COMPARISON OF TOBACCO EXTRACTS: A STRATEGY BASED ON GCxGC-MS ANALYSIS AND CHEMOMETRIC DATA PROCESSING

Benoît Pezous[a], Isabelle Rivals[a], Didier Thiébaut[b], Patrick Sassiat[b], Sreedhar Mallipattu[b], Béatrice Teillet[c], Jérôme Vial[b]

[a]Équipe de Statistique Appliquée, ESPCI ParisTech, 10 rue Vauquelin, 75005 Paris, France.
[b]Laboratoire Environnement et Chimie Analytique, ESPCI ParisTech, UMR CNRS UPMC PECSA, 10 rue Vauquelin, 75005 Paris, France.
[c]Groupe Imperial Tobacco France, SEITA, Centre de Recherche, 4 rue A. Dessaux, 45 404 Fleury les Aubrais, France.

Abstract

The extraction of tobacco, obtained either by Likens Nickerson or supercritical fluid extraction, generates samples that contain a high number of different compounds. The characterization and the comparison of these tobacco extracts require the use of a powerful analytical technique. Thanks to its high peak capacity enabling the separation of several hundreds of compounds in a single run, comprehensive two-dimensional gas chromatography (GCxGC) is now considered as the most efficient technique for the analysis of complex volatile mixtures.

However, taking into account all the information obtained in the whole 2D chromatogram for comparison purposes is far from trivial. A simple transposition of traditional data processing based on individual peak integration would be tedious and time consuming, given the large number of peaks often observed. Thus, the data processing is often reduced to a rather subjective visual examination of color plots to determine which spots are similar and which are different. This is the reason why a strategy based on chemometrics is proposed for the global comparison of the GCxGC chromatograms.

The idea is to perform a data processing based on picture comparison. Each point of the color plot is taken as a response and multivariate analysis tools are used to determine to what extent these responses differ or not from a sample to another. Nevertheless, this strategy failed to give satisfactory results because the comparison is blurred by variability of retention times, especially along the second chromatographic dimension. To handle this problem, a preliminary alignment of the chromatograms is carried out. The effectiveness of a time alignment strategy, Dynamic Time Warping, is demonstrated. In addition, discriminant components were obtained from a comparison of pictures based on point to point correlation, and identified with Mass Spectrometry data.

The proposed approach was successfully applied to the comparison of the volatile fraction of tobacco extracts to discriminate various types of tobaccos.

INTRODUCTION

The properties of tobaccos are influenced by genetics, agricultural processes, soil type and nutrients, climatic conditions, plant disease, stalk position, harvesting and curing procedures. A change in any of these factors can markedly alter the chemical composition of leaves and thus affect smoking quality. The total number of chemical constituents in tobacco leaves exceeds 4000 [1]. Understanding the relationship between the chemical composition and the flavour and the aroma of tobacco requires an analysis of these components. Then, the results may be used to classify and select different types of tobacco, for comparison purposes, for quality control, to discover new compounds or to characterize the chemical classes of compounds. The coupling of a rapid extraction technique and comprehensive GCxGC was shown to allow a fast analysis of representative tobacco samples (Burley, Virginia and Oriental) [2]. The comprehensive GCxGC is a two dimensional technique (2D) that is suitable for to the analysis of complex volatile samples including thousands of compounds. Nevertheless such analyses generate a large amount of data that can not be processed as usually done for 1D technique. To our knowledge there is no general strategy described in the literature to perform a global comparison of GCxGC chromatograms. In this paper, it will be shown how signal processing techniques combined with multivariate and correlation analyses allow to select the responses that differ from a sample to another. Then, these responses will be related to corresponding chemical compounds thanks to the coupling between GCxGC and Mass Spectrometry data.

EXPERIMENTAL

GCxGC apparatus

A trace GCxGC system from Thermo-Electron Corporation (Courtaboeuf, France) equipped with a Merlin Microseal injector (Merlin Instrument Company, CA, US) was used. It was fitted out with a double jets carbon dioxide cryogenic modulator, and a split/splitless injector. To avoid discrepancies related to a poor trapping of the compounds in the modulator, the two jets were placed closer to the column than in the original configuration. The set of columns presenting the best compromise both in terms of separation and ageing was as follows. The

