HAM QUALITY EVALUATION ASSISTED BY GC-IMS

L. Fernandez1, A. Martín-Gómez2, M. Mar Contreras2, M. Padilla 1, S. Marco1, L. Arce2
1Institute for Bioengineering of Catalonia (IBEC). Torre I. 08028-Barcelona. Spain.
2Department of Analytical Chemistry. Annex C-3 Building. Campus of Rabanales. Institute of Fine Chemistry and Nanochemistry. University of Cordoba. 14071-Córdoba. Spain.

ABSTRACT

In recent years, Gas Chromatography-Ion Mobility Spectrometry (GC-IMS) has been successfully employed in food science as a control technique for the prevention of fraud according to food and labeling regulations. In this work, we propose the use of GC-IMS technique to assess the quality of Iberian ham with regard to the Iberian Pig's diet (either nourished with feed or with acorns). For this purpose, we have acquired a dataset composed of 53 samples of Iberian ham from different food providers using a commercial GC-IMS (FlavourSpec, from G.A.S. Dortmund, Germany). Intensive signal pre-processing for GC-IMS was applied to the raw data. This dataset was employed to create 4 Partial Least Squares Discriminant Analysis (PLSDA) models corresponding to different train/test partitions of the dataset. Nearly perfect classification rates (above 91 %) were obtained for each partition of the dataset, denoting the high power of GC-IMS to characterize food samples.

Index Terms— *GC-IMS, Food Science, ham quality, classification, PLSDA*

1. INTRODUCTION

Gas Chromatography – Ion Mobility Spectrometry (GC-IMS) combines the ability to separate chemical mixtures exhibited by chromatographic columns with the excellent levels of sensitivity and selectivity characteristic of ion mobility spectrometers. GC-IMS technique allows fast qualitative and quantitative analyses of the Volatile Organic Compounds (VOCs) present in the headspace of solid and liquid samples. In addition to this, GC-IMS technology is more economical and requires less maintenance than its classical counterpart: Gas Chromatography-Mass Spectrometry (GC-MS). In consequence, GC-IMS has become an interesting alternative to CG-MS for the analysis of alimentary samples [1-2]. Unfortunately, GC-IMS also partakes of two disadvantages of GC-MS that hinders its application in food sample categorization: First, relatively large datasets must be acquired to obtain reliable predictions due to the hyphenated nature of the technique. And second, tolerances in the instrumental parameters, changes in environmental conditions, and sensor noise can cause profound differences in data samples, with independence of the class they belong to.

In this work, we have used GC-IMS to extract the characteristic signatures of two distinct types of Iberian ham qualities. To this end, we have acquired a collection of 53 ham samples from pigs nourished with feed or with acorns during their fattening period, and obtained from 16 different food providers. We have applied a signal pre-processing stage on the raw GC-IMS data to minimize the amount of sample variance non-related to ham quality that would eventually lead to misclassification. This stage included: noise filtering, baseline removal, peak alignment (both in retention and drift time) and decimation. The corrected data was employed to create 4 Partial Least Squares Discriminant Analysis (PLSDA) classification models, each of them from a different training/test partition of the dataset. Finally, we have compared the Classification Rate (CR) of the models to ensure the generality of the predictive results, no matter how the data samples were sorted.

