CHAPTER 4

Data Acquisition, Visualization, and Analysis

Stephen E. Reichenbach

Contents

1. INTRODUCTION

An introduction to informatics for comprehensive two-dimensional gas chromatography (GC×GC) should begin with the strikingly beautiful and

Comprehensive Analytical Chemistry, Volume 55
ISSN: 0166-526X, DOI 10.1016/S0166-526X(09)05504-4

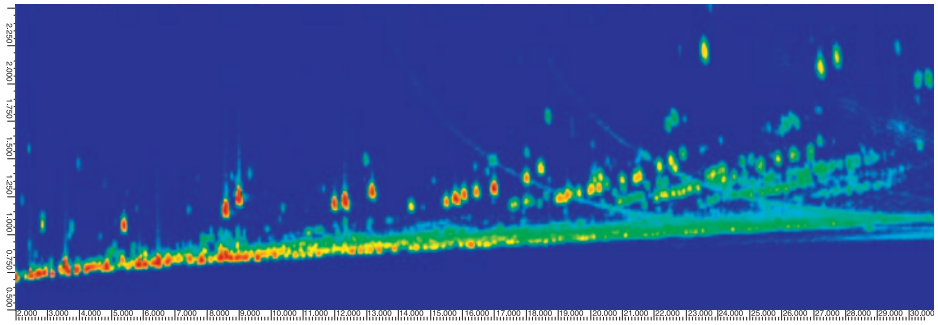

Figure 1 GC×GC data from a gasoline analysis visualized as a digital image. Only a portion of the data is shown. (This and other figures were generated with GC Image® software [1]. Data supplied by Zoex Corporation.)

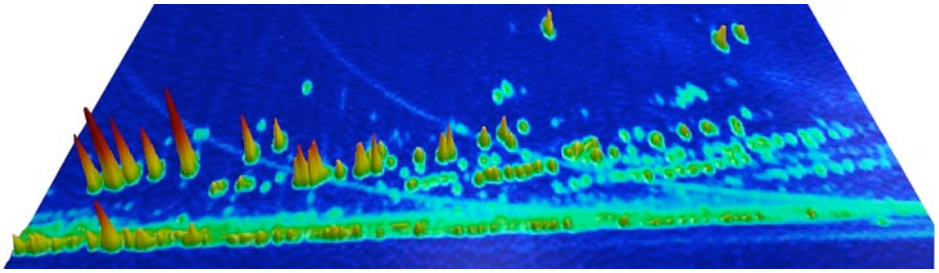

Figure 2 GC×GC data visualized as a three-dimensional surface. A subregion of the data from Figure 1 is shown.

complex pictures of data visualization. Whether viewed as a pseudocolorized two-dimensional image, as in Figure 1, or as a projection of a three-dimensional surface, as in Figure 2, GC×GC visualizations impress even observers lacking chromatographic expertise with their colorful and multitudinous features. Chromatographers recognize, within these pictures, complex patterns embedding a wealth of multidimensional chemical information. The richness of GC×GC data is immediately apparent, but the size and complexity of GC×GC data pose significant challenges for chemical analysis.

This chapter examines methods and information technologies for GC×GC data acquisition, visualization, and analysis. The quantity and complexity of GC×GC data make human analyses of GC×GC data difficult and time-consuming and motivate the need for computer-assisted and automated processing. GC×GC transforms chemical samples into raw data; information technologies are required to transform GC×GC data into chemical information.

The typical data flow is a sequence of: acquiring and storing raw data, processing data to correct artifacts, detecting and identifying chemical peaks, and analyzing datasets to produce higher-level information (including quantification) and reports. In applications for which the analysis is fairly well understood and routine, information technologies may fully automate this process.

However, because GC×GC is so powerful, it frequently is used for analyses that are not well understood or are not routine. In such cases, information technologies must support semi-automated processing, visual interpretation, and interactive analysis.

This chapter addresses the following fundamental tasks in transforming GC×GC data into chemical information:

- Acquiring and formatting data for storage, access, and interchange.
- Visualizing multidimensional data.
- Processing data to remove acquisition artifacts and detect peaks.
- Identifying chemical constituents.
- Analyzing datasets for higher-level information and reporting.

2. DATA ACQUISITION

Although GC×GC is a true two-dimensional separation, the process serializes the data — producing data values in a sequence. In GC×GC, the first column progressively separates and presents eluates to the modulator, which iteratively collects and introduces them into the second column, which then progressively separates and presents eluates to the detector. As explained in detail in Chapter 2, in the detector, the analog-to-digital (A/D) converter samples the chromatographic signal at a specified frequency. In concept, this operation is similar to how some optical systems create an image with as few as one detector by progressively scanning the detector(s) across the two spatial dimensions, but, in GC×GC, the two dimensions are the two retention times. Then, the digitized data and relevant metadata (information about the data) are stored in a file with a defined format for subsequent access.

2.1 Modulation and sampling

The modulation frequency and the detector sampling frequency typically are under user control. Setting these frequencies (subject to the limitations of the hardware) involves trade-offs between resolution and other constraints. The desire for high resolution suggests that the modulation and sampling rates should be as rapid as possible. A Gaussian peak is not band-limited, so truly sufficient sampling is not possible. Therefore, higher modulation and sampling rates provide greater information capacity and increased resolution for detecting co-eluted peaks. However, the modulation frequency must allow adequate intervals for separations in the second column, and the sampling frequency involves a trade-off in data size (i.e., higher sampling frequencies generate more data) and diminishing returns in selectivity and precision. Full consideration of these and other issues (such as duty cycle and noise) in setting the modulation and sampling frequencies involves instrumental and application-specific concerns that are beyond the scope of this chapter, but consideration of the data suggests general guidelines.

Experimental and theoretical studies [2] suggest that the modulation rate should be at least one cycle per two times the primary peak standard deviation σ_1 (i.e., the standard deviation of the peak width from the first column separation), which translates to at least four modulation cycles over $8\sigma_1$ (the effective width of peaks from the first-column separation). The considerations for GC×GC detector frequencies are similar to those for traditional one-dimensional chromatography, for which a rate of at least one sample per peak standard deviation is recommended [3,4], that is, eight samples over $8\sigma_2$ (the effective width of peaks from the second-column separation). With these considerations, Murphy et al. [5] recommend that method development begin with determining the shortest time for adequate chromatographic separation in the second column and then a first-dimension method be used that provides peak widths of at least four times the modulation interval. With the wide variety of chemical mixtures and analytical goals for GC×GC, a broad range of modulation and sampling frequencies are used. Modulation cycles from 2 to 20 seconds (s) and sampling frequencies from 25 to 200 hertz (Hz) are not unusual. Again, however, the application should be considered; slow modulation and sampling rates relative to peak width may be sufficient for applications that require only quantification of well-separated peaks, and fast modulation and sampling rates relative to peak width may be required for applications that involve compounds that are difficult to separate.

A common problem in GC×GC data processing is inadequate sampling of the first-column output; that is, the modulation period is too long with respect to the first-column peak widths, or, put another way, the first-column chromatography produces peaks too narrow for the modulation period. Of course, if the modulation period is constrained by the time required for second-column separations, then broadening the peak widths from the first column may require longer runs (thereby increasing cost). Inadequate sampling of the second-column output is less commonly problematic because most detectors used for GC×GC are fast and most laboratories typically use detector sampling rates that exceed what is required for the analysis (and so generate more data than may be necessary). However, as explained in Chapter 2, some types of detectors — for example, quadrupole mass spectrometer (qMS), atomic emission detector (AED), and electron capture detector (ECD) — may be challenged by the acquisition speeds required for GC×GC.

2.2 Digitization and coding

GC×GC systems use an A/D converter to map the intensity of the chromatographic signal to a digital number (DN). Among the many types of detectors used with GC×GC, the major distinction is between detectors that produce a single number at each time sample of the chromatogram, such as a flame-ionization detector (FID) and a sulfur chemiluminescence detector (SCD), and multichannel detectors that produce multiple values (typically, over a spectral range) for each time sample, such as a mass spectrometer (MS). In either case, each DN is represented with a limited number of bits indicating a value in a limited range with limited precision.

