jove

Chromatographic Fingerprinting by Template Matching for Data Collected by Comprehensive Two-Dimensional Gas Chromatography

Federico Stilo[1], **Chiara Cordero**[1], **Carlo Bicchi**[1], **Daniela Peroni**[2], **Qingping Tao**[3], **Stephen E. Reichenbach**[3,4]

[1] Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino [2] SRA Instruments [3] GC Image LLC [4] Computer Science and Engineering Department, University of Nebraska

Corresponding Authors

Chiara Cordero
chiara.cordero@unito.it

Stephen E. Reichenbach
reich@cse.unl.edu

Abstract

Data processing and evaluation are critical steps of comprehensive two-dimensional gas chromatography (GCxGC), particularly when coupled to mass spectrometry. The rich information encrypted in the data may be highly valuable but difficult to access efficiently. Data density and complexity can lead to long elaboration times and require laborious, analyst-dependent procedures. Effective yet accessible data processing tools, therefore, are key to enabling the spread and acceptance of this advanced multidimensional technique in laboratories for daily use. The data analysis protocol presented in this work uses chromatographic fingerprinting and template matching to achieve the goal of highly automated deconstruction of complex two-dimensional chromatograms into individual chemical features for advanced recognition of informative patterns within individual chromatograms and across sets of chromatograms. The protocol delivers high consistency and reliability with little intervention. At the same time, analyst supervision is possible in a variety of settings and constraint functions that can be customized to provide flexibility and capacity to adapt to different needs and goals. Template matching is shown here to be a powerful approach to explore extra-virgin olive oil volatilome. Cross-alignment of peaks is performed not only for known targets, but also for untargeted compounds, which significantly increases the characterization power for a wide range of applications. Examples are presented to evidence the performance for the classification and comparison of chromatographic patterns from sample sets analyzed under similar conditions.

Introduction

Comprehensive two-dimensional gas chromatography combined with the time-of-flight mass spectrometric detection

(GC×GC-TOF MS) is nowadays the most informative analytical approach for the chemical characterization of complex samples[1,2,3,4,5]. In GC×GC, columns are

serially connected and interfaced by a modulator (e.g., a thermal or valve-based focusing interface) that traps eluting components from the first dimension (^{1}D) column before their re-injection into the second dimension (^{2}D) column. This operation is done within a fixed modulation time-period (P_M), generally ranging between 0.5–8 s. By thermal modulation, the process includes cryo-trapping and focusing of the eluting band with some benefits for the overall separation power.

Although GC×GC is a two-dimensional separation technique, the process produces sequential data values. The detector analog-to-digital (A/D) converter obtains the chromatographic signal output at a certain frequency. Then, data is stored in specific proprietary formats which not only contains the digitalized data but related metadata (information about the data) as well. The A/D converter employed in GC×GC systems helps in mapping the intensity of the chromatographic signal to a digital number (DN) as a function of time in the two analytical dimensions. Single-channel detectors (e.g., flame ionization detector (FID), electron capture detector (ECD), sulfur chemiluminescence detector (SCD), etc.) produce single values per sampling time, whereas multichannel detectors (e.g., mass spectrometric detector (MS)) produce multiple values (typically, over a spectral range) per sampling time along the analytical run.

To visualize ^{2}D data, elaboration starts with rasterization of a single modulation period (or cycle) data values as a column of pixels (picture elements corresponding to detector events). Along the ordinate (Y-axis, bottom-to-top) the ^{2}D separation time is visualized. Pixel columns are sequentially processed so that the abscissa (X-axis, left-to-right) reports ^{1}D separation time. This ordering presents the ^{2}D data in a right-handed Cartesian coordinate system, with the ^{1}D retention ordinal as the first index into the array.

Data processing of ^{2}D chromatograms gives access to a higher level of information than raw data, enabling ^{2}D peak detection, peak identification, extraction of response data for quantitative analysis, and cross-comparative analysis.

The ^{2}D peak patterns can be treated as the sample's unique fingerprint and detected compounds as minutiae features for effective cross-comparative analysis. This approach, known as template-based fingerprinting[6,7], was inspired by biometric fingerprinting[6]. Automatic biometric fingerprint verification systems, in fact, rely on unique fingertip characteristics: ridge bifurcations and endings, localized and extracted from inked impressions or detailed images. These characteristics, named minutiae features, are then cross-matched with available stored templates[8,9].

As mentioned above, every GC×GC separation pattern is composed of ^{2}D peaks rationally distributed over a two-dimensional plane. Each peak corresponds to a single analyte, has its informative potential, and can be treated as a single feature for comparative pattern analysis.

Here, we present an effective approach for chemical fingerprinting by GC×GC-TOF MS featuring tandem ionization. The goal is to comprehensively and quantitatively catalog features from a set of chromatograms.

Compared to existing commercial software or in-house routines[10,11] that employ a peak-features approach, template-based fingerprinting is characterized by high specificity, efficiency, and limited computational time. In addition, it has an intrinsic flexibility that enables the cross-alignment of minutia features (i.e., ^{2}D peaks) between severely misaligned chromatograms as those acquired by different instrumentation or in long-time frame studies[12,13,14].

The basic operations of the proposed method are described briefly to guide the reader to a good understanding of the ^{2}D pattern complexity and information power. Then, by exploring the instrument output data matrix, chemical identification is performed and known targeted analytes located over the two-dimensional space. The template of targeted peaks is then built and applied to a series of chromatograms acquired within the same analytical batch. Metadata related to retention times, spectral signatures, and responses (absolute and relative) are extracted from re-aligned patterns of targeted peaks and adopted to reveal compositional differences in the sample set.

As an additional, unique step of the process, a combined untargeted and targeted (UT) fingerprinting is also performed on pre-targeted chromatograms to extend the fingerprinting potential to both known and unknown analytes. The process produces a UT template for a truly comprehensive comparative analysis that can be largely automated.

As a final step, the method performs the cross-alignment of features in two parallel detector signals produced with high and low electron ionization energies (70 and 12 eV).

The protocol is quite flexible in supporting analyses of a single chromatogram or a set of chromatograms and with variable chromatography and/or multiple detectors. Here, the protocol is demonstrated with a commercially available GC×GC Software suite (see **Table of materials**) combined to a MS library and search software (see **Table of Materials**). Some of the necessary tools are available in other software and similar tools could be implemented independently from descriptions in the literature by Reichenbach and co-workers[15 , 16 , 17 , 18 , 19]. Raw data for the demonstration is derived from a research study on extra-virgin olive (EVO) oil conducted in the authors' laboratory[14]. In particular,

the volatile fraction (i.e., volatilome) of Italian EVO oils is sampled by headspace solid phase microextraction (HS-SPME) and analyzed by GC×GC-TOF MS to capture diagnostic fingerprints for quality and sensory qualification of samples. Details on samples, sampling conditions, and analytical set-up are provided in the **Table of Materials**.

Steps 1–6 describe pre-processing of the chromatograms. Steps 7–9 describe processing and analysis of individual chromatograms. Steps 10–12 describe template creation and matching, which are the basis for cross-sample analysis. Steps 13–16 describe applying the protocol across a set of chromatograms, with steps 14–16 for UT analysis.

Protocol

1. Importing raw data

NOTE: This creates a two-dimensional raster array for visualization and processing.