first column was an apolar capillary column VF-1 ms, Varian (Les Ulis, France), 15 m x 0.25mm, 1.0µm. This column was connected to a DB 1701 1.5m x 0.1m, 0.1µm from Agilent Technologies (Waldbronn, Germany). Connections between columns were made using deactivated presstight connectors from Restek (Evry, France). The flow rate of the carrier gas was 1mL/min and the injector was set at 240°C. In order to have a preconcentration of the solutes at the beginning of the column by recondensation, a cold trapping was applied to the splitless injection. The temperature program used started at 40°C for 40 s, then an increase at 60°C/min was applied up tol 70°C, and after a hold for of 3min, 2.5°/min were applied up to 240°C. The injected volume was 2µL. The injection was carried out in splitless mode with a surge of 400kPa during 40s. A typical modulation period of 5s was used. Detection was carried out with the quadripolar mass spectrometer DSQI (Thermo-Electron). The transfer line was set at 250°C. Classical electron ionization (70eV) was used; only the mass range was limited to 40-240 m/z so that the acquisition frequency (around 30 Hz) was compatible with GCxGC data. Excalibur software was used for data acquisition; then, data were imported into Hyperchrom S/W software for the visualization of 2D chromatograms. Hyperchrom S/W offers the possibility to export the 2D chromatogram as a text file. The matrix obtained could then be read into Matlab (R2008b, The MathWorks, Natick, MA, USA) for data processing.

Gases

Liquid CO_2 was of industrial grade and purchased from Air Liquide (Le Plessis Robinson, France). Pure gases, i.e. helium (99.9995%) and CO_2 (99.999%, for supercritical fluid extraction) were purchased from Messer (Asnières, France).

Tobacco extracts

Three types of tobacco were considered: Burley, Virginia and Oriental. For each type, several different samples were available, corresponding to different batches or different origins. One extract was available for each sample.

FVN data set

A first set of tobacco extracts was provided by the Imperial Tobacco group. They were obtained by the "Likens Nickerson" [3] process directly from tobacco leaves cut into small pieces. At the end of the process, extracts were in hexane.

SFE data set

Another set of extracts was generated at the LECA by supercritical fluid extraction. Extractions were performed with a Suprex SFE Prepmaster GA apparatus (Pittsburgh, PA, US). The experimental protocol consisted in putting tobacco samples into a 5 mL SFE cell. Then extractions were performed in static mode for 5 min, and in dynamic mode for 30 min with a CO_2 density of 0.4 at a temperature of 150°C [4]. Extracted compounds were collected after the pressure was released by bubbling in 3 successive vials, each filled with 3 mL of hexane: ethyl acetate (50:50 v/v) mixture. The injected sample corresponded to the first 3 mL, as it was checked that no compound was present in the following vials. Two extracts were available per sample. Figure 1 and Table 1 summarize the sources of variability of the used data sets. In order to limit chromatographic variability, all samples were analyzed in the shortest possible period of time.

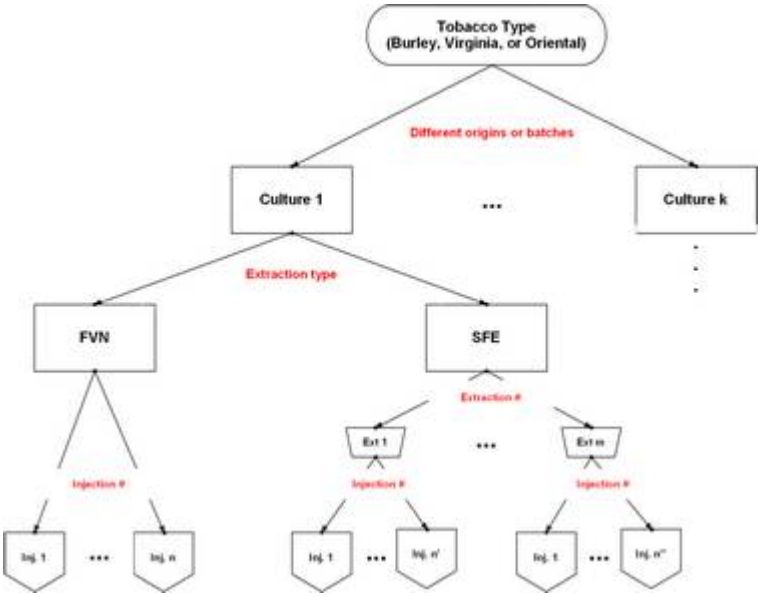

Figure 1. Sources of variability in the two data sets.

Tobacco Type	Burley		Virginia		Oriental	
Extraction Type	**FVN**	**SFE**	**FVN**	**SFE**	**FVN**	**SFE**
Number of different samples	4	4	4	4	4	12
Number of extractions	1	2	1	2-3	1	2
Number of Injections	2-3	2-5	2-3	2-3	2-3	2-3

Table 1. Sources of variability and their modalities.

