Journal of Chromatography A, 1226 (2012) 140–148

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Review

Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography

Stephen E. Reichenbach[a,*], Xue Tian[a], Chiara Cordero[b], Qingping Tao[c]

[a] University of Nebraska – Lincoln, Computer Science and Engineering Department, Lincoln, NE 68588-0115, USA
[b] Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via P. Giuria 9, I-10125 Torino, Italy
[c] GC Image, LLC, PO Box 57403, Lincoln, NE 68505-7403, USA

ARTICLE INFO

Article history:
Available online 26 July 2011

Keywords:
Comprehensive two-dimensional gas chromatography (GC × GC)
Comprehensive two-dimensional liquid chromatography (LC × LC)
Non-targeted analysis
Cross-sample analysis
Feature generation and matching
Pattern recognition

ABSTRACT

This review surveys different approaches for generating features from comprehensive two-dimensional chromatography for non-targeted cross-sample analysis. The goal of non-targeted cross-sample analysis is to discover relevant chemical characteristics (such as compositional similarities or differences) from multiple samples. In *non-targeted analysis*, the relevant characteristics are unknown, so individual features for all chemical constituents should be analyzed, not just those for targeted or selected analytes. *Cross-sample analysis* requires matching the corresponding features that characterize each constituent across multiple samples so that relevant characteristics or patterns can be recognized. Non-targeted, cross-sample analysis requires generating and matching all features across all samples. Applications of non-targeted cross-sample analysis include sample classification, chemical fingerprinting, monitoring, sample clustering, and chemical marker discovery. Comprehensive two-dimensional chromatography is a powerful technology for separating complex samples and so is well suited for non-targeted cross-sample analysis. However, two-dimensional chromatographic data is typically large and complex, so the computational tasks of extracting and matching features for pattern recognition are challenging. This review examines five general approaches that researchers have applied to these difficult problems: visual image comparisons, datapoint feature analysis, peak feature analysis, region feature analysis, and peak-region feature analysis.

Contents

1. Introduction

The goal of non-targeted cross-sample analysis is to discover relevant chemical characteristics (such as compositional similarities or differences) from multiple samples. Some applications of non-targeted cross-sample analysis are:

- **Classification**. Given a sample from an unknown class and exemplary samples from a set of known classes, determine the class of the unknown sample. For example, given samples of cancerous tumors labeled by grade, determine the tumor grade for an ungraded sample [1].
- **Chemical fingerprinting**. Given a sample from an unknown source and exemplary samples from multiple known sources,

* Corresponding author. Tel.: +1 402 472 5007; fax: +1 402 472 7767.
E-mail addresses: reich@cse.unl.edu (S.E. Reichenbach), xtian@cse.unl.edu (X. Tian), chiara.cordero@unito.it (C. Cordero), qtao@gcimage.com (Q. Tao).
URL: http://www.gcimage.com (Q. Tao).

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.07.046

determine the source of the unknown sample. For example, given a sample of environmental pollution from an unknown source and labeled samples from several possible sources of the pollution, identify the source for the pollution [2]. Fingerprinting is a type of classification problem except that each class is restricted to a single source, whereas the general classification problem allows each class to have multiple similar sources.
- **Monitoring**. Given a sequence of samples, identify samples that have uncharacteristic differences with other samples, e.g., for quality assurance. Monitoring also can be used to discover trends in sample sequences, even recognizing subtle changes if they are progressive or cyclical. For example, use a time-sequence of samples from an environmental oil spill to track and understand the weathering processes on oil constituents [3].
- **Clustering**. Given a set of samples, partition subsets such that samples within each subset are relatively similar and samples in different subsets are relatively dissimilar. For example, given multiple samples from oil reservoirs, use clustering to determine the number of distinct reservoirs [4].
- **Marker discovery**. Given a set of exemplary samples from known classes, determine the chemical characteristics that are most relevant for distinguishing the classes. For example, given samples of tumors labeled by grade, determine which characteristics (i.e., biomarkers) are most useful in distinguishing different tumor grades [1].

Non-targeted cross-sample analysis should evaluate each and every constituent in each and every sample. For *non-targeted analysis*, the relevant chemical characteristics are not known, so the analysis should generate characteristic *feature(s)* for each and every constituent. Typically, detector intensities or mass spectral (total and/or selected ion) intensities are used as characteristic features because they indicate the analyte concentrations (or amounts) and provide information for chemical identification. *Cross-sample analysis* should compare the same chemical characteristics across multiple samples, so it is necessary to correctly match the corresponding features that characterize the same analyte in different samples. For example, peak matching would establish which peaks in different samples result from the same analyte. Typically, other features, such as retention times and/or mass spectral signatures, are used to match the characteristic features.

Non-targeted cross-sample analysis requires comprehensive, selective, matched, accurate features. If the features are not comprehensive, then relevant characteristics may not be analyzed. If the features are not selective, then relevant trace constituents may be obscured by more prevalent but less relevant constituents. If the features are not matched, then the analysis is confounded by incorrect comparisons. If the features are not accurate, then the analysis may be unable to detect subtle differences.

Comprehensive two-dimensional gas chromatography (GC × GC) and related techniques are well-suited for non-targeted cross-sample analysis because they offer increased separation capacity, higher dimensional structure–retention relationships, and improved signal-to-noise ratio (SNR), compared to traditional one-dimensional chromatography. Comprehensive two-dimensional chromatography preserves separations at each stage and submits the entire sample to analysis, providing for comprehensive features. Increased separation capacity enables more selective features. The higher dimensional structure relationships can be exploited for better matched features. And, the improved SNR increases the quantitative accuracy of characteristic features.

Comprehensive two-dimensional chromatography offers unprecedented information on compositional characteristics of complex samples, but the size and complexity of the data makes data analysis to extract that information a challenging problem. The most relevant features for a particular cross-sample analysis may be related to trace constituents and/or unidentified compounds. Relevant patterns may involve subtle relationships among multiple features. So, the goal of non-targeted cross-sample analysis is to extract and analyze all of the information that could be relevant. In some sense, it is the ultimate information processing challenge.

The typical data processing sequence for non-targeted cross-sample analysis is:

1. Preprocess individual chromatograms.
2. Generate features for each chromatogram.
3. Match features across chromatograms.
4. Recognize relevant patterns.

The purpose of this review is to examine various approaches that researchers have applied to Steps 2 and 3 — feature generation and matching — but Steps 1 and 4 merit a brief discussion. Preprocessing (Step 1) involves operations (e.g., baseline correction [5–8], peak detection [9–13], coeluted peak detection [14–25], and alignment [24,26–36]) that prepare data for further analysis, but which are not specific to non-target cross-sample analysis. Therefore, general preprocessing methods can be used for these operations. In pattern recognition (Step 4), the matched comparative features are analyzed to recognize relevant characteristics or patterns among samples. Such pattern recognition is not specific to chromatographic analysis and so can be performed with various general-purpose methods, including statistical methods such as principal component analysis (PCA), analysis of variance (ANOVA), and discriminant function analysis (DFA), and machine-learning methods such as support vector machines (SVM), neural networks, and decision trees [1,4,31,35–58]. Of course, research continues to improve methods for preprocessing and pattern recognition and to evaluate their effectiveness for non-targeted cross-sample chromatographic analysis, but that research is not the focus of this review.