1. Launch the image software.

2. Select **File | Import**; navigate to and choose the raw data file acquired by GC×GC-TOF MS system named "VIOLIN 101.lsc" (**Supplementary File 1**); then, click **Open**. The chromatogram opens in this software.
   NOTE: Raw data file format depends of the instrument manufacturer. The software imports a variety of file formats listed in the user's guide.

3. In the **Import** dialog, set the **Modulation Period (P_M)** to **3.5** s; then, click **OK**.
   NOTE: Some acquisition software may not record the modulation period.

4. Select **File | Save Image As**; navigate to the desired folder; enter the name "Oil 1 RAW.gci" (**Supplementary File 2**); then, click **Save**.

2. Shifting the modulation phase

NOTE: This puts all peaks in each modulation cycle into the same image column, including the peaks that wrap around the end of the modulation period into the void time of the next modulation period[20].

1. Select Processing | Shift Phase.

2. In the **Shift Phase** dialog, set the **Shift Amount** to -0.8 s; then, click **OK**.

3. Baseline correction[21]

1. Select **Graphic | Draw Rectangle**.

2. Click-and-drag to draw a rectangle in the image where no peaks are detected.

3. Select **Tools | Visualize Data**; note the mean and standard deviation of the detector signal, here, 21.850 ± 1.455 SD unit-less digital number (DN); then, close the tool.

4. Select **Processing | Correct Baseline**.

4. Coloring the chromatographic image using a value map and color map[20]

1. Select **View | Colorize**.

2. In the **Colorize** dialog, select the **Import/Export** tab; select the **#AAAA** (**Supplementary File 3**) custom color map provided as supplementary material; then, click **Import**.

3. On the **Value Mapping** controls, set the value range to the minimum and maximum values; then, click **OK**.

5. ^{2}D peaks (i.e., blobs) detection for analytes[18]

1. Select **Processing | Detect Blobs** with the default settings; then, observe that some peaks are split and there are spurious detections.

2. Select **Configure | Settings | Blob Detection**; then set **Smoothing** to 0.1 for the first dimension and 2.0 for the second dimension and set **Minimum Volume** (i.e., threshold for the summed values) to 1.00 E6; then, click **OK**.

3. Select **Processing | Detect Blobs** with the new settings; then, observe the improvements.

6. ^{2}D peaks filtration

NOTE: This is done to automatically remove meaningless detections due to column bleeds along the ^{1}D and strikes or tailings along the ^{2}D.

1. Select **Processing | Interactive Blob Detection**.

2. Note the blob detection settings; then, click **Detect**.

3. In the **Advanced Filter** builder, click **Add**; then, in the **New Constraint** dialog, select **Retention II**; then, click **OK**.

4. In the **Constraint** sliders, set the minimum and maximum ^{2}D retention times for the filter to reduce the number of false peaks without losing true peaks.

5. Click **Apply**; then, click **Yes** to save to the detection settings with the new filter.

NOTE: More advanced tools may be required to deal with particular detection problems, such as ion-peak detection or deconvolution for co-elutions[19].

7. Linear retention indices calibration

NOTE: Perform this step[22] (I^T) for the specific retention times across the set of retention index (RI) standards (typically *n*-alkanes).

1. Select **Configure | RI Table | Retention Index (Col I)**.

2. On the **RI Table Configuration** dialog, click **Import**; then, select the RI calibration file (in CSV format with name, retention time, and retention index) named "LRI table.csv" – (**Supplementary File 4**).

3. Select **File | Save Image A**. Navigate to the desired folder; enter the name "Oil 1 LRI CALIBRATED.gci" (**Supplementary File 5**); then, click **Save**.

8. Searching for the peak spectra in the NIST17 MS library [23]

1. Select **Configure | Settings | Search Library**.

2. In the **Search Library** dialog, set **Type of Spectrum** to Peak MS, **Intensity Threshold** to 100, **NIST Search Type** to Simple (Similarity), **NIST RI Column Type** to Standard Polar, and **NIST RI Tolerance** to **10**; then, click **OK**. NIST MS Search offers many other settings that are set to the defaults here.

3. Select **Processing | Search Library for All Blobs**.

9. Review and correct analyte identifications

1. On the tool palette, set the cursor mode to **Blob | Select Blobs**.

2. In the **Image** view, right-click on the desired peak.

3. On the **Blob Properties** dialog, inspect blob properties; then, click **Hit List**.

4. Inspect the hit list; then, if the identification is incorrect, select the checkmark beside the correct identification.

5. In the **Blob Properties** dialog, enter the **Group Name** to designate chemical class and any other desired metadata; then, click **OK**.

6. Select **File | Save Image As**; navigate to the desired folder; enter the name "Oil 1 COLORIZED for Template construction.gci" (**Supplementary File 6**); then, click **Save**.

   NOTE: This file is included in the supplemental archive, which can be opened for step 10.

10. Create a template with targeted peaks[15]

1. In the Image view (still in **Select Blobs** mode from step 9.1), select the desired peaks with a click on the first peak and **CTRL + click** on the additional peaks.

2. On the tool palette, click the **Add to Template** button.

3. When the template is complete, select **File | Save Template**; specify the folder and file name; then, click **Save**.

4. Select **File | Close Image**.

   NOTE: At this point, these instructions continue with the template created to include the desired target peaks, available as "Targeted tamplate.bt" (**Supplementary File 7**).

11. Match and apply the template

NOTE: Matching recognizes the template pattern in the detected peaks a new chromatogram. Applying the matching sets identifications and other metadata in the new chromatogram from the template.

1. Select **File | Open Image**; navigate to and select the "Oil 2 COLORIZED.gci" (**Supplementary File 8**) chromatogram file (which is pre-processed); then, click **Open**.

2. On the tool palette, set the cursor mode to **Template | Select Objects**.

3. Select **Template | Load Template**.

4. In the **Load Template** dialog, click **Browse**; navigate to and select the targeted peaks template "**Targeted template.bt**" (**Supplementary File 7**); then, click **Open**.

5. In the **Load Template** dialog, click **Load**, and then **Dismiss**.

6. In the Image view, right-click on a template peak; then, inspect its object properties, including the qCLIC and reference MS.

7. Select **Template | Interactive Match and Transform Template**.

8. In the **Interactive Match** interface, click **Match All**; then, review the matching results both in the table and in the image, in which each template peak is marked with unfilled circles and, if a match is made, there is a link to a filled circle for the detected peak.

9. Edit the matches as desired; when satisfied, click **Apply** to transfer metadata from the template to the chromatogram.

   NOTE: Matching constraints, such as the qCLIC, help match the correct pattern among the detected peaks of the new chromatogram. Constraint parameters include the type of MS signature used as template reference (*peak MS* or *blob MS*) and the threshold values for spectral similarity (Direct Match Factor (DMF) and Reverse Match Factor (RMF)). Here, parameters are set

based on previous studies[13, 14] to limit false negative matches: *peak MS* and DMF and RMF similarity threshold 700.

12. Transform the template for substantially different chromatography

NOTE: This step is not necessary unless chromatographic conditions vary substantially causing the template to be misaligned with a new chromatogram, such as can be the case over long-term studies or after a new column is installed. In such cases, the template can be geometrically transformed in the chromatographic retention-times plane to better fit the new chromatogram[12, 13]. In this example, the peak patterns of the template and chromatogram are similar, but differ in the retention-times geometry, such as would be seen for different chromatographic conditions.