BASICS ON GC x GC

Comprehensive GCxGC has been largely published [2 and 5-10]. The separation is operated on two GC columns, one being apolar, the other polar. Thanks to a modulator, each peak eluting from the first column is cut into several parts. Then, each cut is re-injected in the second column, allowing a fast separation. It results in very sharp peaks (100-200 ms width). Figure 2 explains how raw acquisition data are converted into 2D chromatograms: two overlapping peaks x and y emerging from the first column are resolved in GCxGC. Raw data of a GCxGC run is a large series of high speed second dimension chromatograms, which are usually stacked side by side to form a two-dimensional chromatogram with one dimension representing the retention time on the first column and the other, the retention time on the second column. Visualization is usually done by means of color or contour plot to indicate the signal intensity.

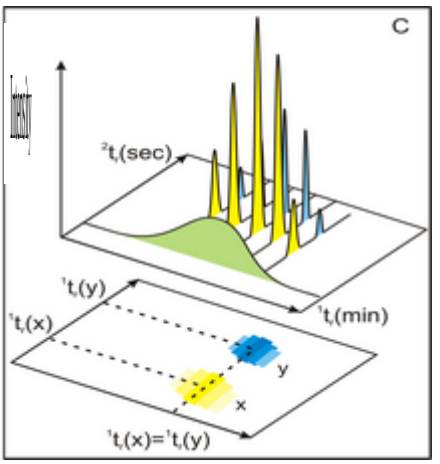

Figure 2. Visualisation of a GCxGC chromatogram.

DATA PROCESSING

We have considered that interpreting and comparing GCxGC chromatograms can be similar to analysing and comparing images. Indeed, each point of the chromatogram is a pixel of an image and is characterized by three numbers: two time coordinates (retention times along the first and second chromatographic dimensions) and a value for signal intensity. Thus, if each pixel is considered as a response, multivariate analysis can be used to compare samples in the space defined by these responses. In this paper, Principal Component Analysis (PCA) [11] has been used to check the relevance of the different processing steps applied to the GCxGC data: the more the different types of tobacco are separated on a PCA score plot, the more the studied processing of the data is relevant. We have already demonstrated the interest of combining time re-alignment algorithms and multivariate analysis [12]. In this paper, data pre-processing, peak alignment, and discriminant pixels selection have been studied and optimized thanks to PCA score plots. Matlab and C routines were developed for these processing steps.

Data pre-processing

To compensate for the inherent variability inherent to GC injection, a pre-processing of the data was required.

Background correction

Correcting the baseline of a unidimensional signal is a classic pre-treatment before analyzing chromatograms. In GCxGC unidimensional signals turn into images, a baseline correction amounts to a background correction. Our method of background subtraction was inspired from a DNA microarray preprocessing algorithm [13].

First, the image is divided into N rectangular zones (typically N = 60 or N = 80). Then the pixels of each zone i are ranked and the lowest 2% are chosen as the background of the studied zone i. Nevertheless a simple union of the backgrounds found would not be a satisfactory estimate of the whole image background. That is the reason why a smoothing adjustment is performed. For this purpose, we compute distances between each pixel of the image and the N zone centers. A weighted sum is then calculated based on the reciprocal of a constant plus the

square of the distances to all the zone centers. The smooth backgrounds obtained in this manner were subtracted to the chromatograms.

Normalization

A simple normalization on the mean intensity of the chromatograms has been operated so that the mean of each chromatogram equals 1.

Dynamic Time Warping

Because of the short separation time along the second chromatographic dimension (about 5s), the analysis along the latter dimension is very sensitive to experimental conditions (pressure, temperature…). Consequently peak misalignment may occur along this separation (see Figure 3). A comparison of chromatograms where peaks have different positions on the second dimension would be meaningless. Thus, to align the peaks, Dynamic Time Warping (DTW) was used. This method has been developed to address speech recognition issues [14], and Wang & Isenhour first proposed in 1987 [15] to apply it to chromatographic signals.

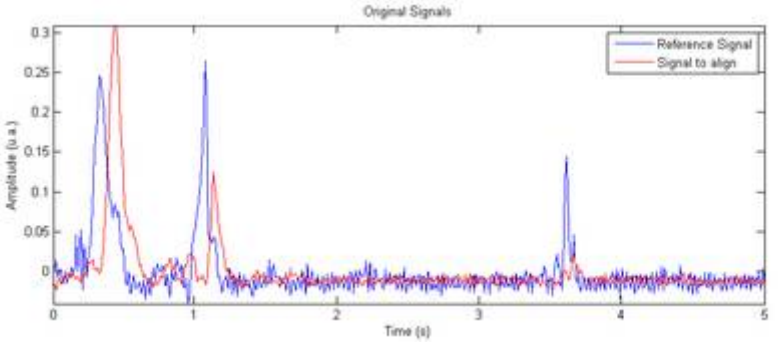

Figure 3. Unaligned signals (both come from Burley samples).