Because GC×GC can produce large datasets, GC×GC systems often employ data compression in their file formats. Sampling at 200 Hz, a detector for single values with a 48-bit dynamic range (as supported by Agilent's IQ data file format [6]) produces data at the rate of 4.3 megabytes/hour (MB/h). Most programming languages must perform arithmetic on 48-bit values with 64-bit long integers or 64-bit double-precision floating-point numbers. Mass spectrometers can produce data at sub-1 GHz (e.g., one 8-bit spectral intensity per nanosecond), a data rate of greater than 1 gigabyte/sec (GB/s). In order to more efficiently store data, GC×GC systems may compress the data. For example, because data values are correlated with neighboring values in the sequence, Agilent's IQ data file format implements a second-order backward differential coding that compresses values from a 48-bit range to 2 bytes. Even more aggressive compression commonly is used for MS data. For example, ORTEC's FastFlight-2TM [7] can accumulate successive spectra in hardware and output only the summed spectra for a much smaller data rate. In a MS with GHz raw speed, summing 100 transient spectra in 100 K channels generates 100 spectra per second (compared to 10,000 raw spectra per second). The FastFlight2 also offers a lossless compression mode that uses fewer bytes to represent smaller values and a lossy compression mode that detects and encodes only the spectral peaks in the MS data — a process sometimes called centroiding because each spectral peak is represented by a single centroid indicating the center, intensity, and sometimes the peak width.

2.3 File formats

Most GC×GC systems use a proprietary data file format, which affords vendors a high degree of control (e.g., to implement data compression), but which poses a barrier and inconvenience for sharing or processing data across systems. Currently, there is no standard format for GC×GC data, but GC×GC data can be shared using nonstandard text files or existing standards for gas chromatography (GC) data. GC×GC data can be converted to text, for example, ASCII-format comma-separated values (CSV), but the resulting files are non-standard and are larger than binary or compressed data files. The ASTM has issued Analytical Data Interchange (ANDI) standards for chromatography [8] and MS [9]. These standards lack some requirements for GC×GC metadata (e.g., a metadata element for the modulation cycle) but can be used to communicate raw data and other chromatographic metadata. These standards were developed primarily for data interchange and lack some desirable features for more routine use. Another limitation of the ANDI standards is that the network Common Data Form (netCDF) [10], upon which the standards are built, was defined for 32-bit computing systems, limiting their usability for data larger than 2 GB. The ASTM has sanctioned an effort to develop a new format standard for analytical chemistry data, the Analytical Information Markup Language (AnIML) [11,12], utilizing the eXtensible Markup Language (XML) [13]. Standard formats for analytical chemistry data facilitate data portability and interchange, but despite such considerations proprietary GC formats have continued to dominate the market.

3. VISUALIZATION

Visualization is a powerful tool for qualitative analysis of GC×GC data (e.g., to troubleshoot the chromatography). Various types of visualizations are useful: two-dimensional images provide a comprehensive overview, three-dimensional visualizations effectively illustrate quantitative relationships over a large dynamic range, one-dimensional graphs are useful for overlaying multivariate data, tabular views reveal the numeric values in the data, and graphical and text annotations communicate additional information. This section explores some of the methods and considerations in the various types of visualizations.

3.1 Image visualizations

3.1.1 Rasterization

A fundamental visualization of GC×GC data is as a two-dimensional image. GC×GC data, which is acquired sequentially, can be reorganized as a raster — a two-dimensional array, matrix, or grid of picture elements called pixels — in which each pixel value is the intensity of the detector signal. As a two-dimensional array of intensities, GC×GC data has many similarities with other types of digital images and so many methods and techniques from the field of digital image processing can be applied or adapted for GC×GC data visualization and processing.

The standard approach for rasterization is to arrange the data values acquired during a single modulation cycle as a column of pixels, so that the ordinate (*Y*-axis, bottom-to-top) is the elapsed time for the second-column separation, and then to arrange these pixel columns so that the abscissa (*X*-axis, left-to-right) is the elapsed time for the first-column separation. This ordering presents the data in the commonly used right-handed Cartesian coordinate system, with the first-column retention time as the first index into the array. Other orderings are possible but less commonly used. The problems of correctly synchronizing the columns of data with the modulation cycle and of modulation cycles that are not evenly divisible by the detector sampling-interval are examined in Section 4.1.

3.1.2 Colorization

For presentation as an image, the pixels are colorized; that is, the GC×GC values are mapped to colors of the display device. Scalar values, such as single-valued GC×GC data, can be colorized simply on an achromatic grayscale, familiar from so-called black-and-white images. Scalar values can be extracted from multi-spectral data in various ways, for example, by adding all intensities in each spectrum to compute the total intensity count (TIC) of the data point or by taking the value in a selected ''channel'' of the spectrum. A grayscale mapping typically is defined by setting a lower bound, below which values are mapped to black; an upper bound, above which values are mapped to white; and a function to map values between the bounds to shades of gray, with brightness increasing with value. Linear, logarithmic, and exponential mapping functions are useful for different effects: linear mapping treats gradations at all intensity levels similarly;

logarithmic mapping emphasizes gradations nearer the lower bound; and exponential mapping emphasizes gradations nearer the upper bound. Although grayscale colorization provides a straightforward ordering of values from small to large that is intuitively meaningful, humans may be able to distinguish fewer than 100 distinct grayscale gradations [14]. Therefore, grayscale images cannot effectively communicate many differences among values over a large dynamic range such as is common for GC×GC data.

Pseudocolorization takes advantage of the differing sensitivities in human vision for different frequencies of light [14]. These differing sensitivities enable ''color'' perception, with greater selectivity than for grayscale. Because humans have trichromatic vision based on three types of color receptors (cones), a trichromatic color model is sufficient for image colorization. Various trichromatic color models have been developed. RGB (with values for red, green and blue) and HSV (with values for hue, saturation, and brightness value) are widely used color models for digital imaging.

Pseudocolorization maps data values with three independent functions for the three color components. The mapping functions for the color components typically are not monotonically nondecreasing (as grayscale mapping functions typically are), so discerning relative values in a pseudocolor image is not as straightforward as with grayscale (for which brighter means larger). However, a good pseudocolor scale can communicate a clear ordering of values. For example, topographic and temperature images commonly use a pseudocolor scale sometimes called cold-to-hot, which has a mapping from small to large that progresses through blue, cyan, green, yellow, and red, with intermediate colors. In Figure 1, the color scale has the smaller values of the background colorized dark blue and the larger values of the peaks colorized with the cold-to-hot scale to show increasing values. This mapping is easily interpreted because it is familiar. Pseudocolor images can present many distinguishable colors, but there is a trade-off between having a pseudocolor scale with an ordinal progression that is simple to understand and the number of gradations that can be discerned: an easily understood scale visually differentiates a smaller number of gradations, and a scale that visually differentiates a larger number of gradations makes the value ordering more difficult to understand.

Pseudocolorization offers better visualization than grayscale for gradations across a wide dynamic range of values, but to be effective the mapping still must allocate color variations to the value range according to the presence of gradations. Specifying pseudocolorization interactively can be tedious and difficult, so automated determination of pseudocolor mapping is useful. Gradient-Based Value Mapping (GBVM) [15] is an automated method for mapping GC×GC data values onto a color scale, for example, the cold-to-hot scale. For a given dataset, GBVM builds a value-mapping function that emphasizes gradations in the data while maintaining ordinal relationships of the values. The first step computes the gradient (local difference) at each pixel. Then, the pixels (with computed gradients) are sorted by value, and the relative cumulative gradient magnitude is computed for the sorted array. The GBVM function is the mapping from pixel value to the relative cumulative gradient magnitude of the sorted

array. GBVM is effective at showing local differences across a large dynamic range.