This review describes five different types of features that have been used for non-targeted cross-sample analyses with comprehensive two-dimensional chromatography: visual images, datapoints, peaks, regions, and peak-regions. Visual images present chromatograms using various methods for two-dimensional data, including pseudo-colorization, contour plots, and three-dimensional projections. Datapoint analyses treat each datapoint as a feature, allowing chromatograms to be compared intensity by intensity. Peak-based approaches attempt to separately integrate the intensities from multiple datapoints that are induced by each individual analyte. Regional features aggregate datapoints in separate regions of the two-dimensional chromatographic plane. Peak-region methods attempt to define a region for each individual analyte.

Some examples of previous research illustrate each approach to generating and matching features for two-dimensional chromatographic analyses, with most research involving GC × GC. The order of presentation roughly follows the historical development. The discussion of each approach presents advantages and problematic issues. Other authors have provided more general reviews of GC × GC and related technologies and provide a broader context for this review [59–77].

2. Visual features

The earliest non-targeted cross-sample analyses with comprehensive two-dimensional chromatography were conducted without benefit of software specifically designed for operating on two-dimensional chromatographic data. Therefore, most early cross-sample comparisons were primarily qualitative visual comparisons using general-purpose software. In particular, two-dimensional chromatograms can be regarded as digital images of the chromatographic plane. Digital images are two-dimensional arrays of intensities and the datapoint intensities of two-dimensional chromatograps are represented naturally in two-dimensional arrays arranged so that the abscissa (X-axis, left-to-right) is the elapsed time for the first-column separation and the ordinate (Y-axis, bottom-to-top) is the elapsed time for the second-column separation. Then, digital image visualization and processing methods can be used for two-dimensional chromatograms.

In 1990, Bushey and Jorgenson [78] demonstrated comprehensive two-dimensional liquid chromatography LC × LC and showed data from a UV detector as surface plots with three-dimensional projection to two dimensions. They presented side-by-side visualizations of reconstituted serum from a human and from a horse, but did not make explicit comparisons of the samples.

Blomberg et al. [79] showed side-by-side two-dimensional contour plots of GC × GC data from a flame ionization detector (FID) for distillation fractions of a heavy catalytic cracked cycle oil before and after hydrogenation to illustrate the conversion of olefins and sulfur compounds. Their results showed that “a clear distinction between different products is visible immediately” [79, p. 544]. For perspective on the computers of the time, they used a computer with 100 MHz processor, 32 megabytes of memory, and generic scientific data processing and visualization software. The authors noted the need for more automated processing to characterize and compare samples: “The vast amount of data generated, necessitate that considerable effort has to be put in software and hardware developments for automated interpretation” [79, p. 544].

Gaines et al. [2] presented GC × GC–FID data from an oil spill sample and from two potential sources for the spill as pseudo-colorized images with a cold-to-hot color scale for qualitative visual comparison. Their goal was to demonstrate GC × GC for oil spill source identification, an application of fingerprinting. The visual comparison allowed them to note that one of the sources exhibited considerably fewer peaks in the heavy aromatic region than the spill, which suggested that it was not the source for the spill. They also made selected quantitative comparisons for fingerprinting, as described here in subsequent sections.

Reddy et al. [80] used a side-by-side sequence of pseudo-colorized images to visualize GC × GC–FID data from progressively weathered samples of a fuel oil standard for comparison to an image of data from a sample of a decades-old fuel oil spill. Their goal was to understand progressive changes in the oil. The visual comparisons allowed them to observe that 70% evaporative weathering of the standard was required to effect the same level of reduction of naphthalenic compounds observed in the oil spill sample, but that level of weathering also removed other components that still were present in the oil spill sample. They were able to conclude that evaporative weathering could not solely account for the GC × GC pattern observed in the oil-spill sample and that other factors, such as water washing, preferential biodegradation, and microbial degradation were required to explain the actual weathering of the oil spill.

Others have used visual comparisons for similar purposes. Janssen et al. [81] visualized LC × GC–FID data for samples of edible oils and fats as two-dimensional bubble plots with circles indicating detected peaks (with dot locations determined by retention from LC and carbon number from GC and dot areas determined by intensity). Perera et al. [82] showed a region of GC × GC–FID data as contour plots to fingerprint headspace volatiles from plant samples. Hope et al. [83] used contour plots to compare total intensity counts (TICs) of data from GC × GC with time-of flight (TOF) mass spectrometry (MS) for pre and post harvest lawn grass extracts. Shellie et al. [39] used GC × GC–TOFMS to analyze mouse spleen samples, then (a) visually compared averaged chromatograms from obese mice to averaged chromatograms from control mice, (b) computed the difference between the averaged chromatograms and showed images of the positive and negative values, (c) compared bubble plots for averaged peaks, and (d) compared bubble plots for relative weighted differences of averaged peaks (dividing by the average standard deviation among sample groups).

Hollingsworth et al. [32] developed software methods for automatically aligning chromatograms using reference peaks, normalizing intensities, and visualizing the differences by various image-based methods, including time-loop flicker (switching between images) and colorized differences. Fig. 1 illustrates a small chromatographic region with benzene, toluene, ethylbenzene, and xylene (BTEX) peaks and a visualization of the differences between two aligned chromatograms. Nelson et al. [84] and Wardlaw et al. [85] used these methods to illustrate weathering of an oil spill and oil seep. Cordero et al. [51] used these methods to compare chromatograms from coffee samples. Such visualizations of pointwise differences provide a segue to the next approach for non-targeted multi-sample analyses — pointwise feature analysis.

Visual comparisons continue to be used both as a preliminary tool and as an investigatory and confirmatory method for automated methods. However, visual analyses are insufficient in several respects: the approach is not quantitative, subtle differences and complex patterns may not be visible, and the approach is not well suited for cross-sample analysis with large sample sets.

3. Datapoint features

Quantitative pointwise comparison is a natural progression from visual image comparison. In a pointwise approach, chromatograms are compared point-by-point (or in imaging terms pixel-by-pixel). With this approach, each datapoint is a feature and the datapoint features at the same retention times are implicitly matched.

In 2002, Johnson and Synovec [37] used quantitative datapoint features (i.e., the chromatographic intensities at each datapoint) of GC × GC–FID data to recognize patterns in different jet fuel mixtures. Their first experiments involved five replicates for each of nine different mixtures of two fuels for a total of 45 chromatograms each with 120 K datapoints. Their second experiments involved three replicates for each of 13 different classes for a total of 39 chromatograms each with 105 K datapoints. The potential relevance of each feature was computed by ANOVA, as the Fisher f ratio — the variance between classes divided by the variance within classes. Then, features were selected based on a f-ratio threshold that yielded good class separation in the space defined by the first two components of PCA. In this way, they reduced the number of features to a few hundred, which gave good PCA separation of classes and good organization in a K-means dendrogram.

Mohler et al. [40] and Pierce et al. [41] applied PCA to GC × GC–TOFMS datapoint intensities at selected mass-to-charge (m/z) channels to show class separations for yeast [40] and plant [41] samples. Pierce et al. [42] analyzed organic acid metabolites in urine samples with GC × GC–TOFMS by computing the f ratios at every mass-to-charge (m/z) channel of each chromatographic datapoint and then summing the f ratios along the m/z dimension (i.e., for each datapoint). Then, they selected peaks with features (i.e., datapoints) having the largest weighted and unweighted f-ratio sums. For peaks indicated by the f-ratio sums, the ratios of the peak volumes between samples from non-pregnant women to samples from pregnant women indicated that those components significantly differentiated between the two classes.