DTW algorithm

Let us represent the signals by R (Reference) and by S (Signal to align on R). These signals are sequences of length *n*:

$$R = r_1, r_2, \ldots, r_n$$
$$S = s_1, s_2, \ldots, s_n$$

To align these signals, we construct an n-by-n matrix D containing the Euclidean distances between the points r_i and s_j:

$$D(i, j) = (r_i - s_j)^2$$

A warping path W is a set of matrix elements that defines a mapping between R and S. The k^{th} element of W is defined as $w_k = (i,j)_k$. We have hence $W = w_1, w_2, \ldots, w_K$, with $n \leq K \leq 2n - 1$. The cost of the warping W between R and S is defined as the cumulated distance:

$$J(R, S) = \sum_{k=1}^{K} D(w_k)$$

An optimal warping W_{opt} minimizes this cost, i.e.:

$$W_{opt} = \arg\min_{W} \left(J(R, S) \right)$$

Furthermore, the optimal path must satisfy several conditions:

- boundary conditions: $w_1 = (1,1)$ and $w_K = (n,n)$,
- path continuity: given $w_k = (i,j)$, then $w_{k-1} = (i',j')$ with $i - i' \leq 1$ and $j - j' \leq 1$,
- path monotonicity : given $w_k = (i,j)$, then $w_{k-1} = (i',j')$ with $i - i' \geq 0$ and $j - j' \geq 0$.

This path can be found using dynamic programming. The dynamic programming matrix γ is computed recursively according to:

$$\gamma(i, j) = \min\left(\gamma(i, j-1) + D(i, j), \gamma(i, j) + 2 \times D(i, j), \gamma(i-1, j) + D(i, j)\right)$$

Once the entire matrix is filled, the optimal warping path is calculated in reverse order from $w_K = (n,n)$ to w_1 (see Figure 4). A simple synchronization algorithm then realigns the signal S on the reference R (see Figure 5).

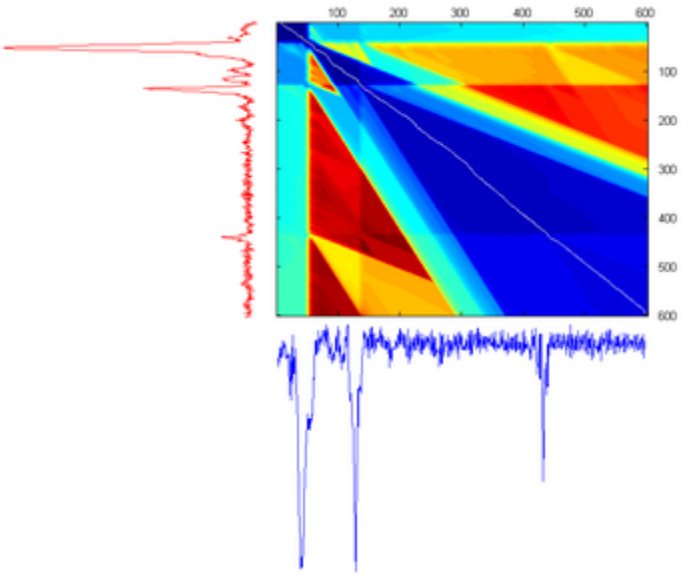

Figure 4. Dynamic programming matrix γ and warping path.

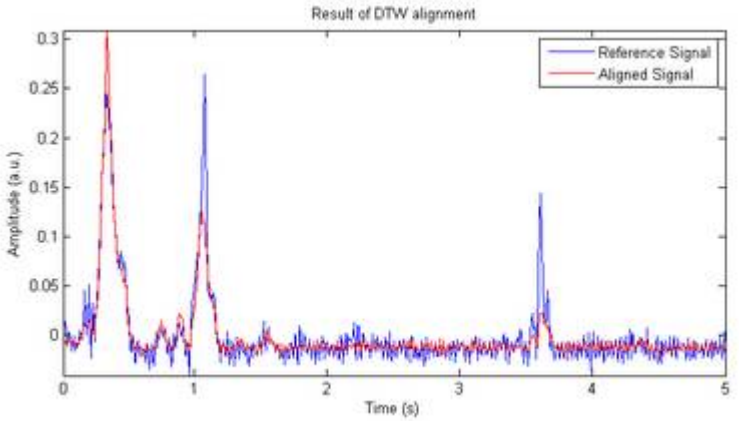

Figure 5. Result of the alignment.

Additional constraints

In order to improve the performance of the algorithm, additional constraints have been applied to the computation of the warping path. Indeed, two major pitfalls may occur during the construction of the path.

First, the path may follow a staircase trajectory, presenting long vertical and horizontal thresholds; this is quite inconvenient because these unrealistic thresholds will be present in the aligned signal. In order to solve this problem, a direct constraint on the slope of the warping path has been applied, by simply preventing the succession of two horizontal or two vertical segments.