2. EXPERIMENTAL

We used a commercial GC-MS (*FlavourSpec*, G.A.S., Dortmund, Germany), which enabled direct sampling of the headspace. The instrument was coupled to an autosampler unit (CTC-PAL, CTC Analytics AG, Zwingen, Switzerland). Ham samples from 16 different food providers (Pedro Diego, Selectos Peñaranda, Sierra de Monesterio, 5 Dehesas, Concurso, La Barrica, Romero Torres, Benito, Delicious Dia, Sierra de Azuaga, Tronco Bretán, Navidul, Delicias de Hurtada, Concurso, Piedra, and Ibérico del Valle) were generated and acquired with following experimental protocol: We placed pieces of ham (1 g) in a 20 mL glass vial closed with magnetic caps. After 20 min of incubation at 70 ºC, we injected 100 µL of the sample headspace into the heated injector (80 ºC) of the GC-MS equipment. Sample volatiles were conveyed from the injector to the capillary column (100% polythylene glycol, Agilent Technologies, CA, USA) by a constant carrier gas flow (10 mL/min of N_2). These volatiles were eluted at 80 ºC for 200 seconds and were driven to the ionization chamber (tritium ionization source) prior to the IMS detection. Finally, we set the drift gas flow and tube temperature, respectively, to 150 mL/min and 65 ºC. Data was acquired by the commercial software LAV version 2.0.0 (G.A.S., Dortmund, Germany). As a result, we obtained a matrix of dimension L x M (9990 x 3000). This process was repeated until the completion of our dataset (N x L x M, for N=53). First sample of the dataset was selected as a reference signal pre-processing (see Fig.1).

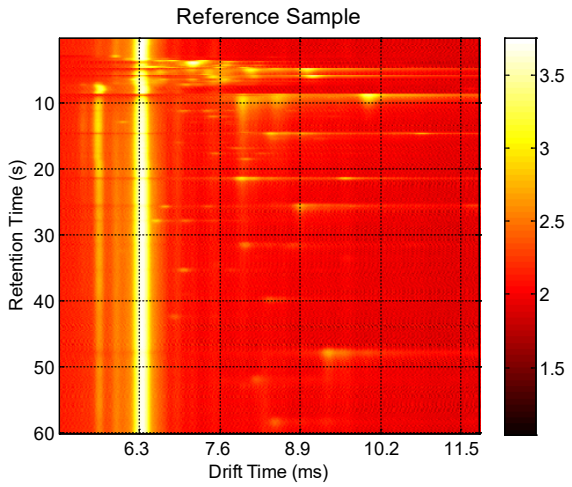

Fig. 1. Data matrix of the reference ham sample. Light/Dark tones denote High/Low intensity values (in decimal logarithmic scale).

3. METHODS

3.1. Signal pre-processing

After data acquisition stage, we applied several signal pre-processing steps to the raw data to facilitate the labor of the later classifier. The explicit sequence of signal pre-processing steps was: 1) Peak alignment, on the retention time axis; 2) Denoising, 3) Baseline removal, 4) Peak alignment, and 5) Selection of the region of interest, on the drift time axis; and finally, 6) Decimation, on the retention time axis.

1) We performed peak alignment on the retention time axis with the aim of standardizing the time at which each of the different chemical compounds (ions) were leaving the chromatographic column that was coupled to our IMS instrument. This task was accomplished by detecting the retention time of ions that were common in all the ham samples, and after that, correcting their position with respect a reference sample. To do it, we first added the intensity values in a sample along drift time dimension, for all the samples in the dataset. That provided N time sequences of length L, which presented a few common peaks. The common peaks were then aligned in each sample by means of cubic splines interpolation.

2) Noise contribution along the drift time dimension was decreased by applying a Saviltky-Golay filter (length, n=19 time samples; polynomial degree, deg=2; and derivative order, der=0). Filter specifications were selected so as to maximize the Signal to Noise ratio of the spectra under the constraint of not reducing the height of the Reactant Ion Peak (RIP) in more than 1%.

3) Baseline removal along the drift time axis was a necessary step in the pre-processing of GC-IMS signals because the area or height of the peaks in a spectrum was directly related to the number of ions of each species present in the sample. Thus, to properly compare the composition of our set of samples, their baseline had to be previously corrected. In particular, we used a variation of the Asymmetric Least Squares method for baseline removal known as *psalsa* less prone to overestimate the baseline under peaks with high intensity values [3]. *Psalsa* algorithm depends on three parameters: a penalty on the second derivative (λ), a penalty on the value of the baseline with respect the value of the signal (p), and third parameter (K), that modifies substantially the value of p in the event of large intensity peaks. These parameters were selected by visual inspection on the reference sample ($\lambda = 10^{-4}$, $p = 5 \cdot 10^{-3}$, and $K = 200$).

