodríguez, C.Ruiz-Samblas, L. Valverde-Som,E. Perez-Castano, A.Gonzalez-Casado. *Anal. Chim. Acta* 909 (2016) 9-23.

3.S.D. Johanningsmeier, G.K.Harris, and C.M. Klevorn. Annu. *Rev. Food Sci. Technol.* 2016. 7:413–38.

MATERIALS & METHODS

Sample acronym	Sample Name	Species	Treatment
BRA	BRAZIL LA2	Arabica	Natural
COL	COLOMBIA CL1	Arabica	Washed
JAV	JAVA WB1 MB	Robusta	Washed
UGA	UGANDA STD	Robusta	Natural
PNG	PAPUA NG Y	Arabica	Washed
INDIA	INDIA ARAB CHERRY	Arabica	Natural
INDO	INDONESIA EK1	Robusta	Natural
KAFA	ETIOPIA KAFA GR. 3	Arabica	Natural

Table 1 List and characteristics of the coffee samples used in this study.

Samples

Coffees samples, consisting of roasted coffee ground to suit a coffee-filter machine, were kindly supplied over a period of 9 months by Lavazza Spa (Turin, Italy). Forty coffee samples originating from eight countries (Ethiopia, Papua New Guiney, Colombia, Brazil, India, Indonesia, Java, and Uganda) differently treated after harvesting (natural and washed samples), belonging the species *Coffea Arabica* L. (Arabica) and *Coffea canephora* Pierre (Robusta), were analyzed (**Table 1**). The roasting degree of each sample was carefully measured by ground bean light reflectance, with a single-beam Neuhaus Neotec Color Test II instrument (Genderkesee, Germany) at a wavelength of 900 nm on 25-30g of ground coffee. Roasting degree was set at 55°Nh, in order to be close to the international standardization protocol for cupping. Samples were roasted within 24 hours prior to cupping, and left for at least 8 hours to stabilize. The coffee brew was prepared from 18g of coffee powder and 300mL of water, using a "Xlong" coffee filter machine.

Volatiles sampling

HS-SPME sampling was carried out with a QP2010 GC-MS system equipped with an autosampler combi-PAL AOC 5000 Autoinjector (Shimadzu - Milan, Italy). 1.5 g of powder were weighed in a septum-sealed gas vial (20mL); the resulting headspace was sampled through the SPME fiber for 40 minutes at 50°C with an agitation speed of 350rpm. Tridecane (C13) in Dibuthylphtalate (DBP), was used as internal standard, were purchased from Sigma-Aldrich (Milan-Italy). The internal standard was pre-loaded onto the fiber (Wang, O'Reilly, Chen, & Pawliszyn, 2005) in advance by sampling 5µL of a 1000mg/L solution of *n*-C13 in DBP into a 20mL headspace vial for 20 min at 50°C, agitation speed of 350rpm. Each sample was analyzed in two technical replicates

Analysis Conditions

HS-SPME-GC-MS chromatographic conditions: injector temperature: 230°C; injection mode, splitless; carrier gas, helium (2mL/min); fiber desorption time and reconditioning, 5min; column, SGE SolGelwax (100% polyethylene glycol) 30 m x 0.25 mm dc x 0.25 µm df (SGE- Melbourne, Australia); temperature program, from 40°C (1min) to 200°C at 3°C/min, then to 250°C (5min) at 10°C/min. MS conditions: ionization mode: EI (70eV); scan range: 35-350 amu; ion source temperature: 200°C; transfer line temperature: 250°C.

Data processing

Data were collected with a Shimadzu GCMS Solution 2.5SU1. The original chromatographic profiles were organized into a matrix format X (I×J), where each replicate represented one sample. The chromatogram alignments were performed according to Malmquist and Danielsson (1994) criteria. The chromatograms were divided into different regions and, for each region, Orthogonal projection approach (OPA) and evolving factor analysis (EFA) was used to determine zero component and proper segmented regions. Savitzky- Golay second order derivative (second order polynomial with a seven points window) was applied on whole data sets.
The regression methods used for data treatment were partial elaborations and done with MATLAB software (version 7.8, MathWorks, Natick, MA, USA). Basic statistics and principal component analysis (PCA) were performed with R 2.15.1 (R Development Core Team) and partial least square-discriminant analysis (PLS-DA) with the Classification toolbox for MATLAB®.