Guo and Lidstrom [46] applied the same approach with GC × GC–TOFMS data to investigate differences in metabolite

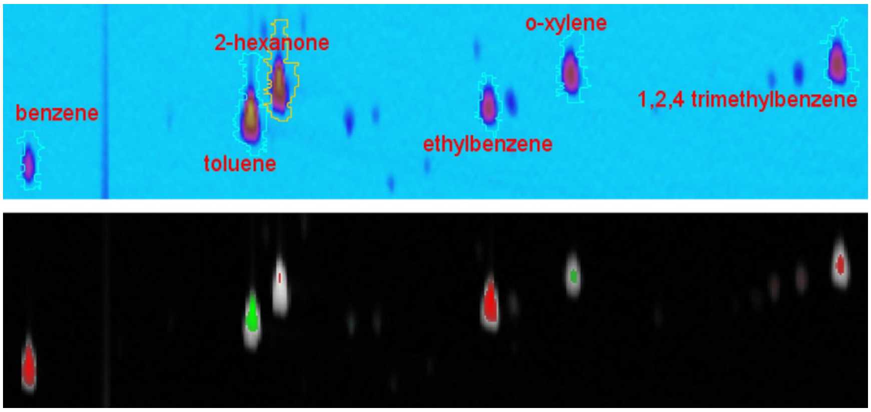

Fig. 1. Top — a pseudocolorized image of a chromatographic region with BTEX peaks. Bottom — a pseudocolorized image of the differences between two aligned chromatograms with red indicating a larger value in the reference image, green indicating a smaller value, and grey indicating nearly equal values [32]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

profiles of methylotrophic bacteria. Mohler et al. [43] used the same approach to GC × GC–TOFMS data for yeast metabolites and then performed the Student's *t*-test as a check on the volumes of the peaks indicated by the summed *f* ratios. Subsequently, Mohler et al. [47] used the ratios of the largest and smallest signals in GC × GC–TOFMS data to distinguish datapoints and then peaks that changed in concert with the dissolved oxygen cycle of yeast. Vial et al. [35,58] used dynamic peak alignment followed by PCA for GC × GC-MS data for several tobacco extracts and later used correlation with class members to assess the discriminatory power of each datapoint to analyze a large set of GC × GC–MS chromatograms for tobacco extracts in three different classes. Gröger et al. [45] used multidimensional scaling, hierarchical clustering, and PCA on datapoint intensities to perform clustering and Fisher criterion to identify discriminating datapoints for illicit drug samples. Gröger and Zimmermann [36] used *t*-tests to select significant datapoint features from selected channels of GC × GC–TOFMS data for partial least-squares (PLS) discriminant analysis (DA). Ventura et al. [57] recently used multiway PCA on GC × GC–FID data for maltene fractions of crude oils.

Hollingsworth et al. [32], Mohler et al. [40,47], Almstetter et al. [34], Gröger and Zimmermann [36], and others have noted the importance of data alignment for datapoint feature analysis. Hollingsworth et al. [32], Almstetter et al. [34], and others have developed alignment algorithms. Gröger and Zimmermann [36] implemented alignment and other preprocessing operations with parallel processing. The scope of this review does not include alignment algorithms.

Chromatographic misalignment and peak shape variations pose serious problems for pointwise cross-sample analysis. The features are individual datapoints, so if there is any misalignment between any pairs of samples, even as small as a fraction of a datapoint interval, then the features are incorrectly matched. Misalignments, both global and local, naturally occur even in well controlled conditions. Analytes normally elute over multiple datapoints, so the effects of small misalignments are mitigated, but misalignment is a fundamental issue that is difficult to eliminate. Like differences due to alignment, peak-shape differences are erroneously seen as quantitative differences in datapoint features. Another issue is that pointwise analysis involves many features and many of those features are highly redundant. Both the number of features and feature redundancy complicate pattern recognition. In view of these issues, it can be argued that datapoint features may be too selective, thereby generating numerous features for slightly varying retention times within individual chromatographic peaks.

4. Peak features

Peak features aggregate multiple datapoints with the goal of characterizing individual analytes (e.g., summing all datapoint intensities that are attributed to each detected peak). Peak features characterize larger, more meaningful chromatographic structures, resulting in fewer features that are less redundant than datapoint features. Peak features also are less sensitive to misalignment and peak-shape variations than datapoint features because peaks typically span many datapoints. However, unlike datapoint features, peak features are not implicitly matched. So, after preprocessing and peak detection, the detected peaks in each chromatogram that are induced by same analyte must be matched. Feature matching is a critical challenge for peak-feature analysis.

Gaines et al. [2] provided an early demonstration of using quantitative characterizations of individual peaks and groups of peaks (i.e., the aggregation of several detected peaks) in GC × GC–FID data to fingerprint samples of an oil spill and potential sources in order to identify the source of the spill. Their analysis used summed intensities of four peaks and nine peak groups that were selected because of their suitability for source determination, so the analysis was not comprehensive, but was quite advanced given the lack of software for two-dimensional chromatography at the time. Also, the selections were performed by hand and so were not automated. Bar charts with the intensities of the selected features showed that one potential source was compositionally more similar to the spill than the other was.

Mispelaar et al. [38,4] used a much larger number of peaks to distinguish samples from different oil reservoirs with GC × GC–FID. Their peak detection found about 6000 peaks per chromatogram. They used retention-time based alignment and filtering to match 3904 peaks, but the results of their multi-variate analysis (MVA) were unsatisfactory. They attributed the poor initial results to an inadequate number of samples with many non-informative peaks and peak detection, quantification, and matching errors. They then selected 292 peaks using an automated criterion for the relative standard deviations (RSDs) between duplicate samples to indicate peak detection and quantification errors. Most of the automatically selected 292 peaks were in regions of the chromatogram with lower peak density. Then, they manually selected 65 peaks for relevance and absence of interference. This small fraction of the peaks (about 1% of the detected peaks) was adequate for clustering the samples according to reservoir, but the feature reduction is indicative of the difficulties of reliable peak detection and matching. Such selective processing could exclude highly informative peaks.

In their work with mouse spleen samples, Shellie et al. [39] matched peaks in each chromatogram to reference data using tolerances on retention times and mass spectral matching similarity. The TIC of each peak that matched the same reference peak was placed on the same row in a matrix with a column each chromatogram. They did not report how many peaks were detected or how many of the detected peaks were matched. Student's *t*-tests were used to indicate the eleven metabolites exhibiting the most significant differences between obese and control mice.

Qiu et al. [44] performed GC × GC–FID on volatile oils from Qianghuo, a traditional Chinese medicine, from five regions. They did not report parameters for rejecting peaks with low SNR nor the number of peaks detected. They developed and implemented peak alignment and matching methods (using retention times relative to reference peaks) to create a matrix with 1544 peaks in fifteen samples. PCA analysis produced three clusters, with separate clusters for samples from two of the five regions. They used variable importance in the projection (VIP) [86, p. 397] to identify potential marker compounds, finding some statistically significant features, then used GC × GC–TOFMS for chemical identification of those compounds.