4) Next, we aligned the drift time axis by means of multiplicative correction that forces the RIP position of every to be the same. This drift time corresponded to the maximum value of the reference sample ($t_d = 6.4$ ms), for a retention time at which all the compounds have left the chromatographic column and the IMS was already clean ($t_r = 190$ s). We note that, due to distortion introduced by our multiplicative correction in the drift time axis, an additional cubic splines interpolation was needed to obtain again a constant sampling period. Figure 2a) shows an example of data misalignment on the drift time axis for three different samples (sample 22, 41, and the reference). After drift time correction, all three samples become aligned, as can be observed in Figure 2b).

In order to reduce the high dimensionality of our samples, we decided to 5) limit the region of analysis to drift times ranging from 5 to 12 ms, since no peaks were found out of this region, and 6) to decrease the resolution of the retention time axis by decimating a factor 5 its sampling period. Finally, we reshaped the data samples so that they could be understandable by the PLSDA classifiers. Basically, we concatenated in a single row all the spectra of sample to create a feature vector. As a measure of goodness for sample correction, we computed the Fisher Score of the Principal Component Analysis (PCA) projection of the data (2 PC's).

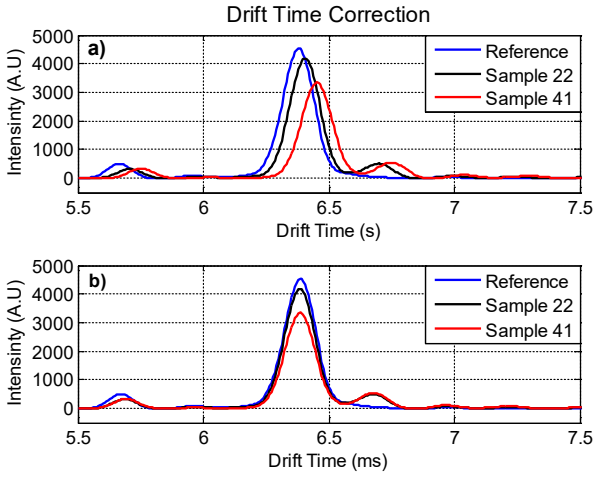

Fig. 2. Position of the reactant ion of samples 22, 41, and the reference sample (t_r =190 ms) for a) before the alignment, and b) after the multiplicative drift time correction.

3.2. Sample Classification

We employed the pre-processed data to build 4 Partial Least Squares Discriminant Analysis (PLDA) models in order to classify ham samples according to the feed received by the Iberian pigs: either nourished with feed (C1) or with acorns (C2). Each of these models corresponded to a different training/test balanced partitions of the dataset (with 40 and 12 samples respectively, for training and test sets). We note that the reference sample was not used for creating the data models. The rationale for creating several classifiers was to ensure that our predictive results were consistent with independence of the sample arrangement and the brand of ham provider. To prevent from overfitting, the level of complexity of PLSDA models (that is their number of latent variables) was limited by cross-validation. More precisely, we used the random subsampling method to divide the training set in two balanced splits. First split was composed of 32 samples, and was used to create PLSDA models, whereas the remaining 12 where utilized for validating them using the classification rate (CR) as a figure of merit. The complexity of the models ranged from 1 to 6 latent variables. This process was repeated 100 times, for each of the possible training/test partitions. Once the complexity of the models was stablished, we created 4 PLSDA models (one per dataset partition) using the whole training set of samples. The predictive performance of these models was checked against their corresponding test set samples employing again the CR as a figure of merit.