RESULTS & DISCUSSION

FIGURE 1

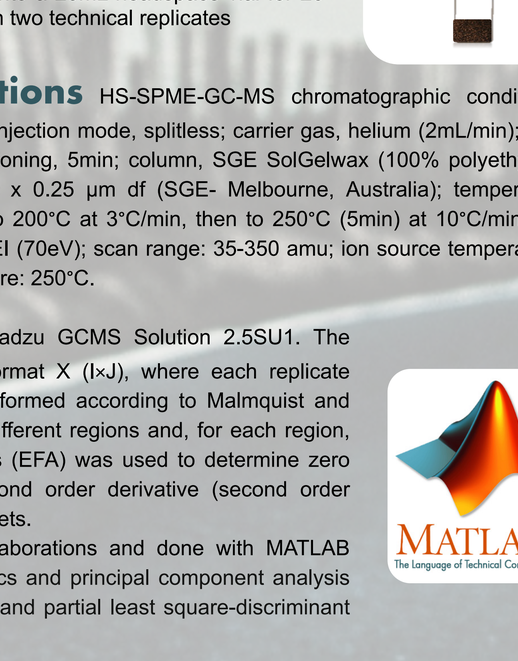

Fig.1
(a) Total ion chromatograms (TICs) of chromatographic fingerprints of sixty coffee samples and **(b)** performance of baseline correction, smoothing and alignment of chromatograms and **(c)** distribution pattern of the coffee samples in the two-dimensional PCA-based factor space of their GC-MS data.

The data provided by a metabolomic approach, mainly from untargeted and fingerprinting strategy, is of great complexity, and correct data treatment is of the utmost importance. All non-processed data were exported to the MATLAB software for further processing and alignment of data sets. Because problems such as base line contribution, noise in a data matrix, variation in peak shape, retention time shifts and co-elution make preprocessing necessary of data before multivariate classification. Smoothing and baseline correction on chromatograms to reduce variation source (baseline/background contribution) that carry no relevant information during multivariate calibration classification alignment, are shown in fig 1.
Unsupervised model based on PCA applied to the samples under study failed into grouping them.
There is no evidence of separation between the classes along the two principal components and there is a large degree of overlapping between samples (between JAV and BRA; INDIA and INDO).
Therefore, the PCA model seems unable to produce a reference "identity" for origins and/or treatments.

An authentic reference is necessary to authenticate a food. In particular when using metabolomics approach for authentication, the comprehensive chemical compounds characterizing for instance the aroma of a food and the representativeness of samples is mandatory to create a model reference of identity of that food, as the identity card to prove a person's identity. Identitation goodness therefore heavy influence reliability and confidence of the following authentication.
This is ever more important for coffee quality control, in particular to define the cost and to standardize further finish-end products.
In PCA, the amount of original variables is reduced to a few independent variables or principal components (PCs) that still are the main information from the original data set. However, each PCA model is generated based on the data demonstrating the highest variation, which might also be distinctly different from separating the classes. Maximum class-separation is thereby not explicitly the objective function of the method. Partial least squares (PLS) procedure performs interdependent PCA decomposition in both X (Independent variable) and Y (dependent variable) matrices. Therefore, each PLS factor (latent variable) is extracted from the independent variables and simultaneously correlated with the variance of the dependent variable enabling to attribute each sample to the appropriate class. A requisite for an accurate and robust sample classification modelling is the availability of a training/calibration set and a validation set. The latter should sufficiently represent the entire dataset in order to provide reliable estimation of the true model predictive ability.

The cross-validation procedure was here adopted. The validation set was obtained by randomly selecting the samples from the whole set. The training (calibration) set was used to calibrate the PLS-DA classification model, whereas the test samples was only used in the final stage to evaluate the true predictive ability of the calibrated model.

FIGURE 2

Step	Comp.	Exp. Var.%	Error rate	Non error rate	Non assigned	Accuracy
Training	18	97	0.03	0.97	0.10	0.96
Cross valid.	18	94	0.06	0.94	0.31	0.93
Predict.	18	100	0.15	0.85	0.28	0.94

Classification of coffee origins

Following pre-processing and alignment of the data, a model was built using the sixty-one training data set and twenty-five prediction samples. The optimal number of latent variables was selected as the number associated to the minimum error (18 LVs), since this approach is preferable in terms of model sensitivity, specificity, interpretation and stability. Figure 2 shows classification for INDO, INDIA and KAFA together with the classification parameters obtained in fitting, cross validation with 5 groups split and on the test set by PLS-DA.

The model performance can be evaluated using some parameters such as sensitivity and specificity. The sensitivity is the model ability to correctly classify the samples, relating the predicted samples to being in a class with the samples that really are in this class. The specificity relates the predicted samples that are not in a class with the samples that actually are not in this class. The confusion matrix collects the outputs of the classification model the classification performance. Table 2 summarizes the model classification performances parameters. Results show a good capability of the model to classify samples origins correctly with a relatively low power for Colombia samples and with low specificity Kafas' sample