Wardlaw et al. [85] developed an algorithm to track peaks between similar samples based on retention times. The algorithm tracked about 1400 of about 4500 peaks in GC × GC chromatograms from oil samples from the reservoir, sea floor, and sea surface.

Analyzing human serum with GC × GC–TOFMS, Oh et al. [87] developed a peak sorting method to recognize peaks from the same metabolite in different chromatograms. Their algorithm used several search criteria with retention times and mass spectra, with options to eliminate non-target peaks. Peaks with low signal-to-noise ratio were discarded during peak detection. The matched peaks showed high correlation for retention times and mass spectra, but only 105 peaks were matched across all fifteen chromatograms, even with five replicates for each of three samples.

Gaquerel et al. [48] used GC × GC–TOFMS to analyze the effect of oral secretions on volatile plant emissions. Peak detection yielded about 600 peaks in each of the 108 samples (subject to a threshold SNR of 10). The authors noted that inconsistencies in the numbers of the detected peaks in each chromatogram complicated matching. In each of three sample periods, the peak set of the chromatogram with the largest number of detected peaks was used as reference data for matching (with the matching procedure developed by Shellie et al. [39]), reducing the number of matched peaks to about 400, which then were corrected for false positives from the alignment and matching procedure. ANOVA followed by another manual check for false positives from the peak alignment and matching was used to select about 15% the peaks for MVA with hierarchical clustering analysis (HCA) and PCA.

Li et al. [49] analyzed blood plasma with GC × GC–TOFMS. They used a mass spectral filter to extract peaks for trimethylsilyated metabolites, then applied a peak alignment method and a peak matching algorithm to create a matrix with 492 metabolites in 79 chromatograms. They tried several modeling methods, including PLS-DA, in which some problems that were attributed to missing values from peak matching were resolved by additional peak filtering. Then, VIP was used to indicate potential biomarkers.

Reichenbach et al. [88] developed Smart Templates^{TM} for peak matching. The template records a prototypical pattern of peaks with retention times and associated metadata, such as chemical identities and compound-group membership. Then, the template pattern is matched to the detected peaks in subsequent chromatograms and the metadata are copied from the template to identify the matched peaks. The matching process explores the space of affine geometric transforms to maximize the number of matched peaks and minimize the residual geometric error. Smart templates employ rule-based constraints (e.g., multispectral matching) to increase matching accuracy. Smart templates also carry other structures, such as text and chemical-structure annotations and polygonal regions (which can be used for region features, described below). They demonstrated the approach and associated methods on urine samples analyzed by LC × LC with a ultraviolet (UV) diode array detector (DAD). Fig. 2 illustrates template peak matching with a template derived from the detected peaks of one chromatogram matched to the detected peaks of another chromatogram.

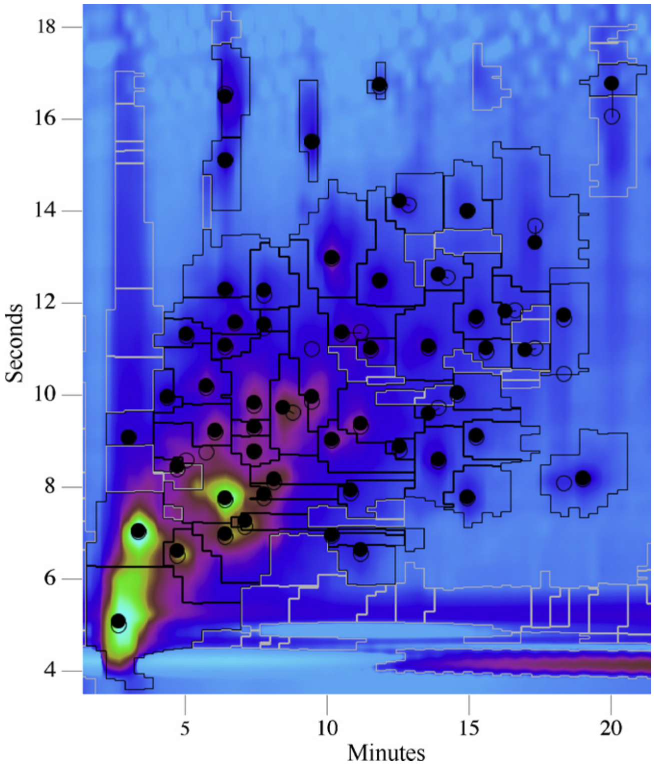

Fig. 2. A pseudocolorized image of an LC × LC chromatogram of a urine sample. The open circles indicate the retention times of the expected peaks recorded in the template. The outlines indicate the detected peaks and the filled circles indicate the retention times of the apexes of the detected peaks that are matched by the template [88].

Cordero et al. [89] analyzed volatile fractions of roasted hazelnuts with GC × GC–MS, then performed peak matching with templates in two different ways. In the first approach, they aligned and summed the chromatograms, then created a feature template comprised by the 411 peaks detected in the cumulative chromatogram. That template then was matched to each individual chromatogram, with matching rates ranging from 68% to 79%. In the second approach, they performed a sequential template matching that used both retention-time patterns and mass spectral matching criteria. At each step of the sequence, unmatched peaks were added to build a comprehensive template. At the end of the sequence, the comprehensive template was matched to each chromatogram and any peak matching with at least two chromatograms were retained in a consensus template. The consensus template contained 422 peaks and the matching rates ranged from 52% to 78%, with 196 peaks matching for all nine chromatograms. For both peak matching methods, the feature fingerprints of samples from nine regions were sifted for the largest normalized intensities and many of the indicated compounds have a known role in defining sensory properties.

Castillo et al. [55] used GC × GC–TOFMS to analyze a variety of samples for metabolomic characteristics. They developed a processing sequence of peak detection, matching, filtering, normalization, and identification. The matching algorithm used a scoring metric to choose some matches over others. For a set of

60 serum samples, almost 15,000 prospective compounds were filtered to 1540 on the basis of matching a sufficient number of chromatograms, then to 1013 compounds by mass spectral and chromatographic constraints. The resulting feature vectors were analyzed by PCA, which separated samples by their storage temperature.

Koek et al. [56] evaluated the analyst and computer time required to process GC × GC–TOFMS datasets for mouse liver samples to produce a table of 170 metabolites in 29 samples. The analysis required approximately 50 h of analyst time and 60 h of computer time, with substantial analyst time required for optimization and construction of the reference target table and dealing with problems of missing peak values. These times are indicative that reliable peak matching, even with recent software for GC × GC, is not yet automated. Subsequently, they evaluated the resulting metabolite profiles with PCA and PCA-DA.

Peak detection errors as well as the inherent ambiguity of matching both contribute to make comprehensive peak matching (i.e., matching all peaks) across many samples intractable. Trace peaks may be detected in some samples, but not in others. Coeluting analytes may be detected as separate peaks in some chromatograms but as one peak in other chromatograms. The peaks of different analytes may be incorrectly matched, especially if constituents differ from sample to sample. To overcome these challenges, researchers filter the peaks that are used for cross-sample analysis. However, such filtering is unreliable and difficult to automate. And, to the extent that peaks are correctly filtered, the analysis is no longer truly comprehensive. Despite extensive research, methods for automated peak matching still are error-prone and/or not comprehensive. Despite these problems, peak features can be effectively used in many applications for non-targeted cross-sample analysis.