Each resolved chemical compound in a sample increases the value in a small cluster of pixels, which, if the colorization effectively shows local differences, are seen as a localized spot with different colors than the surrounding background. If the colorization is not effective over the full dynamic range, spots with small values may not be visible or spots with large values may not show significant relative differences.

3.1.3 Navigation

Standard operations for navigating digital images include panning, scrolling, and rescaling. Rescaling requires resampling the data — creating a displayed image with more pixels to zoom in or a displayed image with fewer pixels to zoom out. (Visualization does not change the underlying data used for later processing.) Enlarging an image by rescaling entails reconstruction, which is the task of rebuilding the signal at resampling points between the data values. Popular methods for digital image reconstruction include nearest-neighbor interpolation, bilinear interpolation, and various methods using cubic polynomial functions for interpolation or approximation [14]. Bilinear interpolation provides a good compromise between quality and computational overhead. It is important to remember that reconstruction estimates signal values and that large zoom factors entail numerous estimates. Therefore, although nearest-neighbor interpolation creates blocky images with less accurate reconstruction, the result makes clear the modulation and sampling rates of the data. Similarly, nearest-neighbor interpolation will show changes in the aspect ratio imposed during rescaling (e.g., to compensate for different sampling rates in the two dimensions, such as undersampling the first-column separation and oversampling the second-column separation). Figure 3 compares bilinear and nearest-neighbor interpolation. Bilinear interpolation shows a spot that more closely represents the continuous peak produced by chromatography. Nearest-neighbor interpolation shows rectangular pixels that make clear the discrete nature of the digitized signal.

3.1.4 Qualitative analysis

Visualization can quickly and clearly show important characteristics of GC×GC data, including problems related to the chromatography. Three such examples are considered briefly here. First, if the retention time of a compound in any second-column separation exceeds the length of the modulation cycle, the associated compound will elute during a subsequent modulation cycle and the peak will appear as a spot that is wrapped around into a subsequent column of pixels in the image. If the retention time is only slightly too long, the spot will appear in the otherwise blank region at the bottom of the image corresponding to the void time of the next second-column separation. This problem can be recognized upon visual inspection, and the chromatographer can change the acquisition settings, for example, lengthening the modulation cycle time or accelerating the second-column separations with a temperature program or shorter column. A second problem sometimes is seen in crescent-shaped trails

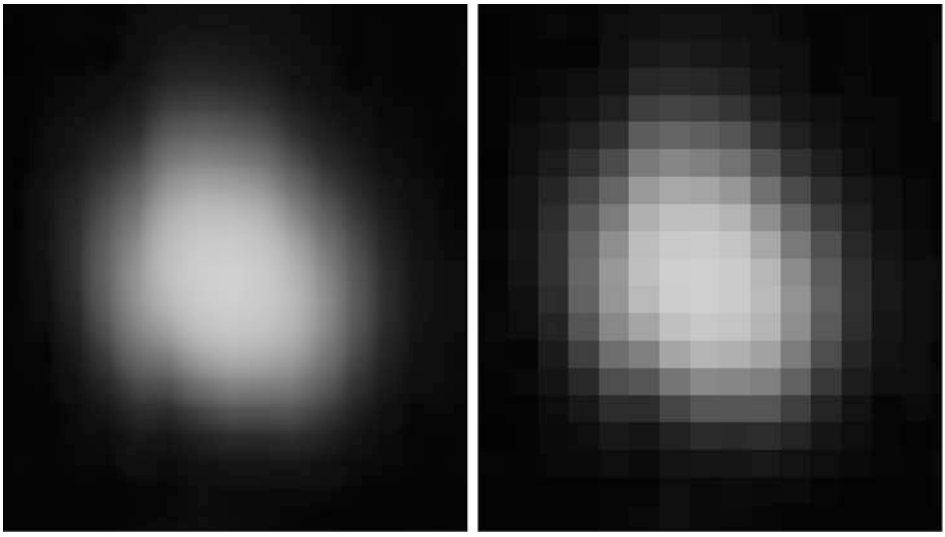

Figure 3 A single GC×GC peak enlarged by bilinear interpolation (left) and nearest-neighbor interpolation (right). Bilinear interpolation yields a truer (i.e., higher fidelity), more pleasing spot; but nearest-neighbor interpolation more clearly shows the individual data points.

that, from left-to-right, slope downward quickly at first and then level out. These artifacts indicate a continuous presentation of eluates from the first column into the second column, perhaps caused by incomplete bake-out (an unclean first column) or by incomplete modulation (i.e., a thermal modulator that is not heated sufficiently to fully release). A third problem seen in visualizations is peak tailing in the second-column separations, which can be caused by various chromatographic issues. Figure 1 illustrates small artifacts of crescent-shaped ''bleed'' and peak tailing. Data visualization enables quick inspection of the data for these and other qualitative issues.

3.2 Other visualizations

3.2.1 Three-dimensional visualizations

Three-dimensional visualizations use many of the same techniques as two-dimensional image visualizations, including rasterization, colorization, navigation, and reconstruction. A three-dimensional visualization is based on a surface, with the surface elevation relative to the base plane given by each pixel's value. The elevation scale can utilize a mapping function (e.g., linear, logarithmic, or exponential functions). Constructing and viewing an artificial surface utilizes many of the techniques of computer graphics. The surface can be rendered in various ways, for example, pseudocolorized at each pixel, colorized with a solid color and illuminated to provide shading, or built as a wire frame. Then, the surface is projected onto a two-dimensional viewing plane for display. A common projection is the perspective view from a single viewpoint. Additional navigation

operations enable the user to rotate the surface in space, in order to view the surface from different perspectives. Figure 2 illustrates a three-dimensional perspective view of a portion of the GC×GC data shown in Figure 1 with values shown as the third dimension (i.e., elevation), with log scaling.

With the added dimension of height, three-dimensional visualizations are better able to show quantitative relationships over a large dynamic range. However, in three-dimensional visualizations, points on the surface can be obscured, and there is no correspondence between the dimensions of the data and the axes of the display, so interactive operations such as point-and-click indexing are more difficult and problematic than with a two-dimensional image. In that sense, different visualizations are complementary, each with its own utilities.

3.2.2 One-dimensional visualizations

One-dimensional graphs are useful for various purposes, including showing slices or integrations of GC×GC data in a graphical format that is familiar to traditional chromatographers. For example, the values in different secondary chromatograms (or rows along the first-column separation) can be rendered as a graph and overlaid to show whether the profiles change over time and/or the results of peak detection in one dimension. Similarly, values in different spectral "channels" of a pixel column (or row) can be graphed and overlaid to show if the multispectral profiles reveal the presence of co-eluted peaks, as illustrated in Figure 4.

3.2.3 Text and tabular visualizations

Some information is best communicated in a text format. For example, the values of the two-dimensional data array can be shown directly as a table, in which each cell displays a numeric pixel value. Visualization features available in spreadsheets are useful for tabular text visualizations. For example, colorization of the text or textboxes can be useful for highlighting different features of the data, such as peak

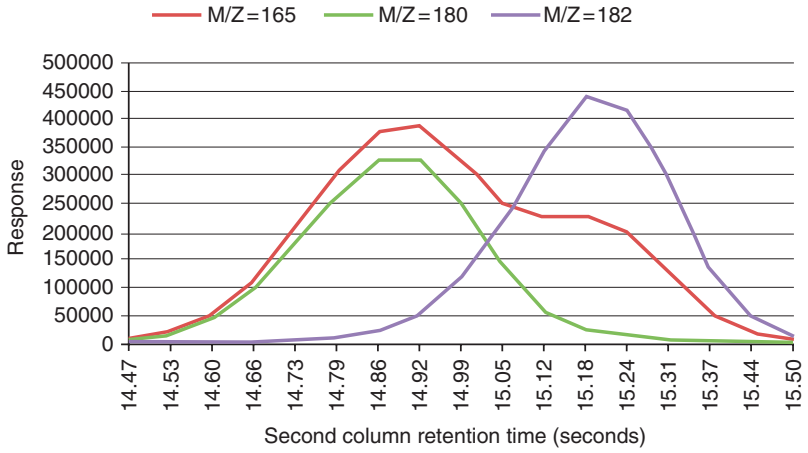

Figure 4 A one-dimensional visualization graphing values in selected-ion channels along a slice through co-eluted peaks.