Second, the path sometimes moves too far away from the matrix diagonal. As proposed by Keogh [16], a windowing on the dynamic programming matrix prevents the path from deviating. It merely consists in confining the path in a pipe centered on the matrix diagonal.

Alignment strategy

All the chromatograms have been realigned with the algorithm presented above. The windowing was performed with a rectangular, 40 pixels wide window. This algorithm was first developed with Matlab, and later implemented in C so as to save computation time (aligning a chromatogram with Matlab takes 1 min, whereas it only takes 3 s in C).

The realignment is operated on each column of each chromatogram. It is done in a supervised fashion inside each tobacco family: for each tobacco type, a reference is chosen on which all the others samples are realigned. We checked that the choice of the reference chromatogram has no influence on the results.

Selection of discriminant pixels

In order to find the pixels (and thus the compounds) that are characteristic of a tobacco type, we developed a method that compute correlation maps between pixel values and tobacco type.

Pixel to pixel correlation

This method is applied to a whole data set at one time. The algorithm scans all the chromatograms at a time pixel by pixel. Let us assume that one wants to find the pixels that are characteristic of Burley tobacco (see Figure 6).

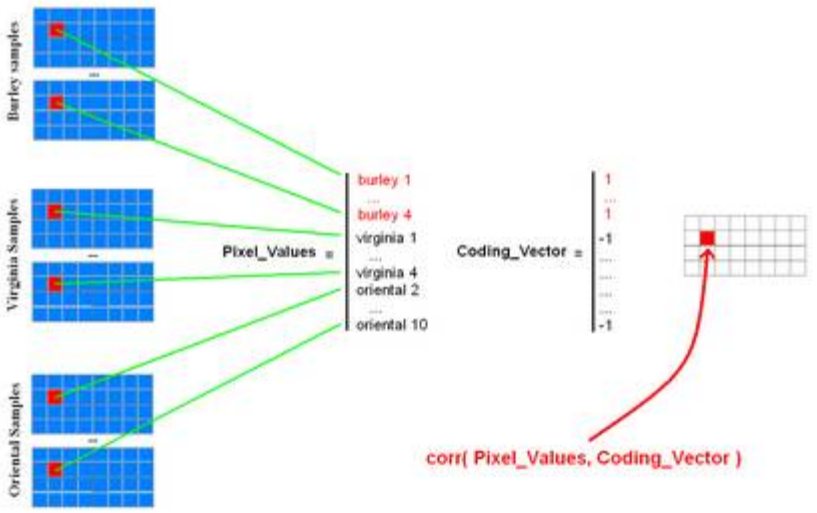

Figure 6. Illustration of the pixel to pixel correlation method.

For each pixel the algorithm computes the correlation between the pixel values vector and the coding vector (these codes belonging to the Burley family). Three cases may occur:

– the correlation is close to 1: the pixel corresponds to a compound *over expressed* in Burley samples;
– the correlation is close to -1: the pixel corresponds to a compound *under expressed* in Burley samples;
– the correlation is close to 0 : the pixel has no discriminant power.

Once all the chromatograms are scanned, a correlation map is obtained. A list of discriminant pixels is obtained by selecting a certain amount of pixels with close to unit correlation (typically 500 pixels).

Retrieving retention times

Due to the realignment, after the previous operation, the coordinates of the pixels can not be directly converted into the retention times of the corresponding compounds. Indeed, because of the DTW the coordinates of the pixels no longer correspond to the retention times. The warping breaks the linear relation between pixels coordinates and retention times. Thus, a reverse warping has to be applied to get back to the original coordinates of a pixel (its coordinates before DTW). This operation is simply made by reversing the warping path of the considered signal calculated during the alignment.

Then, the discriminant pixels have to be short-listed: several pixels may be found on a unique peak. A simple research of local maximum around discriminant pixels allows to find the maximum of the considered peak, and to define a chromatographic peak zone.

RESULTS AND DISCUSSION

The effect and the relevance of each step of the data processing is now evaluated according to the quality of the separation of the tobacco types on PCA score plots.

Effect of the data pre-processing

The background correction and the normalization reduced the differences of intensity between the chromatograms. Consequently, after these operations the dataset has become more homogenous. The results of PCA applied to the pre-processed data are presented on Figure 7. The different tobacco types are quite overlapping (particularly for FVN data), which confirms the need for a peak alignment process.