5. Region features

Region features characterize multiple datapoints (e.g., summing the intensities at all datapoints in each region). Like peak features, region features can characterize larger, more meaningful chromatographic structures than datapoint features, resulting in fewer features that are less redundant. Like peak features, region features are less sensitive to misalignment than datapoint analysis.

For non-targeted analysis, the feature regions should be defined to cover the entire chromatographic space in which analytes are present. When used for cross-sample analysis, the same regions in different chromatograms are implicitly matched, thereby avoiding the matching problem that is inherent with peak features. However, either the chromatograms should be aligned or the regions should be adjusted geometrically so that the same regions in different chromatograms encompass the same analyte(s). As geometric shapes, regions are amenable to geometric transformations to fit different chromatograms in cases of variable retention times.

Two concerns with region features are that a region may encompass more than one analyte and that one analyte may be spread across more than one region. In the first case, selectivity is reduced as compared with peak features (although peak features also may not separate coeluted peaks). In the second case, multiple features for a single analyte are more susceptible to errors related to misalignment as compared with peak features (although peak features also may incorrectly split analyte peaks).

Mispelaar [4,38] created a hand-drawn mesh of contiguous polygons to subjectively encompass different groups of interest in diesel samples and demonstrated the utility of geometric transformations to better fit different chromatograms. Fig. 3 illustrates a similar mesh for GC × GC–FID [90] with automatically drawn vertical lines at linear retention indices based on the *n*-alkanes and hand-drawn lines to separate compound groups. As Mispelaar noted, some prior knowledge of the sample is required to define regions related to its components and component groups. And, as can be seen, there are regions with multiple analytes and analyte peaks spread across multiple regions.

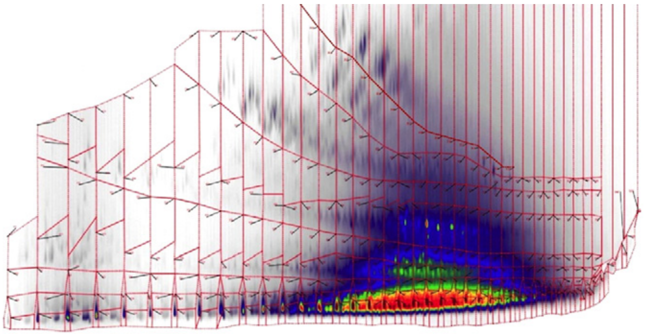

Fig. 3. A mesh of regions with automatically drawn vertical lines at linear retention indices based on the *n*-alkanes and hand-drawn crossing lines to separate compound groups [90].

To quantify weathering of an oil spill by GC × GC–FID, Arey et al. [3] created a grid with region boundaries defined by computed contours of hydrocarbon vapor pressure and aqueous solubility. With this approach, no prior knowledge of the nature of the sample is required, but regions may contain multiple analytes and analyte peaks may straddle multiple regions. To mitigate the effect of misalignment, they used trapezoidal weighting functions at the borders between regions. With contour lines that are roughly orthogonal, the grid can be remapped naturally to a rectangular array and colorized according to intensity for convenient visualization. They applied the analysis to investigate different weathering processes on oil spills, including evaporation, dissolution, biodegradation, photodegradation, and other processes. Wardlaw et al. [85] used these same lines to warp chromatographic images.

To analyze Chinese medicine volatile oils with GC × GC–TOFMS, Qiu et al. [44] used integration in four regions (mostly, but not fully covering the analytes) to compute averages and show differences among five geographical classes. Mullins et al. [91] used seven large regions to characterize compound groups in downhole fluid analysis with GC × GC–FID and GC × GC–TOFMS. They plotted ratios of the summed peak intensities within each region in a spider diagram to visualize similarities and differences. Betancourt et al. [92] used spider diagrams to visualize features for nine large compound-based regions and subdivisions of those regions split by retention indices. Ventura et al. [93] extended the approach to twelve regions. Vaz-Freire et al. [50] divided chromatograms from olive oil samples into twelve rectangular regions, then performed ANOVA and PCA with the regional features.

The principal issue with region features is that selectivity is reduced to the extent that peaks of multiple analytes are included in the same region. For some applications, such as petroleum analysis, the goal may be comprehensive group-type analysis, so loss of selectivity within groups is not problematic. However, the loss of selectivity could be a problem in many applications, especially if a critical trace analyte is in the same region as a predominant analyte that is irrelevant to the application.

6. Peak-region features

The final type of feature surveyed in this review is the peak-region. Peak-region features attempt to define one region per peak. This approach seeks to achieve the one-feature-to-one-analyte selectivity of peak features but with the implicit matching of region features.

Schmarr et al. [53,54] and Reichenbach and co-workers [51,52,1] described similar approaches to defining regions for individual peaks detected across multiple samples. Schmarr and Bernhardt indicated that this general approach is common for 2D gel electrophoresis. After preprocessing, including alignment, the chromatograms are combined (e.g., simply by addition or other fusion operations [94]) to form a single chromatogram that is reflective of all of the constituents in all samples. Then, the boundaries that delineate each peak are recorded as a region in a template. That template is then geometrically mapped back to each chromatogram and each region defines a feature for each chromatogram. The features are comprehensive, accounting for every analyte, and feature matching is implicitly performed by the retention-time mapping.

Schmarr and Bernhardt [53] analyzed 32 samples of volatiles of different fruits by GC × GC–MS. They performed baseline correction with the rolling-ball method, then manually generated warp graphs to determine warping transforms to align 31 chromatograms to a reference chromatogram. Then, each of the chromatograms was aligned by the warping transform and combined using a weighted-mean "union fusion" [94]. They manually detected more than 700 spots indicative of peaks in the fused chromatogram. Then, the spot patterns were mapped back to each chromatogram according to the inverse of its warping transform and the intensities for each region in each chromatogram were computed. The software package that they used was optimized for gel electrophoresis rather than GC × GC, so much of the processing was manual, requiring about 5 h of an analyst's time for the 32 samples. They used HCA and PCA with the resulting peak-region features to cluster samples. The different fruits (apples, pears, and quince) formed clear clusters. The two pear varieties and some of the six apple varieties formed sub-clusters. The mass-spectral signatures were used for compound identification of spots which were statistically relevant for differentiation. Using a similar approach for analyzing red wines subjected to microoxygenation (MOX), Schmarr et al. [54] were able to differentiate MOX treatments and specific varietal and technological effects. They were able to identify areas in the 2D chromatograms that were most responsible for discrimination among different MOX treatments and the loadings of individual aroma compounds suggested a set of markers for the MOX-induced modifications of volatiles.