	21.15	21.20	21.25	21.30	21.35	21.40	21.45
1.430	0.0571	0.1008	0.1730	0.5366	1.5977	0.5145	0.0056
1.425	0.0608	0.1814	0.2747	1.0831	3.1361	1.0161	0.0743
1.420	0.0428	0.3823	0.4438	2.1658	6.0174	1.9412	0.0978
1.415	0.0831	0.7240	0.9287	3.8826	11.2317	3.6697	0.2208
1.410	0.0841	1.3725	1.7997	6.8363	20.1688	6.6537	0.3567
1.405	0.1725	2.5766	3.4523	11.1887	34.9439	11.6780	0.6015
1.400	0.3449	4.7235	6.2445	17.6071	56.5288	19.6141	1.0032
1.395	0.7235	8.4680	11.2150	25.7758	85.6142	31.5504	1.6243
1.390	1.2746	14.2762	19.3033	35.3864	117.9551	47.0344	2.5674
1.385	2.0418	23.0246	31.9252	44.7295	147.8911	64.6653	3.7363
1.380	2.9983	34.3495	49.7217	52.4607	166.2080	80.5460	5.0810
1.375	4.0430	47.7378	71.9066	56.9784	167.6671	90.6433	6.2284
1.370	4.9709	60.2989	95.1407	57.8070	150.7534	91.0964	6.8776
1.365	5.4613	69.1711	113.9597	55.5665	120.4287	81.0968	6.7406
1.360	5.3638	70.7810	122.5308	50.6774	84.3015	63.1797	5.7978
1.355	4.6100	64.3888	117.6683	44.3232	50.4783	41.9836	4.3647
1.350	3.4788	50.9848	100.1199	35.8963	25.2809	23.4741	2.8337
1.345	2.3031	34.7295	74.5050	26.7257	10.3831	10.7412	1.6212
1.340	1.3587	19.7956	47.2256	17.3471	3.7661	4.2601	0.7790
1.335	0.7333	9.4821	24.8462	9.9335	1.2948	1.5094	0.3424
1.330	0.3598	3.8899	10.6588	4.8219	0.4114	0.5122	0.1174
1.325	0.1834	1.4643	3.8885	2.1327	0.0980	0.1950	0.0448
1.320	0.1014	0.5461	1.3180	0.8218	-0.0022	0.0817	0.0131

Figure 5 A tabular visualization of data values in two adjacent peaks with colorization to show primary peak membership.

membership, as shown in Figure 5. Statistical views of the data can be presented simply in a table, and other spreadsheet functions, such as sorting and averaging, are useful for quantitative analysis, which is the subject of the next section.

3.2.4 Graphical overlays and annotations

Graphical overlays are useful for communicating metadata — additional information about the data. For example, in Figure 6, semitransparent bubbles are used to indicate detected peaks. This analysis is for ASTM D5580 Standard Test Method for Determination of Benzene, Toluene, Ethylbenzene, p/m-Xylene, o-Xylene, C9 and Heavier Aromatics, and Total Aromatics in Finished Gasoline by Gas Chromatography [16], so bubbles are activated only for the peaks of interest. The areas of the bubbles are proportional to the peaks' total response, and the colors indicate the chemical group membership of the peak. (Peak detection and identification are described later.) Lines connecting peaks show associations with internal standards for quantitative calibration. Graphical shapes, such as polygons and polylines, are used to indicate chemical groups — in this

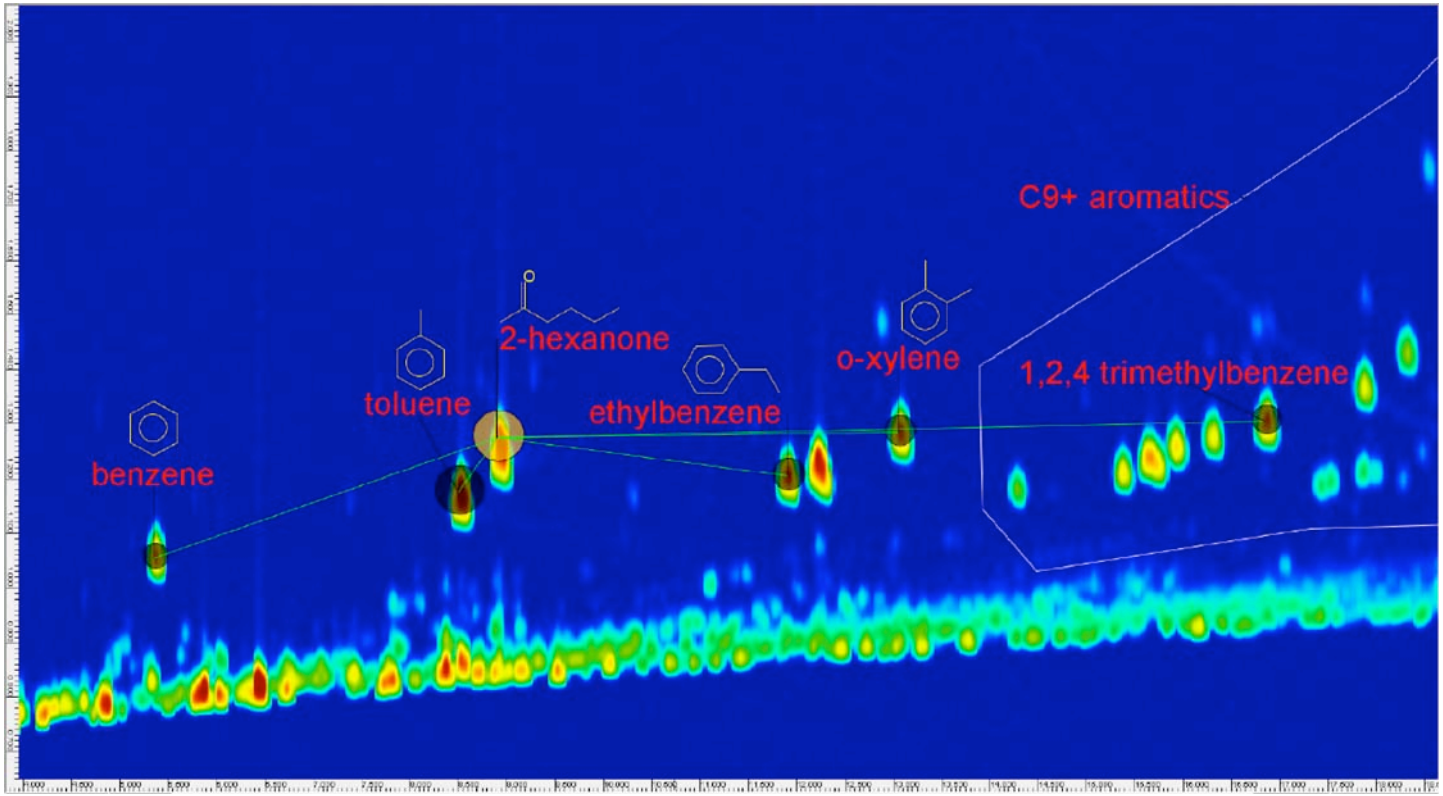

Figure 6 A graphical overlay with semitransparent bubbles for detected peaks of interest, a polygon to indicate the C9+ aromatics, text labels, and graphical chemical structures. A subregion of the data from Figure 1 is shown.

example, the C9+ aromatics. Text labels and chemical structure graphics communicate additional information.