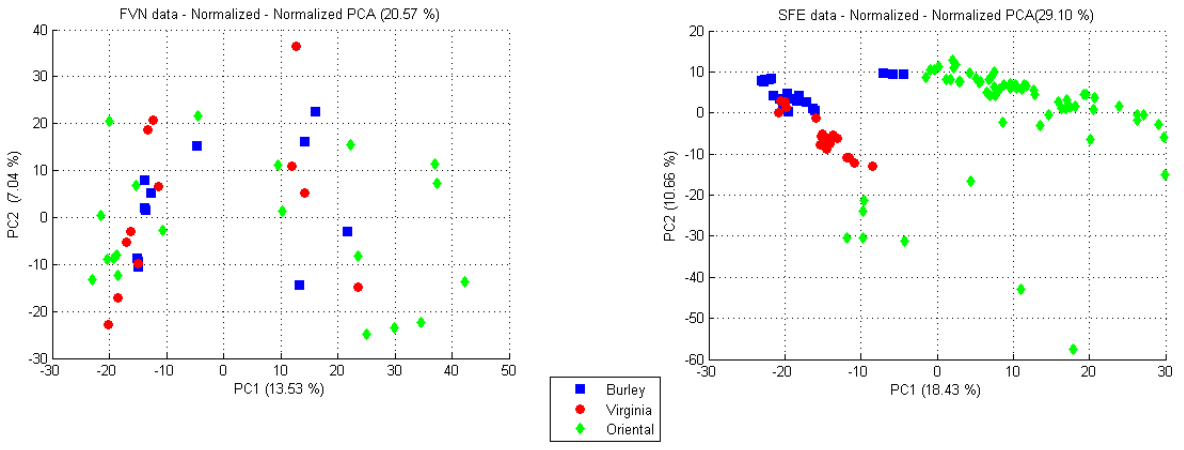

Figure 7. PCA score plot after data pre-processing.

Effect of the peak alignment

Figure 8 shows that the alignment enhances the discrimination between the different kinds of tobaccos. So, the three types are now visibly well separated.

However, if this analysis reveals the differences between the tobacco types, it does not provide information about what are these differences. This is the reason why the algorithm that selects the discriminant pixels has been developed.

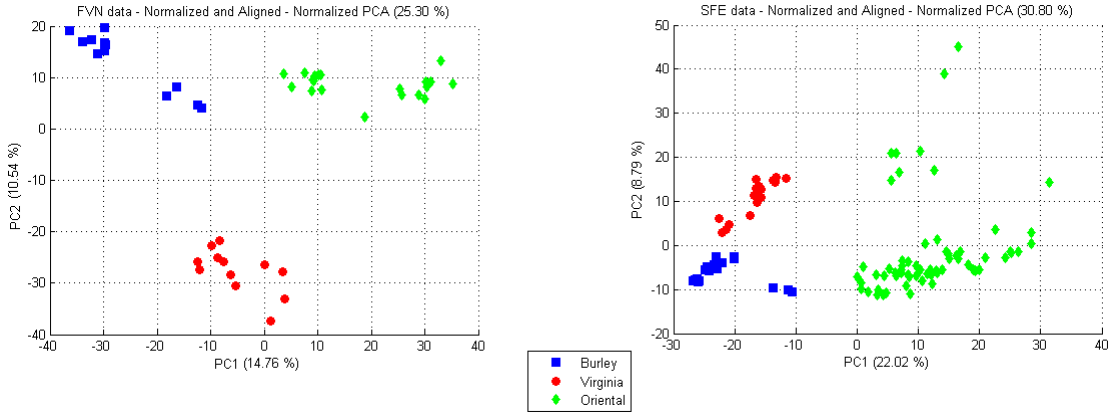

Figure 8. PCA score plots after the peak alignment.

Selecting discriminant pixels

The correlation maps described above have been computed for each data set and for each tobacco type. Then, for each map, the 500 pixels whose correlation is closest to 1 have been selected. After a reverse time warping, we can represent them on the original chromatograms (see Figure 9 and Figure 10).

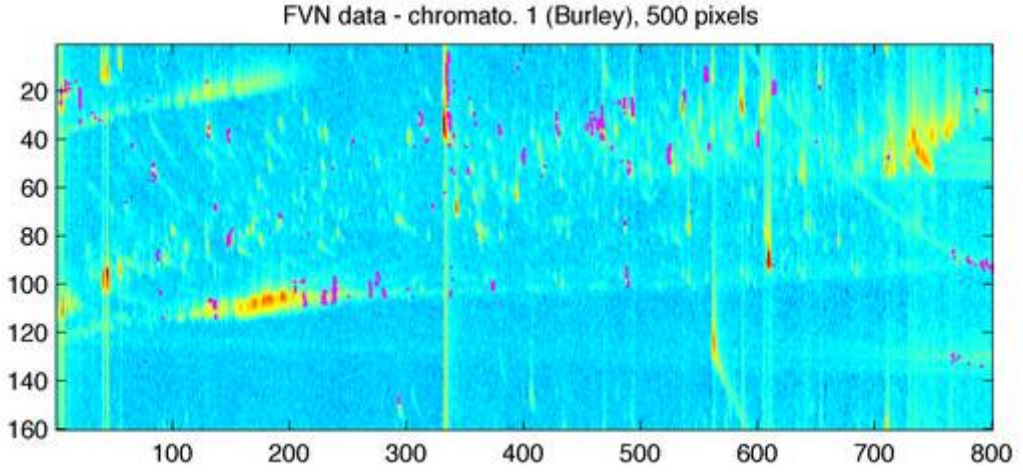

Figure 9. FVN Burley 500 discriminant pixels.