Cordero et al. [51] analyzed samples of coffees and junipers by GC × GC–MS. After preprocessing including peak detection, they identified peaks that could be matched reliably across all chromatograms. These reliable peaks were the basis of a registration template with mass spectral matching rules that then was used to determine a geometric transform to align the chromatograms. After alignment, the chromatograms were summed to create a cumulative chromatogram. In three chromatograms of coffee samples, about 1700 peaks were detected, about half of which were reliable. They manually drew a mesh of about 1100 regions which were combined with the registration peaks to create a feature template that could be matched to individual chromatograms thereby transforming the regions to maintain their positions relative to the reliable peaks. They sifted the features by intensity, standard deviation, and relative standard deviation to select relevant features but did not perform MVA because of the small number of samples. Many of the indicated compounds were known botanical, technological, and/or aromatic markers for coffee. For the analysis of five chromatograms of juniper samples, there were about 100 reliable peaks and 727 peak-regions were drawn. Reichenbach et al. [52] used the same approach for 39 urine samples analyzed by LC × LC. Then, they performed classification with SVM and k-NN, evaluating the performance using cross-validation.

Reichenbach et al. [1] analyzed data from GC × GC with high-resolution mass spectrometry (HRMS) of samples from breast cancer tumors. There were eighteen samples each from different individuals, with six samples each for grades 1–3 as determined by a cancer pathologist. They followed the same approach as Cordero et al. [51] except that the process, including drawing the regions around the peaks detected in the cumulative chromatogram, was performed automatically by newer software. About 3300 peaks were detected in each of the eighteen individual chromatograms, but only thirteen were reliable across all eighteen chromatograms.

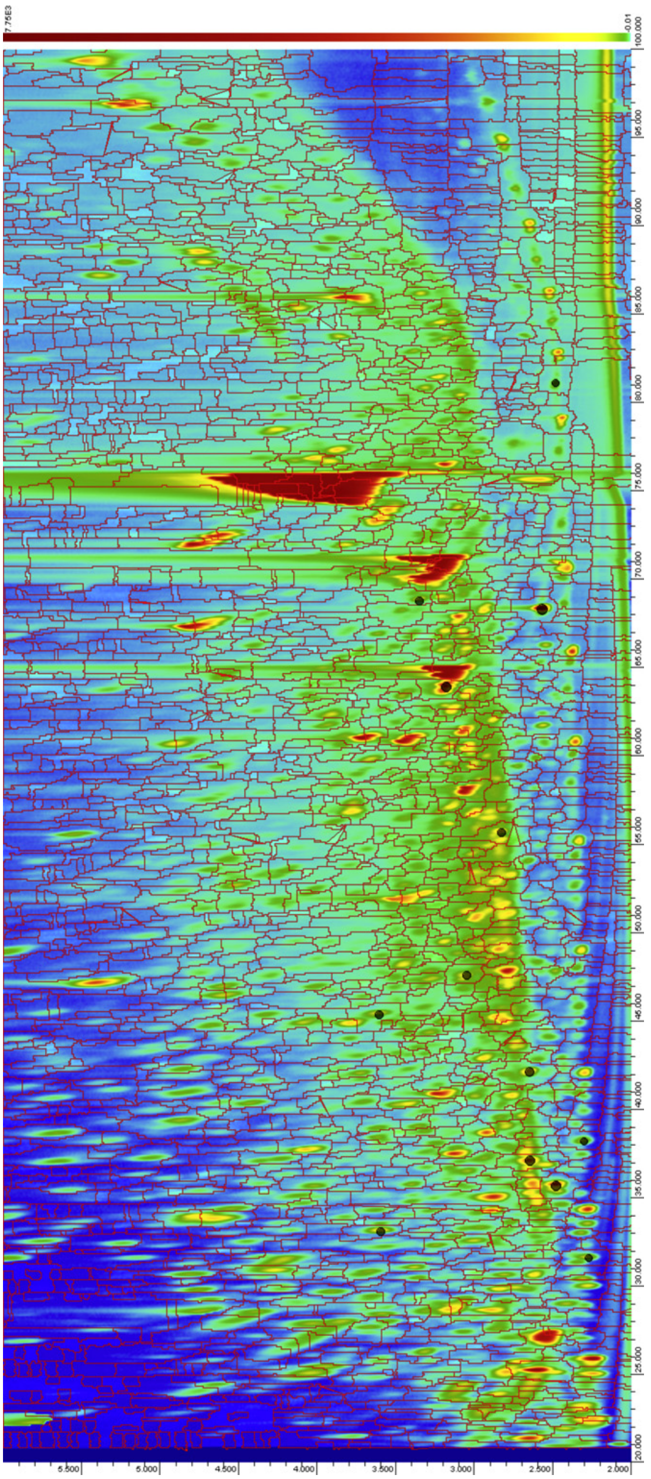

Fig. 4. Cumulative chromatogram for eighteen breast-cancer tumor samples overlaid with the feature template (registration peaks shown with dark ovals and region features shown with red outlines). The color bar shows the logarithmic pseudocolorization mapping [1]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Note that reliability was defined as bidirectional matching between all possible pairs (more than 300 matches for each common peak). In the cumulative chromatogram, more than 3300 peak-regions were defined. Fig. 4 shows the cumulative chromatogram overlaid with black ovals for the reliable peaks used for registration and red outlines for the peak-regions. They applied several machine learning methods with the peak-region features to classify samples by tumor grade and to indicate potential biomarkers for tumor grade which then were investigated using the high-resolution mass spectra.

The peak-region approach is more comprehensive than using reliably matched peak features and is more selective than region features. As with the other feature methods, misalignment is a potential source of errors. As with peak features, peak detection errors, such as unseparated coelutions and incorrectly split peaks, are another source of errors for peak-region features.

7. Conclusion

A common goal of chemical analysis is to compare samples, either for a few specific compounds (targeted analysis), for groups of compounds (group-type analysis), or for all compounds (i.e., non-targeted analysis). The key to comparative analyses is to establish correspondences between features of different data sets, e.g., recognizing that a peak in the data for one sample and a peak in the data for another sample are induced by the same compound. Establishing correspondences — *feature matching* — is necessary before it is possible to perform comparisons and pattern recognition across sample sets.

Targeted analyses and group-type analyses are more straightforward than non-target analyses. In targeted analyses, the compounds of interest are known, so the chromatography can be tailored to provide selectivity for those compounds and the data processing methods can be refined for detecting and recognizing the features for those compounds. For group-type analysis, the method need not be selective of every individual analyte, so many problems of feature generation (e.g., peak unmixing) and matching can be avoided. Comprehensive non-target analyses are more difficult because the most relevant compounds are unknown, so the chromatography and data processing cannot be tuned specifically for individual compounds or for groups of compounds.

Non-targeted cross-sample analysis is especially difficult because it requires the analysis of all analytes in all chromatograms of a sample set. Applications of non-targeted cross-sample analysis include sample classification, chemical fingerprinting, monitoring, sample clustering, and chemical marker discovery. Comprehensive two-dimensional chromatography is a powerful technology for separating complex mixtures and so is well suited for comprehensive non-targeted analysis, but fully extracting chemical information from large and complex datasets is challenging and the subject of ongoing research. And, the difficulty of comparative analyses increases with the size of the sample set.

Feature matching for comprehensive two-dimensional chromatography can be based on retention times, spectral signature, detected intensity, and/or other characteristics of features. Past research on non-targeted cross-sample analysis with comprehensive two-dimensional chromatography has demonstrated the usefulness of qualitative visualization, individual datapoints, detected peaks, chromatographic regions, and comprehensive peak-regions.