4. DATA PROCESSING

Data processing extracts higher-level information from the raw data for further analysis. This section presents an overview of basic processing operations for GC×GC data:

1. Phase correction — synchronize the columns of data points with the modulation cycles.
2. Baseline correction — remove nonsignal baseline offsets in the data values.
3. Peak detection — detect signal peaks induced by separated compounds.

4.1 Phase correction

In rasterizing GC×GC data, it is typical that the starting data point of each secondary chromatogram in the image corresponds to the time that the modulator released its sample into the second chromatographic column. Then, the vertical axis of the image properly reflects the retention time in the second column. Typically, this is performed by the chromatographic system, but if the data acquisition is out-of-phase with the start of the modulation time, phase correction may be required.

Phase correction is the operation of shifting data in the image so that the data point acquired at the start of each modulation cycle (i.e., the start of each second-column separation) is the first pixel in each image column. (Other synchronizations, e.g., starting each column at the holdup time, are possible but less commonly used.) In the data itself, there may be no markers for the start of the modulation cycles, in which case corrective processing requires inference. (If there are such markers, phase correction is simple.) If the modulation and sampling frequencies are known accurately, then it is possible to accurately infer the first data point corresponding to the modulator release in every modulation cycle from the data point corresponding to the modulator release in just one modulation cycle by iteratively adding (or subtracting) the product of the modulation interval and the sampling rate.

For example, a modulation interval of 4 s and a sampling rate of 200 Hz mean the data point for the start of each modulation cycle follows 800 data points after the data point at the start of the previous modulation cycle. Suppose, in this example, the first data point of the first full modulation cycle is not the first pixel in the first image column but is instead the 400th pixel (i.e., in the middle of the first image column). Then, phase correction could be performed by dropping the first 400 pixels of the first image column, corresponding to the data points before the start of the first full modulation and shifting the data. So, given the modulation and sampling frequencies, it is sufficient to know the second-column retention time of any constituent compound and then to identify the peak pixel

for that compound, in order to establish a known mapping between data points and the modulation cycle. From that known point in the modulation cycle, the starting data points for every modulation cycle can be inferred and shifted accordingly.

If the required phase correction is not an integer, two options are possible: (1) round the phase correction to the nearest integer pixel index and accept a timing error of not more than one-half of the sample interval or (2) resample the data so that the resample point is precisely at the start of the modulation cycle. The first option typically is preferred because it maintains the original data, without introducing resampling errors, and is computationally simpler.

A similar issue exists if the product of the modulation interval and the sampling frequency is not an integer. In this situation, each pixel column may have a different fractional offset relative to the modulation start time. Then, the fractional phase correction varies among image columns, and so rounding may result in image columns with heights that differ by one pixel. For visualization, but not for subsequent analysis, this requires that a pixel be added to shorter rows (or that a pixel be excised from longer rows), for example, in data for the void time at the start of the separation.

4.2 Baseline correction

In gas chromatography, the signal peaks, induced by constituent compounds in the sample, rise above a baseline level in the output. Under controlled conditions, the baseline level consists primarily of the steady-state standing-current baseline of the detector and column-bleed (which may cause a progressive rise in temperature-programmed runs). Figure 7 illustrates a three-dimensional

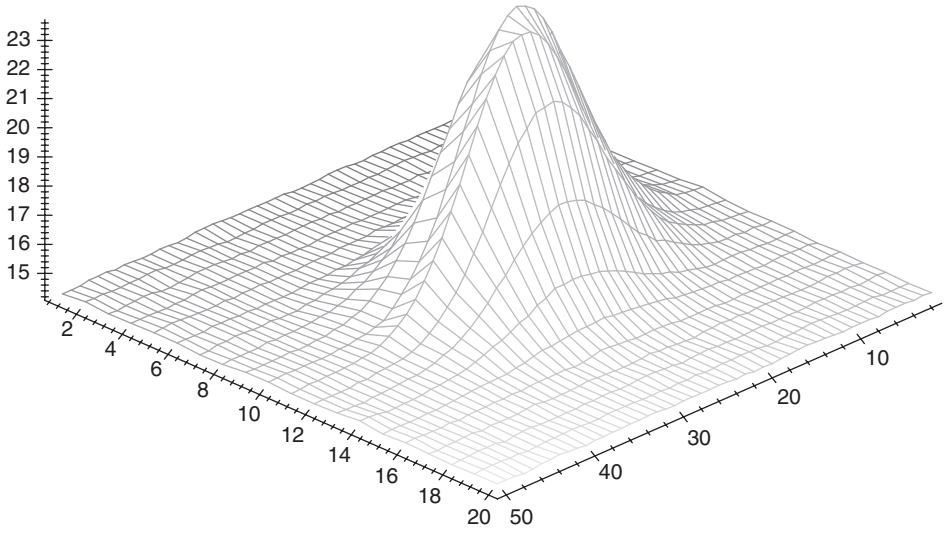

Figure 7 A GC×GC peak on a non-zero baseline.

perspective plot of an isolated peak rising to a maximum value of over 23 picoamps. However, the baseline in the region of the peak is more than 14 picoamps, so the actual maximum peak height induced by the sample compound is less than 10 picoamps. As this example makes clear, accurate quantification of the analyte peaks requires subtraction of the baseline level from the signal.

There are two general approaches for estimating the baseline for correction: (1) estimate the baseline around each peak separately and (2) estimate the baseline across the data comprehensively. The first approach requires that a data point value just outside a peak indicate the baseline level, but this is problematic in regions of the data that are crowded with peaks because the values just outside a peak may be acted upon by neighboring peaks. The second approach requires multiple data point values indicating the baseline level with sufficient frequency that the baseline can be reconstructed.

In GC×GC data, the baseline usually can be observed at many points, for example, during the void time of each second-column separation, even if other regions of the data are crowded with peaks. This is an important attribute of GC×GC for accurate quantification because if the baseline cannot be estimated, then peak integration is less accurate. Typically, the baseline does not change significantly over the brief time of a few modulation cycles, so these observations are sufficient to reconstruct the baseline in a comprehensive fashion.

In a simple model of the GC×GC process, each data point value produced by the system is the sum of:

- A nonnegative baseline offset value that is present even when there is no sample compound detected.
- The signal due to the presence of the detected sample compound(s).
- Random noise fluctuations (including digitization round-off).

Under typical controlled conditions, the baseline offset values change relatively slowly over time, and the signal and noise fluctuate more rapidly over time.

Reichenbach et al. [17] described a method for extracting the GC×GC baseline comprehensively. The first step identifies background regions (i.e., regions without analyte peaks) by locating data points with the smallest values in each second-column chromatogram (or other interval). Then, the local means of the values from data points in the background regions are taken as first estimates of the baseline, and the variances of the values are taken as first estimates of the variance of the noise distribution (which also is present in the background). Then, signal processing filters are used to reconstruct the baseline as a function of the local estimates. Finally, the baseline estimate is subtracted from the signal.