1. Repeat steps 11.2–11.5, except navigate to, select, and load **Targeted template 2.bt** (**Supplementary File 9**).

2. Select **Template | Interactive Match Template**; then, click **Edit Transform**.

3. In the **Transform Template** interface, vary the ^{1}D and ^{2}D scales, translations, and shears to better align the template with the detected peaks; then, click **Transform Template**.

4. With the transformed template, click **Edit Match**; then, repeat steps 11.8–11.9.

13. Perform combined untargeted and targeted analysis across a set of chromatograms

NOTE: A combined untargeted and targeted (UT) template, also referred to as *feature* template[24, 25], when matched to each of a set of chromatograms, establishes correspondences between untargeted and targeted analytes,

then consistent cross-sample features are extracted for pattern recognition.

1. Perform pre-processing (steps 1–6) and UT template matching (steps 11.1–11.9) for all chromatograms in the set (i.e., ^{2}D chromatograms of oils). Alternatively, automate this step with project software or similar software, not described here.

2. Launch the Investigator software.

3. Select **File | Open analysis**; then, select, and open "Feature Jove su 70 eV.gca" (**Supplementary File 10**).

4. Click **OK** to open and examine the results.

5. Click on the **Compounds** tab to review metric values and statistics for specific analytes (i.e., targeted analytes with associated chemical names) or untargeted analytes with (#) identifiers aligned across all chromatograms, then perform the steps below.

    1. Click on the **Attributes** tab to review values and statistics for specific metrics across chromatograms.

    2. Click on the **Summary** tab to review the summary statistics for both compounds and features. If the chromatograms are from different classes, as in this case oils produced from olives harvested in two different regions of Italy, then the **Summary** tab lists Fisher ratio statistics (F and FDR), which provide insights into features for discriminating between classes.

    3. View various charts on all tabs and, if desired, perform Principal Component Analysis (PCA) on the **Attributes** tab.

14. Modify the UT template for parallel MS analysis

NOTE: The analysis was performed with both 70 eV and 12 eV (i.e., high and low) electron ionization energies[26 , 27].

1. Open one of the 12 eV chromatograms, e.g., "Oil 1 12 eV RAW.gci" (**Supplementary File 11**), perform pre-processing (steps 1–6) and load the UT template "UT template 70 relaxed.bt" (**Supplementary File 12**) as described in steps 11.1–11.6. Files are provided as supplementary material.

2. If necessary, adjust the template to fit the detected 12 eV peaks as described in step 12. Here, there is no significant misalignment because the tandem signals are multiplexed. However, it should be noted that because the different ionization settings produce different fragmentations, it is necessary to relax constraints for the qCLIC constraints on DMF and RMF spectral similarity (not demonstrated here).

3. Select **File | Save Template**; specify the folder and file name, e.g., "UT template 12.bt" (**Supplementary File 13**); then, click **Save**.

15. Perform combined untargeted and targeted analysis across 12 eV chromatograms

1. Select **File | Open analysis**; then select and open "Feature Jove su 12 eV.gca" - **Supplementary File 14** file provided.

2. Click **OK** to open and examine results.

3. Click on the **Compounds** tab to review metric values, refer to 12 eV responses and statistics for specific analytes (i.e., targeted analytes with associated chemical

names) or untargeted analytes with (#) identifiers aligned across all chromatograms, then perform the steps below.

1. Click on the **Attributes** tab to review values and statistics for specific metrics across chromatograms.

2. Click on the **Summary** tab to review the summary statistics for both compounds and features at 12 eV. If the chromatograms are from different classes, as in this case oils produced from olives harvested in two different regions of Italy, then the Summary tab lists Fisher ratio statistics (F and FDR), which provide insights into features for discriminating between classes.

3. View the various charts available on all the tabs and, if desired, perform **Principal Component Analysis (PCA)** on the **Attributes** tab.

Representative Results

GC×GC-TOF MS patterns of high-quality extra-virgin olive oil volatilome exhibit about 500 ^{2}D peaks above a signal-to-noise ratio (SNR) threshold of 100. Such a threshold was defined by previous investigations on food volatiles[14, 27] as the minimum relative signal over threshold to obtain reliable spectra for cross-comparative analysis. Components are distributed over the chromatographic space according to their relative retention in the two chromatographic dimensions, and specifically based on their volatility/polarity in the ^{1}D and volatility in the ^{2}D. Here, column combination is polar × semi-polar (i.e., Carbowax 20M × OV1701).

The ^{2}D pattern shows a high degree of order. Relative retention patterns for homologous series and classes are shown in **Figure 1A** with annotations (graphics for groups and bubbles for peaks) for linear saturated hydrocarbons (black), unsaturated hydrocarbons (yellow), linear saturated aldehydes (blue), mono-unsaturated aldehydes (red),

polyunsaturated aldehydes (salmon), primary alcohols (green), and short-chain fatty acids (cyano).

Detected ^{2}D peaks can then be identified by comparing the average MS spectrum extracted from the entire ^{2}D peak (*blob* spectrum) or from the largest spectrum (*apex* spectrum). **Figure 2** illustrates the output of the *apex* spectrum search for blob 5 and returns a high similarity match (first 10 hits) for *(E)*-2-hexenal. Databases explored are those pre-selected by the analyst in step 8 of the method.

The identification is validated by active retention indexing. The experimental I^T value was calculated for the ^{2}D peaks, so that at this stage the library search prioritizes results with coherent values of tabulated I^T. Tolerance windows can be customized based on analyst experience, reliability of reference database values according to stationary phase, and analytical conditions applied. New tools for smart calibration of linear retention indices without experimental calibration with *n*-alkanes, have been recently developed and discussed in a study by Reichenbach et al[19].

The collection of identified ^{2}D peaks (i.e., targeted peaks) can be adopted to build a template of targeted peaks to promptly establish reliable correspondences between the same compound across all sample chromatograms. The collection of targeted template peaks is visualized in **Figure 1B**. Red circles correspond to the 196 targeted compounds, including two Internal Standards (IS) linked to template peaks with connection lines. IS are used for response normalization and connection lines help to visualize which of the included IS will be adopted to normalize each ^{2}D peak/blob response.

In **Figure 1B**, filled circles indicate positive matches between template peak and the actual pattern while empty circles are for template peaks for which the correspondence was not

verified. False negative matches can be limited by appropriate selection of threshold parameters, reference spectra and constraint functions[13 , 14 , 18 , 19]. For complex patterns with multiple co-elutions, ion peak detection functions that are based on spectral deconvolution are advisable and could be a valid option[19]. Template peak metadata are shown in the enlarged panel of **Figure 1B** for (*E*)-2-hexenal.

The specificity of template matching relies on the possibility to apply constraint functions that limit positive correspondence to those candidate peaks that, falling within the search window of the algorithm, have MS spectral similarity above a certain threshold. In this case, in step 11, similarity thresholds[23] were set at 700 according to previous experiments aimed at defining optimal parameters limiting false negative matches[14]. Highlighted areas of the template peak properties in **Figure 1B** show the information about the reference MS spectrum string and the qCLIC constraint function (i.e., (Match("<ms>") >= 700.0) and (RMatch("<ms>") >= 700.0)).