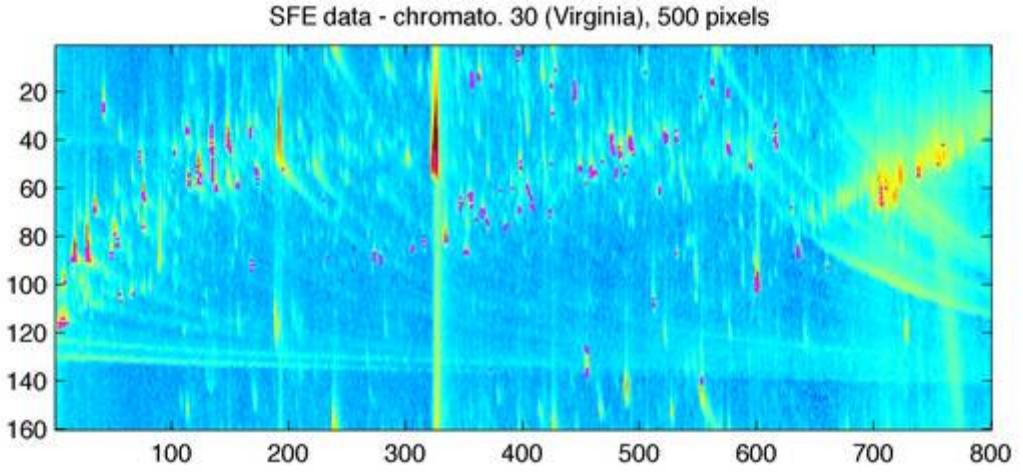

Figure 10. SFE Virginia 500 discriminant pixels.

Then, the selected pixels may be used as descriptors for the PCA: instead of taking the 128000 pixels of the chromatograms as descriptors, only the union of the 1500 (3 kinds of tobaccos x 500 pixels for each kind)

selected pixels is kept (see Figure 11). The separation between the tobacco types is obvious: The pixels selection has enhanced the differences between the chromatograms of each type. It proves the relevance of our pixel selection algorithm.

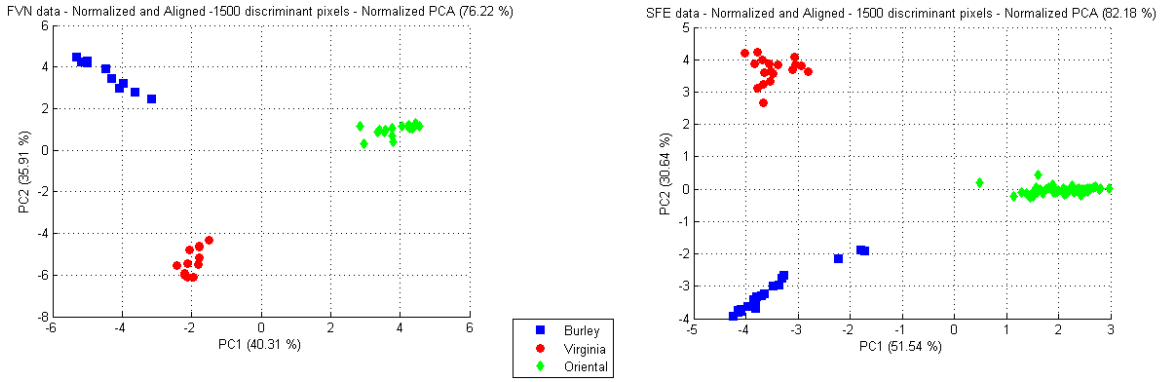

Figure 11. PCA after discriminant pixel selection.

Peak assignment

Once the discriminant pixels have been found, they have been short listed into peak areas. A peak selection algorithm has been applied so as to provide a list of discriminant peaks. Then, this list of peaks can be linked to Mass Spectrometry data in order to identify the discriminant compounds. An example of such a peak attribution is shown in Figure 12.