Figure 8 shows two examples of baseline correction: with a blank sample (top) and a diesel sample (bottom). On the left, images of the data before baseline correction are shown with a narrow grayscale range of 1.0 picoamp from black to white. As can be seen in both images, but especially the blank data, there is a temperature-induced increase in the baseline from left to right such that the baseline at the right is nearly 1.0 picoamp greater than the baseline at the left. On the right, images of the data after baseline correction are shown with an even narrower grayscale range of 0.1 picoamp from black to white centered about

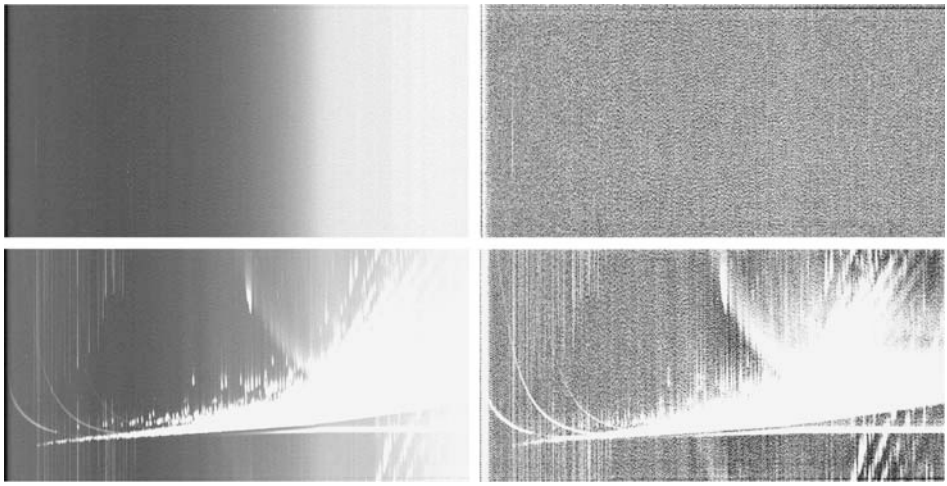

Figure 8 Baseline correction for a blank sample (top) and diesel sample (bottom), before baseline correction (left, with a grayscale range of 1.0 picoamp, 14.5 to 15.5) and after baseline correction (right, with a grayscale range of 0.1 picoamp, from −0.5 to 0.5).

0.0 picoamp. As can be seen in the images after baseline correction, the baseline is removed, and the remaining background values consist of near zero-mean noise with variance less than 0.1 picoamp. The baseline correction is successful not only for the blank run, but also for the diesel sample in which signal obscures much of the baseline.

For systems producing multichannel data, such as GC×GC–MS, the baseline can be estimated in each channel using the same method. Baseline correction for centroided multispectral data is difficult because the centroiding process removes many (or all) of the background values. Therefore, baseline correction should be performed before or at the same time as spectral centroiding (but, unfortunately, that is not always done).

4.3 Peak detection

Blob detection is the process of aggregating peaked clusters of pixels. The term *blob*, from the digital image processing literature, means a cluster of pixels that are brighter (or darker) than their surround. For GC×GC data, it is useful to distinguish blobs from analyte peaks, because a detected blob might be formed from several co-eluted analyte peaks, or a single analyte peak might be detected incorrectly as several blobs (e.g., due to false minima introduced by noise). After blob detection, peak detection may require unmixing blobs resulting from co-elution and merging blobs resulting from incorrectly split peaks.

Two alternative approaches for GC×GC blob detection are: (1) use traditional one-dimensional chromatographic peak detection along each second-column chromatogram and then form two-dimensional blobs from the unions of adjacent one-dimensional peaks [18,19] or (2) perform detection in both dimensions

simultaneously. The first approach, of relying on one-dimensional chromatographic peak detection, builds on an accepted methodology but does not fully utilize all available relevant information as it detects peaks in one dimension without reference to the other dimension. The second approach requires a two-dimensional algorithm but can use all available relevant information in each step of the detection.

The drain algorithm for two-dimensional blob detection in GC×GC data [20] is an inversion of the watershed algorithm [21]. The approach is a ''greedy'' dilation algorithm that proceeds by starting blobs at peak tops and iteratively adding smaller pixels bordering the blobs until there are no more smaller, positive-valued pixels in the surrounds. This process can be understood conceptually by picturing the image as a relief map with larger values having higher elevation (i.e., as a three-dimensional surface as in Figure 2). The surface is placed under enough ''water'' to submerge the highest elevation; then, the water is progressively ''drained.'' As the draining proceeds, peaks appear as ''islands'' and are distinguished with unique blob identification numbers. As more water drains, islands (blobs) expand as lower-lying pixels around the ''shore'' are exposed. When the water between two islands disappears, then a border between blobs is set. When the water level reaches zero, the process is stopped (as negative values are due to noise fluctuations below the baseline). In order to prevent noise from being detected as spurious peaks, blobs that are too small — either in number of data points, apex value, total blob intensity, and/or other criteria — can be ignored.

The example in Figure 9 illustrates the drain algorithm. The intensity of the data point is the base number (values up to 99), and the subscript indicates the order (1–12) in which the data points are added to a blob (dark gray for Blob 1 or light gray for Blob 2). In A, the data point with largest value, 99, starts Blob 1, and then the data points ordered by values 95, 88, and 80 are added to Blob 1 because they neighbor another data point previously assigned to Blob 1. In B, the data point with value 77 starts a new blob, Blob 2, because it is the next largest value and is not adjacent to a data point in any other blob. Then, the data point with value 72 is added to Blob 2. In C, the data points with values 63 and then 61 are added to Blob 2 and Blob 1, respectively, based on their adjacencies to previously assigned data points. In D, the data points with values 42, 38, and 34 are assigned, in order, to Blobs 1, 1, and 2. Where a data point is adjacent to more than one previously assigned data points, the data point is assigned to the same blob as its largest neighbor.

95_2	61	34	77
99_1	71	38	72
80_4	88_3	42	63

95_2	61	34	77_5
99_1	71	38	72_6
80_4	88_3	42	63

95_2	61_9	23	77_5
99_1	71_7	38	72_6
80_4	88_3	42	63_8

95_2	61_9	24_{12}	77_5
99_1	71_7	38_{11}	72_6
80_4	88_3	42_{10}	63_8

Figure 9 Data points, with intensity shown as the base number, are assigned in order of their intensity, with order shown as the subscript, to a blob (dark gray for Blob 1 or light gray for Blob 2). Snapshots of the assignment process are shown from left to right.

One type of error for any blob detection algorithm is oversegmentation — detection of multiple blobs that should be detected as a single peak. This problem can be caused by noise-induced false minima within a peak or other acquisition artifacts. Various approaches can be used to reduce or eliminate oversegmentation. For example, smoothing can be implemented before detection as a convolution with a two-dimensional Gaussian spot whose width is parameterized according to the variance of the noise: a wider blurring function is appropriate for more noise and a narrower blurring function is appropriate for less noise. Too little blurring does little to correct incorrectly split peaks, whereas too much blurring can cause the opposite problem of incorrectly merged peaks.

Undersegmentation, in which multiple analyte peaks are detected as one blob, occurs if analyte peaks are so close in time that there are no minima between them (or small minima are removed by smoothing). For example, a small co-eluted peak may appear as a shoulder on the larger peak. Even if there are minima between overlapping peaks, the watershed algorithm does not ''unmix'' the peaks; it simply delineates the minima between them. As described in Chapter 5, numerical methods may be used to unmix co-eluting peaks. For example, if each peak has a consistent shape with respect to every row and with respect to every column of the data, then unmixing can be seen as the task of inverting (or deconvolving) a separable, bilinear system for single-valued data or tri-linear model for multichannel data. However, the inversion problem is ill conditioned, and the peak shapes and data are subject to noise and other variables, so the unmixing problem is difficult. MS data can be especially useful for unmixing co-eluting peaks that have differing spectra. Even with MS data, unmixing nearly coincident peaks may require external information (e.g., the spectra of the coincident peaks).

Various chromatographic conditions can cause problems for peak detection algorithms. For example, if the temperature for the second-column separation changes rapidly relative to the modulation cycle, then the apexes of one-dimensional peaks in consecutive second-column separations of a single compound may be offset from one another. For a two-dimensional method such as the drain algorithm, the two modulations may be detected as two separate peaks if the shift is two or more samples. Similarly, a one-dimensional method may fail to join the two one-dimensional peaks. Smoothing, described above, may ameliorate this problem. Chromatographic solutions include more rapid modulations, a slower temperature program, and/or a slower sampling rate. As discussed in Section 2.1, long modulation cycles or slow sample rates relative to (respectively) the first-column and second-column peak widths yield narrower troughs between co-eluting (or nearly co-eluting) peaks, which can lead to undersegmentation as the separate peaks become more difficult to discern. In this case, chromatographic solutions include more rapid modulations, a slower temperature program, and/or a faster sampling rate.