Each type of feature has advantages and disadvantages. Visualization is simple and intuitive, but is not quantitative, important differences may not be visible, and working with large sample sets is difficult. Datapoint features are highly selective and implicitly matched across aligned chromatograms, but they are subject to misalignment errors and generate a large number of features, many of which are redundant. Peak features characterize individual analytes and so are especially consistent with analytical goals, but peak matching is an intractable problem. Region features are more attuned to meaningful analytical characteristics than datapoint features and are easier to match across samples than peak features, but they may not be as selective as datapoint or peak features. Peak-regions define a region for each peak across chromatograms and so aim for selectivity and accurate feature matching, but still are subject to errors from misalignment and peak detection failures.

Future research will refine, compare, and combine these approaches. There has been little research to deeply examine the variables that affect feature generation and matching in the different approaches and to validate performance in cross-sample analyses. Advances in instrument technologies could contribute to improved feature generation and matching, e.g., with increased repeatability and reproducability, greater mass spectrometric accuracy, and more effective column sets. Feature generation and matching might be improved by better preprocessing methods, especially for detection of coeluted peaks, but also for baseline correction and alignment. Likewise, more research is needed to compare the performance of different approaches for feature generation and matching in different applications. Ultimately, a hybrid approach, using a combination of different approaches, may be most effective e.g., peak features for peaks that can be reliably matched, and peak-region, region, or datapoint features for other chromatographic data. Again, such combined approaches require a better understanding of the variables that affect the performance of the different approaches.

Acknowledgements

This work was supported in part by the U.S. National Science Foundation under Award Number IIP-1013180 and by the Nebraska Center for Energy Sciences Research.

References

[1] S.E. Reichenbach, X. Tian, Q. Tao, E.B. Ledford, Z. Wu, O. Fiehn Jr., Talanta 83 (2011) 1279.
[2] R.B. Gaines, G.S. Frysinger, M.S. Hendrick-Smith, J.D. Stuart, Environ. Sci. Technol. 33 (1999) 2106.
[3] J.S. Arey, R.K. Nelson, C.M. Reddy, Environ. Sci. Technol. 41 (2007) 5738.
[4] V. G. van Mispelaar, Chromametrics, Ph.D. thesis, University of Amsterdam, 2005.
[5] S.R. Sternberg, Computer 16 (1983) 22.
[6] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford Jr., J. Chromatogr. A 985 (2003) 47.
[7] Y. Zhang, H.-L. Wu, A.-L. Xia, L.-H. Hu, H.-F. Zou, R.-Q. Yu, J. Chromatogr. A 1167 (2007) 178.
[8] S.E. Reichenbach, P.W. Carr, D.R. Stoll, Q. Tao, J. Chromatogr. A 1216 (2009) 3458.
[9] J. Beens, H. Boelens, R. Tijssen, J. Blomberg, J. High Resolut. Chromatogr. 21 (1998) 47.
[10] Q. Song, A. Savant, S.E. Reichenbach, E.B. Ledford Jr., in: Visual Information Processing, Proc. SPIE., vol. 3808,1999, p. 2.
[11] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Chemom. Intell. Lab. Syst. 71 (2004) 107.
[12] S. Peters, G. Vivó-Truyols, P. Marriott, P. Schoenmakers, J. Chromatogr. A 1156 (2007) 14.
[13] E.J.C. van der Klift, G. Vivó-Truyols, F.W. Claassen, F.L. van Holthoon, T.A. van Beek, J. Chromatogr. A 1178 (2008) 43.
[14] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796.
[15] B.J. Prazen, C.A. Bruckner, R.E. Synovec, B.R. Kowalski, J. Microcolumn Sep. 11 (1999) 97.
[16] C.G. Fraga, B.J. Prazen, R.E. Synovec, J. High Resolut. Chromatogr. 23 (2000) 215.
[17] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Anal. Chem. 73 (2001) 675.
[18] B.J. Prazen, K.J. Johnson, A. Weber, R.E. Synovec, Anal. Chem. 73 (2001) 5677.
[19] A.E. Sinha, C.G. Fraga, B.J. Prazen, R.E. Synovec, J. Chromatogr. A 1027 (2004) 269.
[20] A.E. Sinha, J.L. Hope, B.J. Prazen, C.G. Fraga, E.J. Nilsson, R.E. Synovec, J. Chromatogr. A 1056 (2004) 145.
[21] C.G. Fraga, C.A. Corley, J. Chromatogr. A 1096 (2005) 40.
[22] H. Kong, F. Ye, X. Lu, L. Guo, J. Tian, G. Xu, J. Chromatogr. A 1086 (2005) 160.

[23] J.C. Hoggard, R.E. Synovec, Anal. Chem. 79 (2007) 1611.

[24] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, J. Chromatogr. A 1216 (2009) 4020.

[25] Z.-D. Zeng, S.-T. Chin, H.M. Hugel, P.J. Marriott, J. Chromatogr. A 1218 (2011) 2301.

[26] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154.

[27] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141.

[28] V.G. van Mispelaar, A.C. Tas, A.K. Smilde, P.J. Schoenmakers, A.C. van Asten, J. Chromatogr. A 1019 (2003) 15.

[29] K.J. Johnson, B.J. Prazen, D.C. Young, R.E. Synovec, J. Sep. Sci. 27 (2004) 410.

[30] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, Anal. Chem. 77 (2005) 7735.

[31] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101.

[32] B.V. Hollingsworth, S.E. Reichenbach, Q. Tao, A. Visvanathan, J. Chromatogr. A 1105 (2006) 51.

[33] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Anal. Chem. 80 (2008) 2664.

[34] M.F. Almstetter, I.J. Appel, M.A. Gruber, C. Lottaz, B. Timischl, R. Spang, K. Dettmer, P.J. Oefner, Anal. Chem. 81 (2009) 5731.

[35] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiébaut, B. Teillet, D.N. Rutledge, J. Chromatogr. A 1216 (2009) 2866.

[36] T. Gröger, R. Zimmermann, Talanta 83 (2011) 1289.

[37] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225.

[38] V.G. van Mispelaar, H.-G. Janssen, A.C. Tas, P.J. Schoenmakers, J. Chromatogr. A 1071 (2005) 229.

[39] R.A. Shellie, W. Welthagen, J. Zrostliková, J. Spranger, M. Ristow, O. Fiehn, R. Zimmermann, J. Chromatogr. A 1086 (2005) 83.

[40] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Anal. Chem. 78 (2006) 2700.

[41] K.M. Pierce, J.L. Hope, J.C. Hoggard, R.E. Synovec, Talanta 70 (2006) 797.

[42] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Anal. Chem. 78 (2006) 5068.

[43] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756.

[44] Y. Qiu, X. Lu, T. Pang, S. Zhu, H. Kong, G. Xu, J. Pharm. Biomed. Anal. 43 (2007) 1721.

[45] T. Grögera, M. Schäffer, M. Pütz, B. Ahrens, K. Drew, M. Eschner, R. Zimmermann, J. Chromatogr. A 1200 (2008) 8.

[46] X. Guo, M.E. Lidstrom, Biotechnol. Bioeng. 99 (2008) 929.

[47] R.E. Mohler, B.P. Tu, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, J. Chromatogr. A 1186 (2008) 401.

[48] E. Gaquerel, A. Weinhold, I.T. Baldwin, Plant Physiol. 149 (2009) 1408.