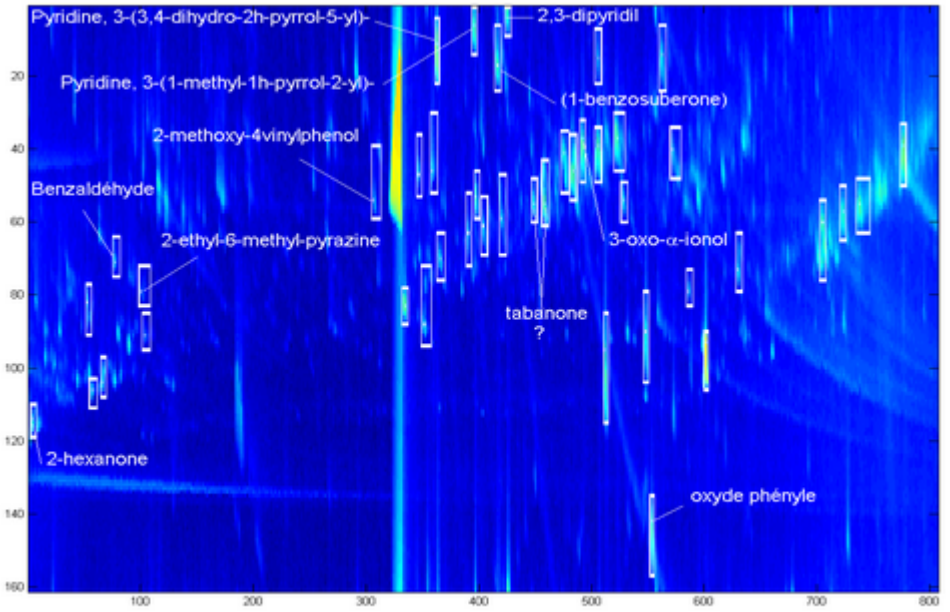

Figure 12. Discriminant peak attribution thanks to MS data for a SFE Burley sample.

CONCLUSION

This study demonstrates the relevance of a chemometric strategy to interpret and compare GCxGC chromatograms. PCA has allowed us to check the quality and the relevance of our different data processing steps. The performance of the data pre-processing, of the peak re-alignment, and of the discriminant pixel selection, have been studied on the yardstick of this multivariate analysis. It revealed that the peak re-alignment and the pixel selection enhance the differences between Burley, Virginia and Oriental tobacco samples. Moreover, the identification of discriminant pixels has been linked to discriminant compounds thanks to Mass Spectrometry data. This opens wide outlooks in (i) the development of an automatic processing strategy to interpret GCxGC chromatograms; (ii) the discrimination of origin and grades inside a tobacco type; (iii) the study of the impact of the variability of extraction process on the effectiveness of the discrimination.

REFERENCES

[1] Leffingwell, J.C., Basic Chemical Constituents of Tobacco Leaf and Differences among Tobacco Types, in Tobacco - Production, Chemistry and Technology, D.L. Davis and M.T. Nielsen, Editors, Blackwell Science, Ltd., 1999, pp. 265-284.

[2] J. Vial, D. Thiébaut, P. Sassiat, M.S. Beldean-Galea, M.J. Gomez Ramos, M. Bouzige, Coresta Congress 2006.

[3] S.T. Likens, G.B. Nickerson, Am. Soc. Brew. Chem. Proc. 22 (1964) 5.

[4] J.Vial, D.Thiébaut, P.Sassiat, M.S. Beldean-Galea, M. J. Gomez Ramos, G. Cognon, S. Mallipattu, B. Teillet, M. Bouzige, J. Chrom. Sci accepted for publication.

[5] J.B. Phillips, J. Beens, J. Chromatogr. A 856 (1999) 331.

[6] J. Dalluge, J. Beens, U. A. T. Brinkman, J. Chromatogr. A 1000 (2003) 69.

[7] M. Adachour, J. Beens, R. J. J. Vreuls, U. A. T. Brinkman, J. Sep. Sci 25 (2006) 438.

[8] T. Gorecki, J. Harynuk, O. Panic, J. Sep. Sci 27 (2004) 359.

[9] P. J. Marriott, T. Massil, H. Hogel, J. Sep. Sci 27 (2004) 1273.

[10] R. Shellie, P. Marriott, Flavour Fragr. J. 18 (2003) 179.

[11] T.W Anderson, An introduction to Multivariate Analysis (1958) (Second edition: 1984). J. Wiley, New York.

[12] J.Vial, et al. J. Chromatogr A (2008), doi:10.1016/j.chroma.2008.09.027.

[13] Statistical Algorithms Description Document: http://www.affymetrix.com/support/technical/whitepapers/ sadd_whitepaper.pdf

[14] H.Sakoe, S. Chiba, IEEE Trans. Acoustics Speech Signal Process. 26 (1978):43-49.

[15] C.P. Wang, T.L. Isenhour, J.Anal. Chem. 59 (1987), pp. 649-654.

[16] E. Keogh, C.A. Ratanamahatana, Knowledge and Information Systems 7 (2005), pp. 358-386.