By applying the template to all chromatograms of a set, one could encounter challenging situations as in the case of partial misalignment of patterns. This can be due to oven temperature inconsistencies, carrier gas flow/pressure instabilities, or because of a manual intervention on the system as in the case of column substitution or modulator loop-capillary replacement[14 , 28]. **Figure 3** shows a situation of a partial misalignment between the targeted template and the actual chromatogram. For minimal misalignments, interactive template transforms (**Figure 3**, control panel) can reposition template peaks for a better fit. Once repositioned, the template can be matched to establish correspondences. In the example, the template (**Figure 3**, step 12) peaks correctly match with the actual ^{2}D pattern. In case of severe

misalignments, not discussed here, the repetition of match-transform-update actions can iteratively adapt the template peaks position to the actual peak pattern[12 , 13 , 14].

Here, the targeted peaks (i.e., known analytes) provide about 40% of the chromatographic result (196 targeted peaks of about 500 detectable peaks on average). The other 60% of compounds, together with the information they bring, are not taken into consideration in targeted analysis. To make the investigation truly comprehensive, consistent cross-alignment of untargeted ^{2}D peaks should also be established. The first application where template matching was extended to all detectable analytes dealt with the complex volatilome of roasted coffee[7]. This process is automated with a software (e.g., Investigator), shown here in steps 14–15.

In this process, pre-targeted images belonging to the sample set under study (20 samples) are used to define reliable peaks by cross-matching of all image patterns[29]. Subsequently, a composite chromatogram is built from which one can identify UT reliable peaks and peak regions (i.e., ^{2}D peaks footprint) in the so-called feature template[17].

For analyses acquired at 70 eV, the process determined 144 reliable peaks with relaxed reliability[29], 76 of which belong to the targeted peaks list. Based on these 144 reliable peaks, the process aligns all chromatograms consistently with the average retention times of the reliable peaks and then combines them to create a composite chromatogram. **Figure 4** shows a list of all samples labeled according to the production region of the oil (left) and the list of reliable peaks/blob volumes in each sample (right).

The untargeted feature template is composed of ^{2}D peaks from analytes detected in the composite chromatogram, shown in **Figure 5A**, that are matched by the reliable-peaks

template (n = 168 – red circles for targeted peaks and green circles for untargeted peaks). The mass spectra of the composite peaks, as well as their retention times, are recorded in the feature template as shown for (*Z*)-3-hexenol acetate in the enlarged area. Peak-regions are shown in **Figure 5B** as red colored graphics; they are instead defined by the outlines of all ^{2}D peaks detected in the composite chromatogram (n = 3578).

When unsupervised pattern recognition by Principal Component Analysis is applied to targeted peaks distribution within the 20 analyzed samples, Sicilian and Tuscany oils cluster separately suggesting that pedo-climatic conditions and terroir impact the relative prevalence of volatiles. The results are shown in **Figure 6A** and the PCA results from the reliable peaks distribution are shown in **Figure 6B**. The two approaches cross-validate that oils from different geographical areas have different, while coherent, chemical signatures whether targeted or untargeted compounds, or both, are mapped.

Finally, the software enables prompt and effective re-alignment of patterns across parallel detection channels. In this application, the re-alignment is proposed for tandem ionization signals. The ion source of the MS multiplexes between two ionization energies (i.e., 70 and 12 eV) at an acquisition frequency of 50 Hz per channel[30]. The two resulting chromatographic patterns are closely aligned while spectral data (i.e., spectral signatures and responses) bring complementary information with different dynamic ranges of response[26,27]. The aligned patterns allow extracting features (^{2}D peaks and peak-regions) with univocal IDs (i.e., chemical names for targeted peaks and unique numbering # for untargeted peaks and peak-regions).