TABLE 2

Model Sample	Fitting		Cross validation		Prediction set	
	Spec	sens	Spec	sens	Spec	sens
BAR	1	1	1	1	1	1
COL	1	0.86	0.88	0.5	1	0.67
JAV	0.98	1	0.88	1	0.94	1
UGA	1	1	0.97	1	1	1
PNG	1	1	0.97	0.75	1	1
INDIA	1	1	0.94	1	1	1
INDO	1	1	1	0.75	1	1
KAFA	0.98	1	0.92	1	0.5	1

Table 2. The specificity and selectivity valued of Fitting, cross validation and prediction steps

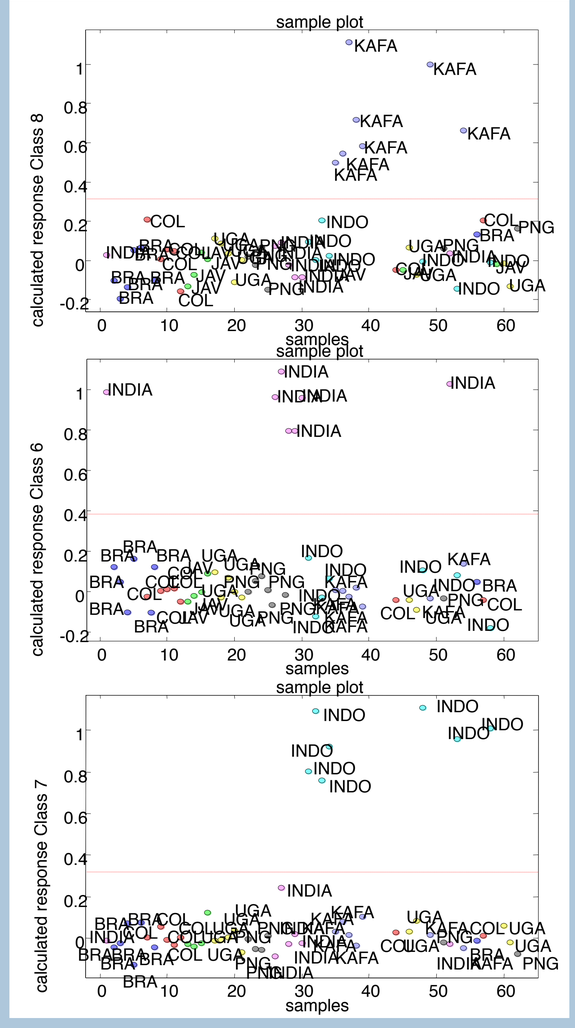

Fig. 2 PLS-DA Classification for INDIA; INDO and KAFA and model parameters obtained in fitting, cross validation with 5 groups split and on the test set by PLS-DA

Classification of coffee treatments

A PLS-DA was developed to discriminate coffee samples according to the post-harvest treatment using untargeted GC-MS data. The best model after removing outlier was obtained by 7 LVs that explained 89.82% with the lowest value of prediction error for each class and highest values of sensitivity and specificity. The model was able to classify the samples within an error rate of classification of 3 % in fitting. The score plot of PLS-DA latent variables, LV1 vs. LV2 are presented in Fig.3 together its confusion matrix. These results show a separation between samples according to the treatments, indicating that all samples were discriminated by the positive part of LV1 and LV2. The results show that coffee samples were correctly classified at 98 % in fitting and 90 % in prediction.

FIGURE 3

Step	Natural	Washed	Not assigned
Fitting			
Natural	43	0	0
Washed	1	18	0
Cross validation			
Natural	40	1	2
Washed	2	17	0

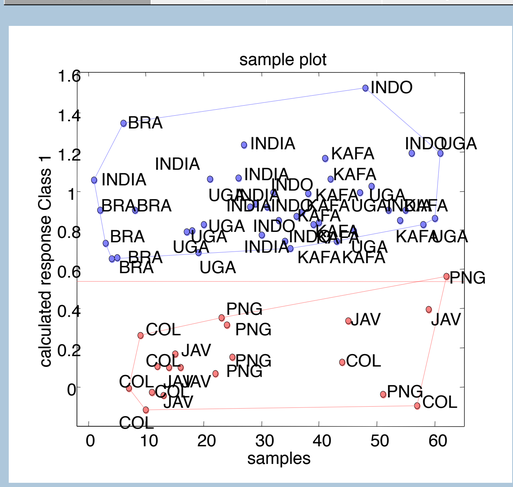

Fig. 3
Coffee post-harvest treatments classification by PLS-DA together with the confusion matrix of fitting model and Cross validation

CONCLUSIONS

These preliminary results show that coffee aroma fingerprint hide chemical information that could be exploited in a coffee quality control to authenticate origins and post-harvest treatments with an unique analytical evaluation. This goal can be reached by building an "identity card" for each origin and treatment and requires a) a high number of representative authentic samples, b) a fast and automatic analytical tools and c) an integrated analytical system in which sample preparation, analysis and qualification of samples through modelling occur in a unique step to obtain a high-throughput screening.