After blobs are detected (or even as they are detected), important statistical features of the blobs can be computed. Most important for quantification, the integration or sum of all of a peak's intensity values is indicative of the relative amount of the compound inducing the peak (subject to the responsivity of

the detector to the compound). Geometrically, the integration under a two-dimensional peak is a volume (with two retention-time dimensions and the response dimension), analogous to the integration under a one-dimensional chromatographic peak as an area (with one retention-time dimension and the response dimension). Quantification that accounts for the detector responsivity is described in Section 6.1.

Many other statistics can be computed. The number of data points (or pixels) in the peak is a measure of its retention-time footprint or area, with two retention-time dimensions. Symmetry can be measured as a ratio of the tailing and leading half-widths in each dimension. Various measures with weighted and unweighted moments indicate the center of the peak, center of gravity, variance in each retention-time dimension, orientation, eccentricity, and so on. The GC Image Users' Guide [22] documents more than 70 GC×GC peak features. These features are useful in identifying unusual and possibly problematic blobs, for example, blobs resulting from co-eluting peaks or blobs resulting from split peaks, which then can be subject to visual inspection and interactive correction.

As experienced chromatographers know, automated peak detection is sometimes erroneous, especially for small peaks that are barely detectable amid noise and co-eluted peaks that are nearly coincident. So, interactive tools are useful, but even human experts may not be able to solve difficult peak detection problems. As described in the next section, complex features can be computed as the combination of elementary features for chemical identification.

5. CHEMICAL IDENTIFICATION

A common analytical goal is an assay with individual compounds or group identities and quantitative concentrations of target constituents. (Compounds belonging to the same chemical group are related to one another in some chemical or physical way.) Accurate quantification involves not only the peak responses, but also the responsivity of the detector because detectors may have differing quantitative responses to the same concentrations of different compounds. Therefore, analyte identification (described in this section) typically is performed before quantification (described in the next section). With single-valued GC×GC data, analyte identification must be based primarily on retention time. With multichannel data, such as from GC×GC–MS, multivariate methods can be used for chemical identification.

5.1 Chemical identification by retention time

A common method for chemical identification in one-dimensional chromatography is to define retention-time windows for peaks of interest. Under repeatable, reproducible, and tightly controlled chromatographic conditions, the peaks for target compounds will fall reliably within fixed retention-time windows. However, narrow windows may be required for peaks with nearby neighboring peaks (to avoid false positives), and, with narrow windows, even

slightly different chromatographic conditions may cause a peak to drift outside its window. Here, ''drift'' is used to characterize a local variation that may be related to more complex systemic variations as might be caused by differing column conditions, temperatures, gas pressure, etc.

Some standard one-dimensional GC methods use reference peaks to help recognize drift [23]. For more widely varying chromatographic conditions, retention times for targets can be related using a linear retention index (LRI) [24], in which retention times are referenced relative to the retention times of marker compounds. A common LRI scheme uses the n-alkanes as marker points with indices equal to 100 times the carbon number (following the Kováts index [24]); then the indices for peaks between marker points are computed using piecewise linear interpolation. If retention-time windows are defined relative to marker peaks that can be located, then any linear retention-time transformation observed in the marker peaks can be applied to the windows used for chemical identification.

Retention-time windows can be used in two dimensions, but the problems of drift exist in both dimensions, with drift in the first dimension possibly inducing drift in the second (related to the temperature program). In an intralaboratory study of GC×GC retention times across separate column sets, chromatographs, and days, Shellie et al. [25] demonstrated highly reproducible peak positions, but with statistically significant drift over separate days and other chromatographic conditions. Ni et al. [26] showed that peak pattern variations over widely varying chromatographic conditions could be modeled well by affine transformations (i.e., translation, scale, and shear). As illustrated in Chapter 3, several approaches have been put forward for two-dimensional indexing [27–31], but none has yet achieved wide acceptance and research continues. A robust approach for dealing with two-dimensional retention-time transformations that can be tailored to specific applications is to locate and identify target peaks relative to the positions of many other peaks in the sample, not just a few standard markers. With this approach, the transformation observed in the pattern of many peaks can be applied to the windows for chemical identification.

Template matching is a powerful extension of the traditional approaches of reference and marker peaks to identify compounds by recognizing peaks in multidimensional separations subject to multidimensional retention-time transformations. A template records the pattern of peaks expected for an analysis, along with information for chemical identification, such as the compound name and/or chemical group for peaks of interest. A template can be built from prototypical data either automatically with all peaks meeting specified criteria (e.g., the largest peaks) or interactively with selected peaks. Templates can be constructed based on peak retention times in one chromatogram or based on averaged peak retention times in several chromatograms. Then, given a template and the set of peaks observed in a sample for analysis, peak pattern matching finds a subset of peaks in the sample data that forms the same pattern as the template. A template-matching algorithm establishes as many correspondences as possible between peaks in the template and peaks in the sample data subject to the allowed retention-time transformation (e.g., shifting or scaling the template) and the allowed retention-time window [32–37]. After peak correspondences are

established, the annotated information (such as compound name or group) from peaks in the template is copied into corresponding peaks in the data. Consequently, all the matched compounds in the template are identified in the data.

Figure 10 illustrates a template constructed from the gasoline analysis in Figure 6. The template from the gasoline analysis is overlaid and matched to the chromatogram of a diesel analysis acquired four years later with a different chromatograph and different columns. This template is a multitype template that contains not only a pattern of expected peaks, shown with open circles, but also other information for annotating and reporting on the data. Polygons define regions in which peaks for chemical groups are expected. Text and chemical structure objects are included to provide annotations for visualizations. Graphical lines are used to visually highlight associations between compounds and the internal standards used for calibration. (However, the internal standard, 2-hexanone, is not present in the diesel sample.) The locations of the matched peaks in the diesel chromatogram are shown with filled circles connected by lines to the nearby, corresponding template peaks (shown with open circles). As can be seen, template matching is an effective method for quickly identifying peaks and chemical groups. Other objects in the template are geometrically transformed according to the transformation of the matched peaks, as can be seen for the shifted polygon and its label. Any errors in template matching can be corrected interactively.

5.2 Multivariate methods for chemical identification

Methods for identifying chemical compounds by multichannel data signatures (such as searching a MS library for a matching multispectral signature) are essentially the same for GC×GC as for GC, but GC×GC, with its superior separation power, can significantly reduce co-elution and so improve the accuracy of chemical identification. With multichannel detectors, different compounds have different multivariate signatures (although signatures of similar compounds can be quite similar). The signatures of unidentified peaks can be compared to the known signatures of compounds of interest, with a mathematical computation of difference or similarity between signatures, to find a match that identifies the compound. The National Institute of Standards and Technology (NIST) distributes a library of MS signatures for more than 163 K compounds and a program for searching the library [38]. This approach can be highly effective for chemical identification, but there are many issues that can cause misidentifications, for example, the unknown compound may not be documented in the library, observed signatures are variable, co-elution mixes signatures. In the presence of variability, co-elution, and noise, the search program may find the wrong match. GC×GC can greatly reduce co-elutions, thereby producing purer signatures that can be better identified.