Template matching allows effective cross-alignment. In this situation, there is not much misalignment, but MS constraints must be relaxed to allow matches for UT peaks. On the other hand, featured UT peak-regions, that have no MS constraints, are promptly matched without any false negative matches. **Figure 5C** shows an enlarged area of a 12 eV chromatogram where the feature template built from 70 eV data is matched. Reliable UT peaks are positively matched because of the lowered qCLIC constraints (e.g., DMF threshold at 600). To note, at 12 eV, there are fewer detected peaks due to the limited fragmentation induced by low ionization energy.
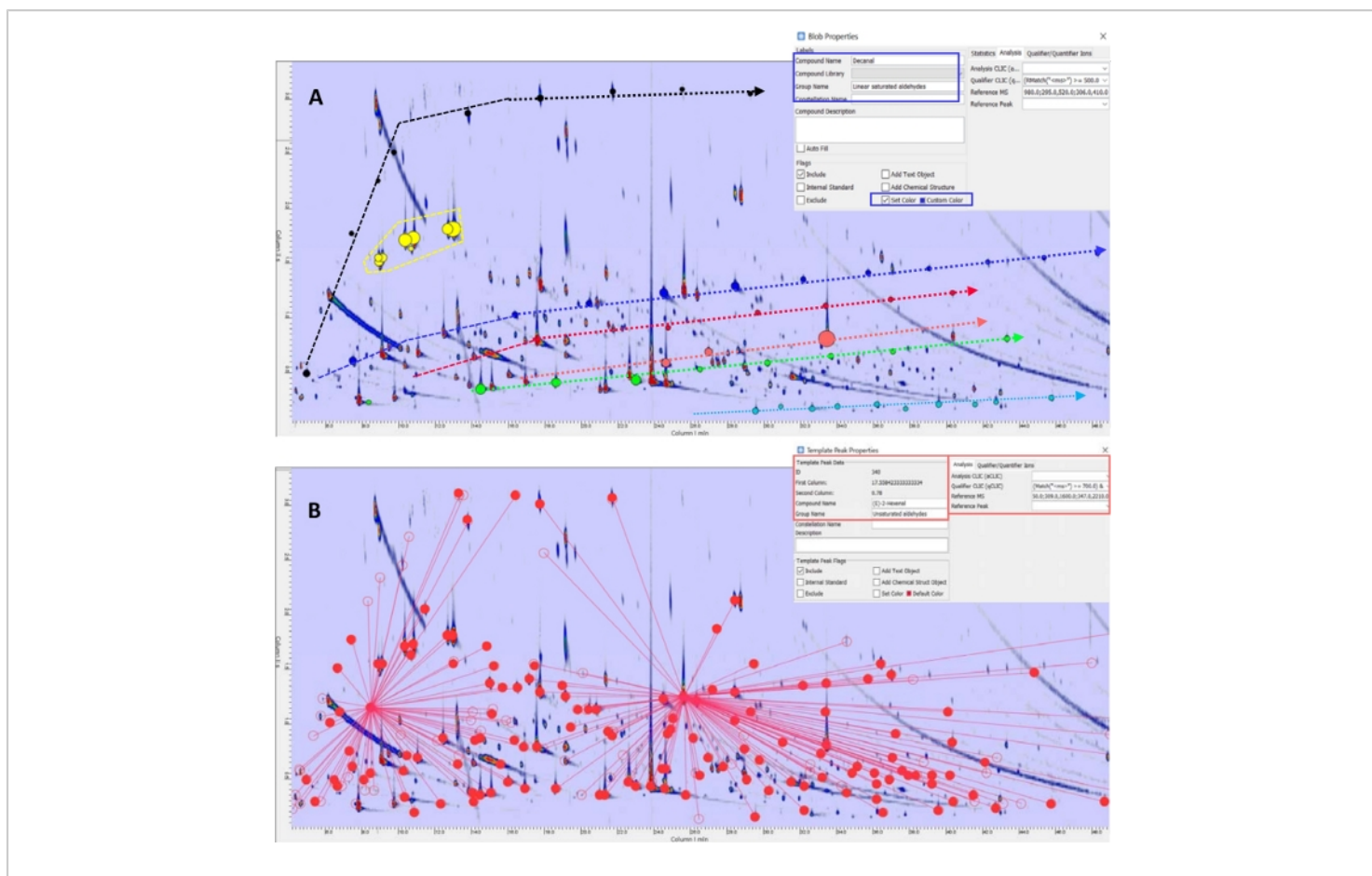
Figure 1: Bidimensional contour plot and targeted template. (**A**) Contour plot of the volatile fraction of an extra-virgin olive oil from Tuscany. Ordered patterns of homolog series and classes are highlighted with different colors and lines: linear saturated hydrocarbons (black line and ^{2}D contours) unsaturated hydrocarbons (yellow), linear saturated aldehydes (blue) mono-unsaturated aldehydes (red), polyunsaturated aldehydes (salmon), primary alcohols (green) and short-chain fatty acids (cyano). (**B**) Overimposed targeted template of known analytes (red colored circles) with connection lines linking Internal Standards (ISs). Panels show ^{2}D peak/blob properties metadata (Decanal) or Template peak properties. Please click here to view a larger version of this figure.
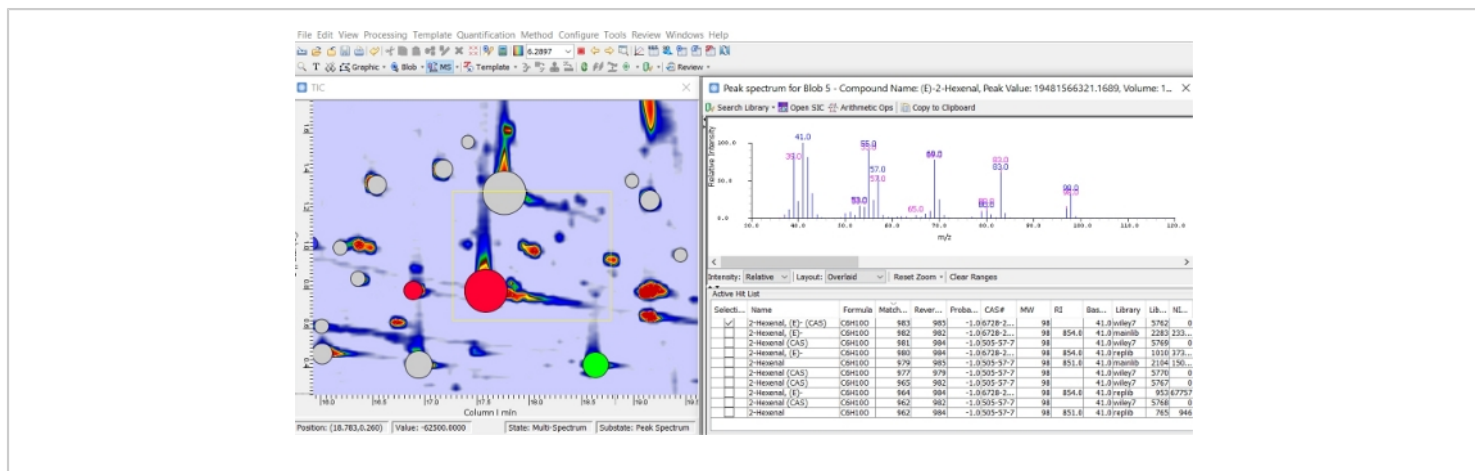
Figure 2: Apex MS search. Output of the apex MS search for blob 5. List of the database entries with the highest similarity match and related metadata available from the library. Please click here to view a larger version of this figure.
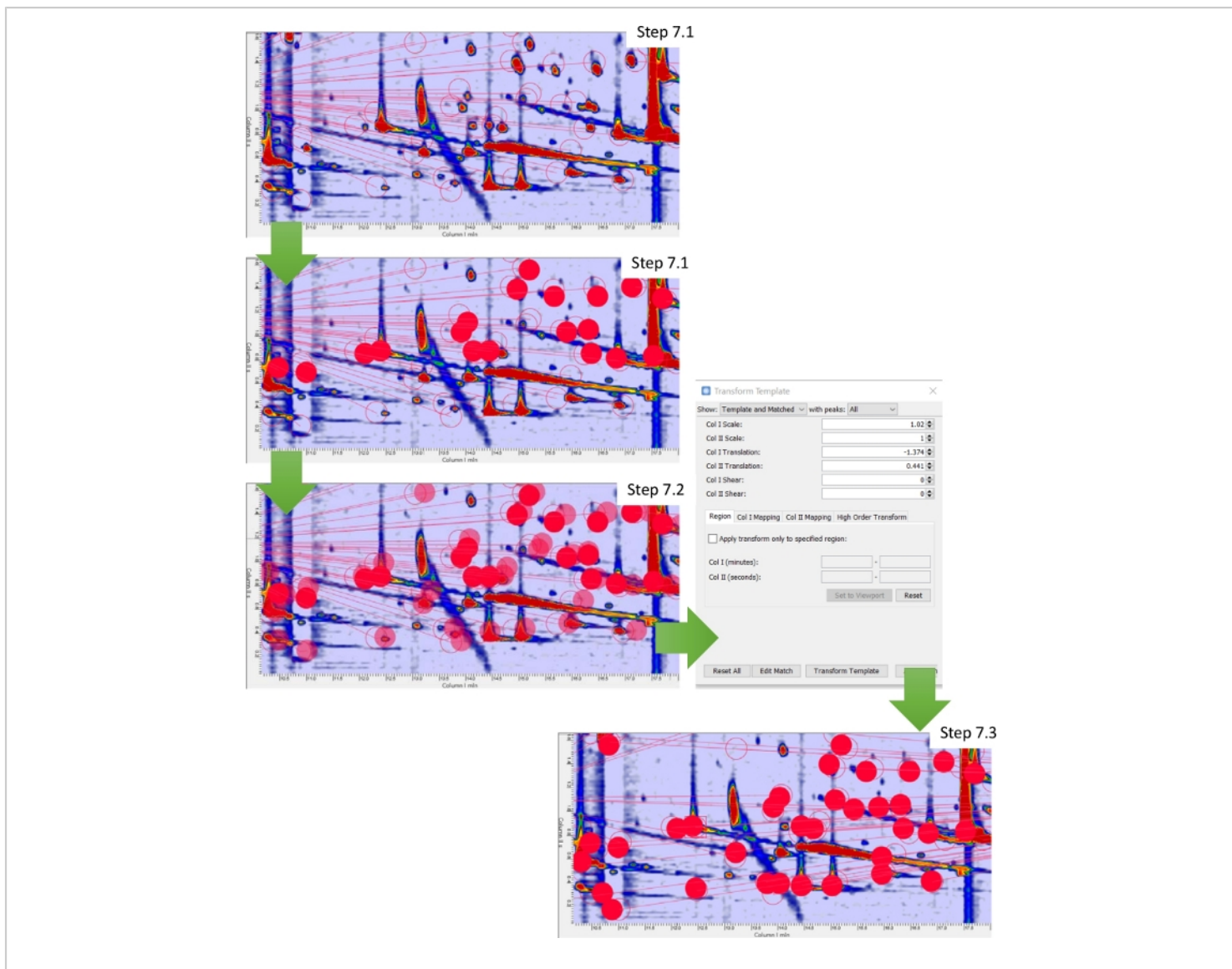
Figure 3: Template realignment. Workflow illustrating the steps that allow re-alignment of the template by transformation.