4. RESULTS

Data correction was performed following our signal pre-processing workflow. The Fisher Score of the PCA data projections increased after signal pre-processing (from 0.8 to 1.2) suggesting that the spurious data variance was downweighed. The cross-validation stage revealed that an increment on model complexity increased model's predictive performance up to 4 latent variables (for each training/ test partition of the dataset). Beyond this point, they experienced a fast degradation in sample class prediction. Consequently, final PLSDA models were created using 4 latent variables. Figure 3 shows the scores plot (first 2 latent variables) for a PLSDA model obtained from the first training/test partition. We represented C1 samples using square markers, while C2 samples with triangular ones. Training and test sets were plotted, respectively, in black and red colors. As can be seen on the figure, the PLSDA classifier obtained a perfect performance identifying samples regarding the Iberian pig's diet. In fact, only the last partition of the dataset exhibited slightly worse results (CR = 91.7 %), denoting the robustness of the corrected dataset to batching effects and to or the distinct processing parameters to manufacture ham carried out by the different food providers.

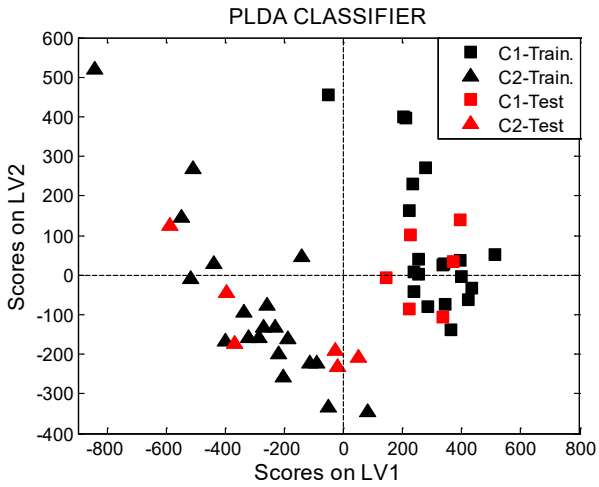

Fig. 3. Scores plot for a 2-latent variable PLSDA model corresponding to the first training/test partition of the dataset. C1 and C2 samples were represented, respectively, with square and triangular markers, while training/test sets in black/red colors.

5. CONCLUSIONS

In this work, we have used a GC-IMS to analyze to distinct types of Iberian ham samples. To reduce sample dissimilarity due to spurious sources of variance, intensive signal pre-processing was applied to the raw data. The corrected data was employed to build 4 different PLSDA model for different partitions of the dataset, obtaining nearly perfect classifications. This result suggest that the dataset is robust to batching effects and to the brand of ham provider.

6. ACKNOWLEDGMENTS

This work has been partially funded by CERCA Programme / Generalitat de Catalunya. The experimental part was carried out thanks to the support from the Government of Spain (DGICyT Grant CTQ2014-52939R).

7. REFERENCES

[1] N. Arroyo-Manzanares, A. Martín-Gómez, N. Jurado-Campos, R. Garrido-Delgado, C. Arce, and R. Arce, "Target vs spectral fingerprint data analysis for avoiding the labeling fraud using head-space gas chromatography-ion mobility spectrometry. *In press*, 2017.

[2] M. Camara, N.Gharbi, A. Lenouvel, M. Behr, C. Guignard, P. Orlewski, and D. Evers, "Detection and quantification of natural contaminants of wine by gas chromatography-differential ion mobility spectrometry (GC-DMS)." *Journal of agricultural and food chemistry,* ACS Publications, Volume 61, Issue 5, pp. 1036-1043, 2013.

[3] S. Oller-Moreno, S. Pardo, L. Jiménez-Soto, J. M. Samitier, and S.Marco, "Adaptative Asymmetric Least Squares baseline estimation for analytical instruments." In *Systems, Signals &Devices (SSD),11ht International Multi-Conference on*, IEEE, pp. 1-5, 2014.