[49] X. Li, Z. Xu, X. Lu, X. Yang, P. Yin, H. Kong, Y. Yu, G. Xu, Anal. Chim. Acta 633 (2009) 257.

[50] L.T. Vaz-Freire, M.D.R. Gomes da Silva, A.M.C. Freitas, Anal. Chim. Acta 633 (2009) 263.

[51] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, S.E. Reichenbach, X. Tian, Q. Tao, J. Chromatogr. Sci. 48 (2010) 251.

[52] S.E. Reichenbach, X. Tian, Q. Tao, D.R. Stoll, P.W. Carr, Talanta 83 (2010) 1365.

[53] H.-G. Schmarr, J. Bernhardt, J. Chromatogr. A 1217 (2010) 565.

[54] H.-G. Schmarr, J. Bernhardt, U. Fischer, A. Stephan, P. Müller, D. Durner, Anal. Chim. Acta 672 (2010) 114.

[55] S. Castillo, I. Mattila, J. Miettinen, M. Ores˘ix˘, T. Hyötyläinen, Anal. Chem. 83 (2011) 3058.

[56] M.M. Koek, F.M. van der Kloet, R. Kleemann, T. Kooistra, E.R. Verheij, T. Hankemeier, Metabolomics 7 (2011) 1.

[57] G.T. Ventura, G.J. Hall, R.K. Nelson, G.S. Frysinger, B.R.A.E. Pomerantz, O.C. Mullins, C.M. Reddy, J. Chromatogr. A 1218 (2011) 2584.

[58] J. Vial, B. Pezous, D. Thiébaut, P. Sassiat, B. Teillet, X. Cahours, I. Rivals, Talanta 83 (2011) 1295.

[59] J. Blomberg, P.J. Schoenmakers, U.A.Th. Brinkman, J. Chromatogr. A 972 (2002) 137.

[60] P. Marriott, R. Shellie, Trends Anal. Chem. 21 (2002) 573.

[61] L. Mondello, A.C. Lewis, K.D. Bartle, Multidimensional Chromatography, Wiley, Chichester, UK, 2002.

[62] J. Dallüge, J. Beens, U.A.Th. Brinkman, J. Chromatogr. A 1000 (2003) 69.

[63] S. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Chemom. Intell. Lab. Syst. 71 (2004) 107.

[64] A.E. Sinha, B.J. Prazen, R.E. Synovec, Anal. Bioanal. Chem. 378 (2004) 1948.

[65] P.Q. Tranchida, P. Dugo, G. Dugo, L. Mondello, J. Chromatogr. A 1054 (2004) 3.

[66] P.Q. Tranchida, P. Dugo, G. Dugo, L. Mondello, Trends Anal. Chem. 26 (2007) 191.

[67] M. Adahchour, J. Beens, U.A.Th. Brinkman, J. Chromatogr. A 1186 (2008) 67.

[68] S.A. Cohen, M.R. Schure (Eds.), Multidimensional Liquid Chromatography: Theory and Applications in Industrial Chemistry and the Life Sciences, John Wiley and Sons, New York, NY, 2008.

[69] P. Dugo, F. Cacciola, T. Kumm, G. Dugo, L. Mondello, J. Chromatogr. A 1184 (2008) 353.

[70] L. Mondello, P.Q. Tranchida, P. Dugo, G. Dugo, Mass Spectrom. Rev. 27 (2008) 101.

[71] O. Amador-Muñoz, P.J. Marriott, J. Chromatogr. A 1184 (2008) 323.

[72] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, J. Chromatogr. A 1184 (2008) 341.

[73] H.J. Cortes, B. Winniford, J. Luong, M. Pursch, J. Sep. Sci. 32 (2009) 883.

[74] L. Ramos, Comprehensive Two Dimensional Gas Chromatography, Elsevier, Oxford, UK, 2009.

[75] J.C. Hoggard, R.E. Synovec, L. Ramos, Comprehensive Two Dimensional Gas Chromatography, Elsevier, Oxford, UK, 2009, p. 107.

[76] S.E. Reichenbach, L. Ramos, Comprehensive Two Dimensional Gas Chromatography, Elsevier, Oxford, UK, 2009, p. 77.

[77] S.E. Reichenbach, L. Ramos, Comprehensive Two Dimensional Gas Chromatography, Elsevier, The Netherlands, 2009, p. 77.

[78] M.M. Bushey, J.W. Jorgenson, Anal. Chem. 62 (1990) 161.

[79] J. Blomberg, P.J. Schoenmakers, J. Beens, R. Tijssen, J. High Resolut. Chromatogr. 20 (1997) 539.

[80] C.M. Reddy, T.I. Eglinton, A. Hounshell, H.K. White, L. Xu, R.B. Gaines, G.S. Frysinger, Environ. Sci. Technol. 36 (2002) 4754.

[81] H.-G. Janssen, W. Boers, H. Steenbergen, R. Horsten, E. Flöter, J. Chromatogr. A 1000 (2003) 385.

[82] R.M.M. Perera, P.J. Marriott, I.E. Galbally, Analyst 127 (2002) 1601.

[83] J.L. Hope, B.J. Prazen, E.J. Nilsson, M.E. Lidstrom, R.E. Synovec, Talanta 65 (2005) 380.

[84] R.K. Nelson, B.S. Kile, D.L. Plata, S.P. Sylva, L. Xu, C.M. Reddy, R.B. Gaines, G.S. Frysinger, S.E. Reichenbach, Environ. Forensics 7 (2006) 33.

[85] G.D. Wardlaw, J.S. Arey, C.M. Reddy, R.K. Nelson, G.T. Ventura, D.L. Valentine, Environ. Sci. Technol. 42 (2008) 7166.

[86] User's Guide to SIMCA-P, SIMCA-P+, Umetrics AB, Version 11.0 edition, 2005.

[87] C. Oh, X. Huang, F.E. Regnier, C. Buck, X. Zhang, J. Chromatogr. A 1179 (2008) 205.

[88] S.E. Reichenbach, P.W. Carr, D.R. Stoll, Q. Tao, J. Chromatogr. A 1216 (2009) 3458.

[89] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, P. Schieberle, S.E. Reichenbach, Q. Tao, J. Chromatogr. A 1217 (2010) 5848.

[90] S. Reichenbach, Q. Tao, D. E. Hutchinson, S. B. Cabanban, H. A. Pham, W. E. Rathbun, H. Wang, in: Pittcon, p. 540.

[91] O.C. Mullins, G.T. Ventura, R.K. Nelson, S.S. Betancourt, B. Raghuraman, C.M. Reddy, Energy Fuels 22 (2008) 496.

[92] S.S. Betancourt, G.T. Ventura, A.E. Pomerantz, O. Viloria, F.X. Dubost, J. Zuo, G. Monson, D. Bustamante, J.M. Purcell, R.K. Nelson, R.P. Rodgers, C.M. Reddy, A.G. Marshall, O.C. Mullins, Energy Fuels 23 (2008) 1178.

[93] G.T. Ventura, B. Raghuraman, R.K. Nelson, O.C. Mullins, C.M. Reddy, Org. Geochem. 41 (2010) 1026.

[94] S. Luhn, M. Berth, M. Hecker, J. Bernhardt, Proteomics 3 (2003) 1117.