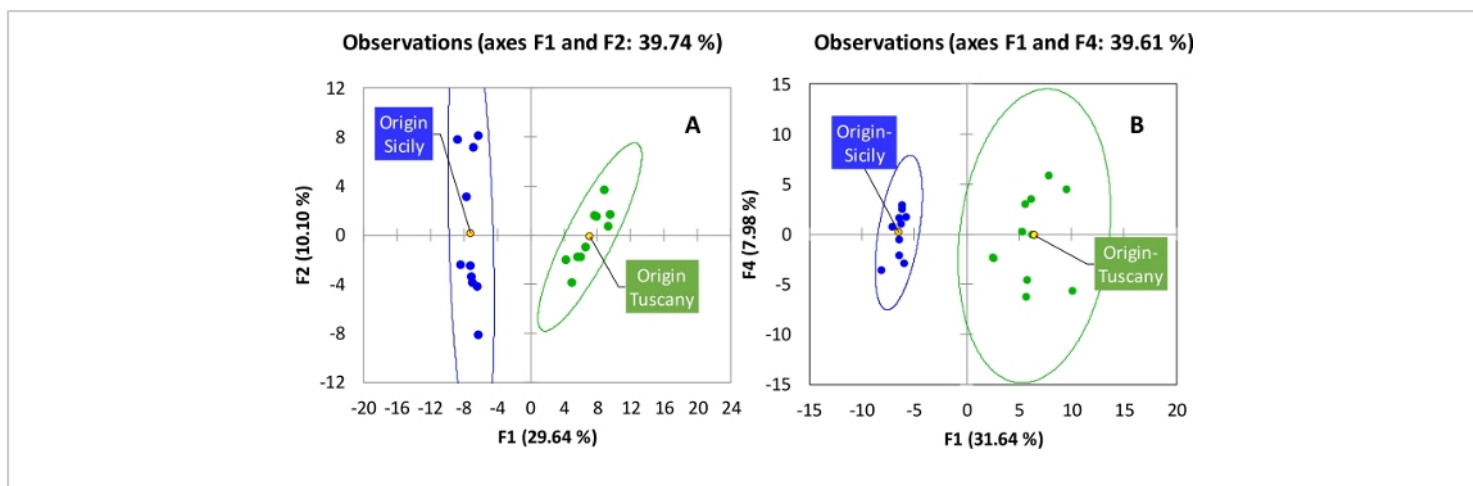Please click here to view a larger version of this figure.
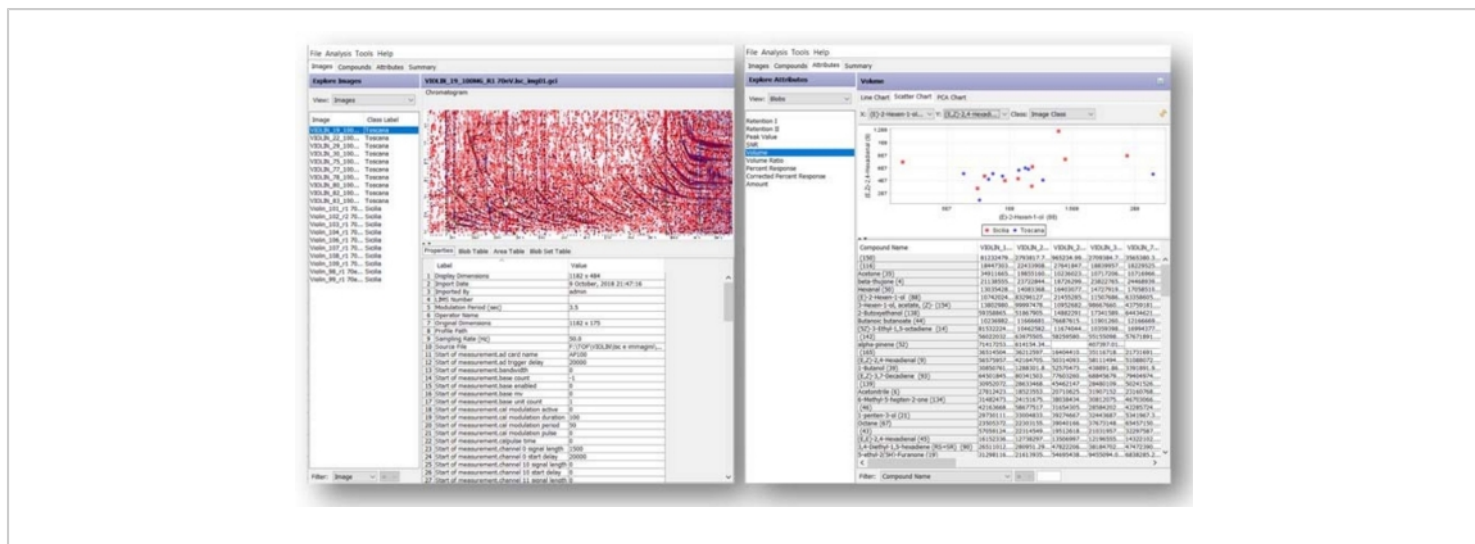
Figure 4: GC Investigator interface. Investigator panel with all selected images labeled according to the production Region of the oil (left) and the list of reliable peaks/blob volumes in each sample (right). Please click here to view a larger version of this figure.
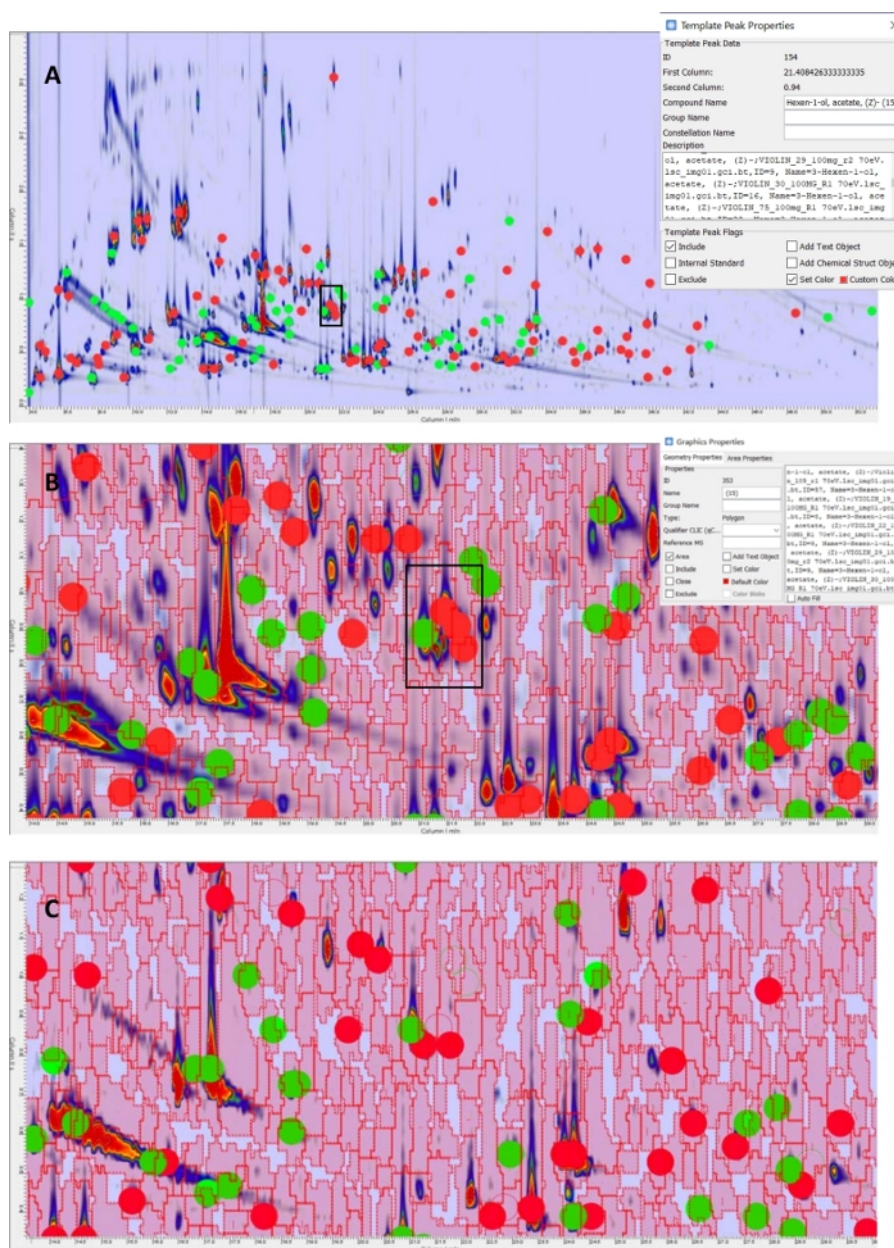
Figure 5: Targeted and UT template. (**A**) Reliable peaks as resulting from the automated processing in step 11; red circles correspond to known analytes while green circles are unknowns. In the superimposed panel, template object properties are shown for the (Z)-3-hexenal. (**B**) Enlarged area that shows the UT peaks (red and green circles) and peak-regions (red graphics) of the UT template matched on a sample oil acquired at 70 eV ionization energy. (**C**) UT template matched on a sample oil acquired at 12 eV ionization energy. Please click here to view a larger version of this figure.

Figure 6: PCA loading plots. They show the natural conformation of samples (oils from Tuscany and Sicily) as they result by (**A**) targeted peaks distribution or (**B**) UT peaks distribution. Please click here to view a larger version of this figure.

Supplemental Files. Please click here to download these files.

Discussion

Visualization of GC×GC-TOF MS data is a fundamental step for an appropriate understanding of the results achieved by comprehensive two-dimensional separations. Image plots with customized colorization allows analysts to appreciate detector response differences and thus the differential distribution of sample components. This visual approach completely changes the analysts' perspective on the interpretation and elaboration of chromatograms. This first step, once understood and confidently used by chromatographers, opens a new perspective in further processing.

Another fundamental aspect of data processing is the accessibility to the full data matrix (i.e., MS spectral data and responses) for all sample points, each of which corresponds to a single detector event. In this respect ^{2}D peaks integration, so that the collection of detector events

corresponding to a single analyte represent a critical step. In the current protocol, ^{2}D peaks detection is based on the watershed algorithm[18] with some adaptations included to improve detection sensitivity in case of partial co-eluting compounds. To make this process more specific, deconvolution must be done, and more sophisticated procedures adopted. This is possible by performing an ion peak detection for MS data; the algorithm processes the data array and isolates the response from single analytes based on spectral profiles[19,31].

An important yet critical step of the protocol, and of any GC×GC-MS data interpretation process, relates to analytes identification. This procedure, proposed in steps 8 and 9, in absence of a confirmatory analysis with authentic standards, must be carefully conducted by the analyst. Automated actions are available in any commercial software; they include MS spectral signature similarity evaluation against the collected reference spectra (i.e., spectral libraries) and evaluation of characteristic ratios among qualifier/quantifier ions. However, additional confirmatory criteria are needed to

disambiguate identification of isomers. The protocol proposes the adoption of linear retention indexes to prioritize the list of candidates; the limit here relates to the availability of retention data and its consistency.

The main characteristic that makes this approach unique is template matching[12, 13, 15, 29]. Template matching enables ^{2}D pattern recognition in a very effective, specific, and intuitive way. It can be set, in terms of sensitivity and specificity, by applying customized threshold values and/ or constraint functions while the analyst can supervise the procedure by actively interacting with transform function parameters. The peculiarity of this process relies on the possibility to cross-align targeted and untargeted peaks information between samples of a uniform batch but also between samples acquired with the same nominal conditions despite medium-to-severe misalignment. Advantages of this operation relate to the possibility to preserve all targeted analytes identifications, which is a time-consuming task for the analyst, and all metadata saved for targeted and untargeted peaks from previous elaboration sessions.

Template matching is also very effective in terms of computational time; low-resolution MS data files consists of about 1–2 Gb of packed data while high-resolution MS analyses may reach 10–15 Gb per single analytical run. Template matching does not process the full data matrix every time but, at first, performs retention-time alignment between chromatograms using template peaks then, processes candidate peaks within the search window for their similarity match with reference in the template. In case of severe misalignment, the most challenging situation, global second-order polynomial transforms performed better than local methods while reducing computational time[13].

For the GC×GC technique to spread widely beyond academia and research laboratories, data processing tools have to facilitate basic operations for visualization and chromatograms inspection; identification of analytes should offer the possibility to adopt standardized algorithms and procedures (e.g., NIST search algorithm and I^T calibration); and cross-comparative analysis should be intuitive, effective and supported by interactive tools. The proposed approach addresses these needs while offering advanced options and tools to deal with complex situations such as analytes co-elution, multiple analytes calibration, group-type analysis, and parallel detection alignment.

The referenced literature well covers many possible scenarios where GC×GC and, more generally, comprehensive two-dimensional chromatography, offer unique solutions and reliable results that cannot be achieved by ^{1}D-chromatography in single run analysis.[5, 32, 33] Although GC×GC is the most powerful tool that increases separation capacity and sensitivity, there are always limitations to separation power, sensitivity, and other systemic capacities. As these systemic limits are approached, data analysis becomes progressively more difficult. Therefore, research and development must continue to improve the analytical tools at our disposal.

Disclosures

Prof. Stephen E. Reichenbach and Dr. Qingping Tao have financial interests in GC Image, LLC. Dr. Daniela Peroni is an employee of SRA Instruments, a distributor of GC Image in Italy and France. Dr. Federico Stilo, Prof. Chiara Cordero, and Prof. Carlo Bicchi declare no conflicts of interest.

JOVE

Acknowledgments

References

1. Tranchida, P. Q. et al. Potential of comprehensive chromatography in food analysis. *Trends in Analytical Chemistry.* **52**, 186-205 (2013).

2. Cordero, C., Kiefl, J., Reichenbach, S. E., Bicchi, C. Characterization of odorant patterns by comprehensive two-dimensional gas chromatography: A challenge in omic studies. *Trends in Analytical Chemistry.* **113**, 364-378 (2019).

3. Cordero, C., Kiefl, J., Schieberle, P., Reichenbach, S. E., Bicchi, C. Comprehensive two-dimensional gas chromatography and food sensory properties: potential and challenges. *Analytical and Bioanalytical Chemistry.* **407**, 169-191 (2015).

4. Adahchour, M., Beens, J., Vreuls, R. J. J., Brinkman, U. A. T. Recent developments in comprehensive two-dimensional gas chromatography (GC × GC) - Introduction and instrumental set-up. *Trends in Analytical Chemistry.* **25** (5), 438-454 (2006).

5. Prebihalo, S. E. et al. Multidimensional gas chromatography: Advances in instrumentation, chemometrics, and applications. *Analytical Chemistry.* **90** (1), 505-532 (2018).

6. Cordero, C. et al. Profiling food volatiles by comprehensive two-dimensional gas chromatography coupled with mass spectrometry: advanced fingerprinting approaches for comparative analysis of the volatile fraction of roasted hazelnuts (*Corylus Avellana L.*) from different origins. *Journal of Chromatography A.* **1217**, 5848-5858 (2010).

7. Cordero, C. et al. Targeted and non-targeted approaches for complex natural sample profiling by GC×GC-QMS. *Journal of Chromatography Sciences.* **48** (4), 251-261 (2010).

8. Maio, D., Maltoni, D. Direct gray-scale minutiae detection in fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **19** (1), 27-40 (1997).

9. Jain, A. K., Hong, L., Pankanti, S., Bolle, R. An identity-authentication system using fingerprints. *Proceedings of IEEE.* **85** (9), 1365-1388 (1997).

10. Parsons, B. A. et al. Tile-based fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC-TOFMS) data using a null distribution approach. *Analytical Chemistry.* **87** (7), 3812-3819 (2015).

11. Pierce, K. M., Kehimkar, B., Marney, L. C., Hoggard, J. C., Synovec, R. E. Review of chemometric analysis techniques for comprehensive two dimensional separations data. *Journal of Chromatography A.* **1255**, 3-11 (2012).

12. Reichenbach, S. E. et al. Alignment for comprehensive two-dimensional gas chromatography with dual secondary columns and detectors. *Analytical Chemistry.* **87** (19), 10056-10063 (2015).

13. Rempe, D. W. et al. Effectiveness of global, low-degree polynomial transformations for GCxGC data alignment. *Analytical Chemistry.* **88** (20), 10028-10035 (2016).

14. Stilo, F. et al. Untargeted and targeted fingerprinting of extra virgin olive oil volatiles by comprehensive two-dimensional gas chromatography with mass spectrometry: challenges in long-term studies. *Journal of Agricultural and Food Chemistry.* **67** (18), 5289-5302 (2019).

15. Reichenbach, S. E., Carr, P. W., Stoll, D. R., Tao, Q. Smart templates for peak pattern matching with comprehensive two-dimensional liquid chromatography. *Journal of Chromatography A.* **1216** (16), 3458-3466 (2009).

16. Reichenbach, S. E. et al. Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGC-HRMS). *Talanta.* **83** (4), 1279-1288 (2011).

17. Reichenbach, S. E., Tian, X., Cordero, C., Tao, Q. Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography. *Journal of Chromatography A.* **1226,** 140-148 (2012).

18. Latha, I., Reichenbach, S. E., Tao, Q. Comparative analysis of peak-detection techniques for comprehensive two-dimensional chromatography. *Journal of Chromatography A.* **1218** (38), 6792-6798 (2011).

19. Reichenbach, S. E., Tao, Q., Cordero, C., Bicchi, C. A data-challenge case study of analyte detection and identification with comprehensive two-dimensional gas chromatography with mass spectrometry (GC×GC-MS). *Separations.* **6** (3), 38 (2019).

20. Reichenbach, S. E. Chapter 4 Data Acquisition, Visualization, and Analysis. *Comprehensive Analytical Chemistry.* 77-106 (2009).

21. Reichenbach, S. E., Ni, M., Zhang, D., Ledford, E. B. Image background removal in comprehensive two-dimensional gas chromatography. *Journal of Chromatography A.* **985** (1-2), 47-56 (2003).

22. van Den Dool, H., Dec. Kratz, P. A Generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography. *Journal of Chromatography A.* **11**, 463-471 (1963).

23. NIST Mass Spectrometry Data Center. NIST Standard Reference Database 1A: NIST/EPA/NIH Mass Spectral Library (NIST 08) and NIST Mass Spectral Search Program (Version 2.0f). *National Institute of Standards and Technology (NIST).* Gaithersburg MD (2005).

24. Magagna, F. et al. Combined untargeted and targeted fingerprinting with comprehensive two-dimensional chromatography for volatiles and ripening indicators in olive oil. *Analytica Chimica Acta.* **936**, 245-258 (2016).

25. Reichenbach, S. E. et al. Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition. *Journal of Chromatography A.* **1595**, 158-167 (2019).

26. Cialiè Rosso, M. et al. Adding extra-dimensions to hazelnuts primary metabolome fingerprinting by comprehensive two-dimensional gas chromatography combined with time-of-flight mass spectrometry featuring tandem ionization: insights on the aroma potential. *Journal of Chromatography A.* **1614** (460739), 1-11 (2020).

27. Cordero, C. et al. Comprehensive two-dimensional gas chromatography coupled with time of flight mass spectrometry featuring tandem ionization: challenges and opportunities for accurate fingerprinting studies. *Journal of Chromatography A.* **1597**, 132-141 (2019).

28. Ni, M., Reichenbach, S. E., Visvanathan, A., TerMaat, J., Ledford, E. B. Peak pattern variations related to comprehensive two-dimensional gas chromatography acquisition. *Journal of Chromatography A.* **1086**, 165-170 (2005).

29. Reichenbach, S. E. et al. Reliable peak selection for multisample analysis with comprehensive two-dimensional chromatography. *Analytical Chemistry.* **85** (10), 4974-4981 (2013).

30. Markes International. Select-EV: The next Generation of Ion Source Technology. *Technical Note.* (2016).

31. Tao, Q., Reichenbach, S. E., Heble, C., Wu, Z. New investigator tools for finding unique and common components in multiple samples with comprehensive two-dimensional chromatography. *Chromatography Today.* February-March (2018).

32. Seeley, J. V., Seeley, S. K. Multidimensional gas chromatography: fundamental advances and new applications. *Analytical Chemistry.* **85** (2), 557-578 (2013).

33. Tranchida, P. Q., Aloisi, I., Giocastro, B., Mondello, L. Current state of comprehensive two-dimensional gas chromatography-mass spectrometry with focus on processes of ionization. *Trends in Analytical Chemistry.* **105**, 360-366 (2018).