Rule-based methods follow another approach for chemical identification with multichannel data. Experienced analytical chemists often use rules to deduce chemical identity [39,40]. In a computer-based system, rules express the reasons or criteria for chemical identification. Welthagen et al. [41] used a rule-based

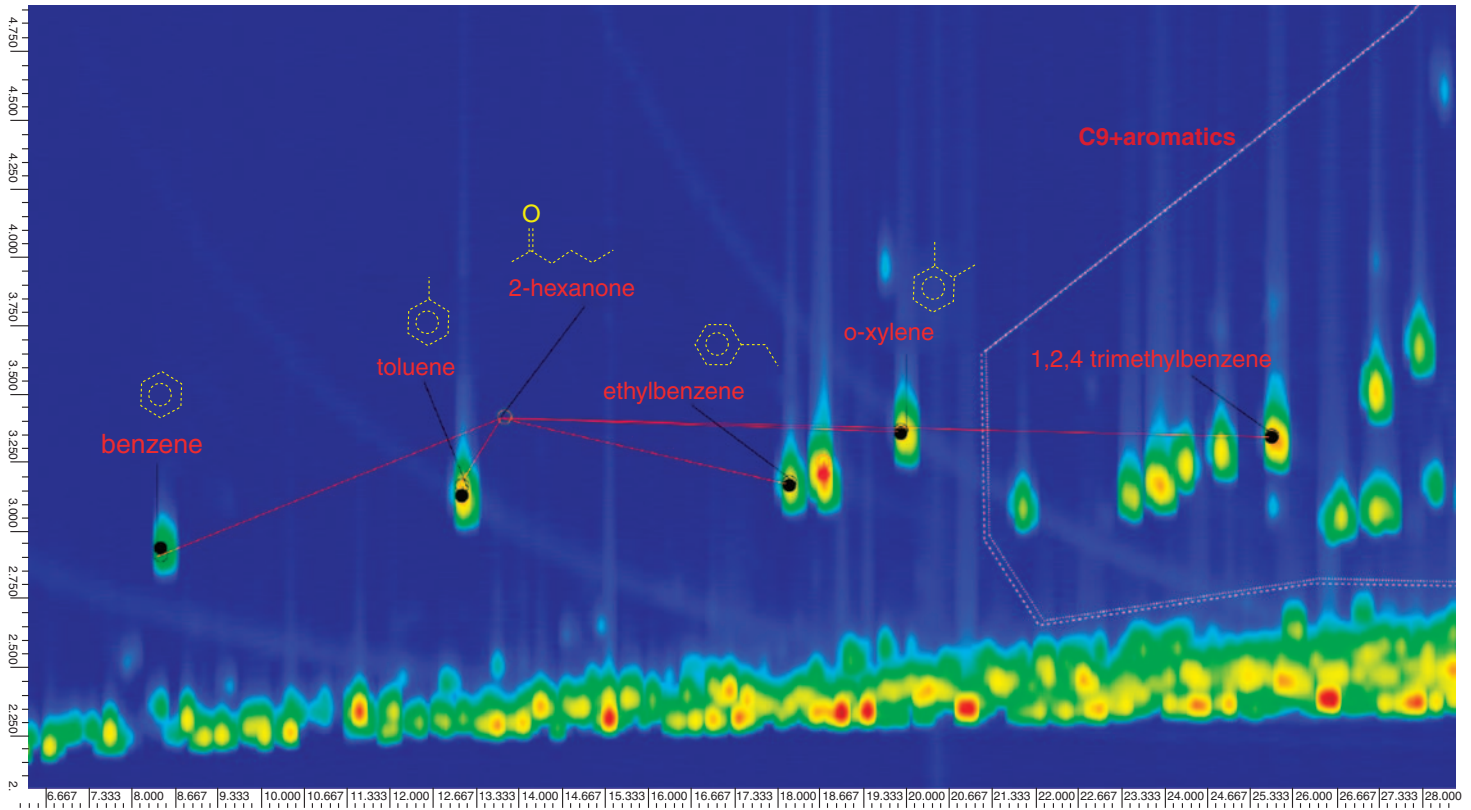

Figure 10 A template from a gasoline analysis is overlaid and matched to peaks in a diesel sample, thereby identifying the peaks and groups of interest. A subregion of the data is shown with open circles, showing the expected peak positions in the template.

approach based on GC×GC retention times and MS signatures to classify chemical groups in the analysis of airborne particulate matter. For example, the rule for polar benzenes with or without alkyl groups in the GC×GC–MS data was:

> The MS intensity at mass-to-charge ratio (m/z) 77 is greater than 25% of the intensity of the MS base peak (i.e., the largest MS intensity), and the retention time in the second column is greater than 2 s.

The Computer Language for Identifying Chemicals (CLIC) [42] defines a grammar for expressing rules for chemical identification based on multidimensional retention times and spectral characteristics, including library search. CLIC offers functions of multidimensional chromatographic retention times, functions of MS characteristics (such as selected-ion intensity counts), functions for MS library search, numbers for quantitative and relational evaluation, and logical and arithmetic operators. The CLIC expression for the above rule for identifying polar benzenes is:

$$(\text{Relative}(77) > 25)\ \&\ (\text{Retention}(2) > 2).$$

This rule can be applied to the spectra of all peaks to determine which are polar benzenes. Even more complicated rules involving selected-ion intensity counts can be derived using classifiers [43,44], and other features can be applied with GC×GC [45].

Rule-based identification works well for multispectral constraints but is less convenient for retention-time constraints (e.g., describing a many-sided polygon to restrict the retention times for a group of peaks in a chemical class). Complex retention-time rules can be more easily expressed graphically, for example, in templates. Rule-based constraints and templates have complementary strengths that can be combined for highly effective chemical identification (as described next).

5.3 Smart Templates

Smart Templates [46] combine retention-time templates with rule-based chemical constraints. Templates express retention-time patterns in a convenient graphical form that is highly visual; CLIC expressions efficiently define rules with an arsenal of functions, constants, and mathematical and logical operators. In complex chromatographic regions, if template matching finds several peaks in the data that are candidates to match a template peak, then a rule associated with the template peak can eliminate incorrect matches. (In this case, the CLIC expression can be applied only to the peaks that are potential matches.) Similarly, if a spectral rule to identify peaks in a chemical group identifies peaks with too widely ranging retention times, then a template polygon with the associated rule can restrict group identification using both the rule and convenient graphical retention-time constraints. The combination is a powerful methodology for chemical identification.

Ledford [47] developed a Smart Template for analysis of diesel aromatics in experimental work on a standard analytical method under development for evaluation by the ASTM. (Vogt et al. [45] developed a similar analysis.) Figure 11 shows an example group analysis with Ledford's Smart Template. The Smart Template uses a retention-time polygon with CLIC expression on the GC×GC–MS data for each chemical group, with distinctions for both chemical classes and carbon number. The result is a descriptive group analysis.

6. QUANTIFICATION AND MULTI-DATASET ANALYSES

Several important analytical problems involve multiple datasets.

- *Sample quantification*: calibrate for quantification by measuring detector responses to different levels of concentrations in multiple chromatograms.
- *Sample comparison*: characterize similarities and differences between datasets, for example, to find anomalies such as might be responsible for a desirable or undesirable trait.
- *Sample classification*: use many GC×GC datasets to characterize sample classes based on within-class commonalities and between-class differences and then classify a sample into one of the classes based on GC×GC analysis.
- *Sample recognition*: establish the identity of a sample's source by pattern recognition comparing a GC×GC dataset against many GC×GC datasets stored in a library to find the best match. This is sometimes referred to as chemical fingerprinting.
- *Sample query*: find a dataset(s) that have specified characteristics from among a repository of many datasets. Other standard database operations, such as insert and delete, are useful for maintaining and using repositories.

6.1 Quantification

After the detector responses for a peak have been integrated, accurate quantification requires consideration of the detector's responsivity to the compound inducing the peak. In this, calibration and quantification of GC×GC peak responses are performed with the same approaches as for GC (including internal calibration, external calibration, and response factors), but research surveys document that the quantitative performance of GC×GC is superior to that for one-dimensional GC [48–50]. In an early report of quantitative performance for GC×GC, Gaines et al. [51] reported two- to fourfold improvements in limits-of-detection for trace oxygenate and aromatic compounds with FID. Lee et al. [52] observed a four- to fivefold increase in sensitivity for GC×GC with FID, which was consistent with their model predicting both peak response enhancement of roughly 20-fold from peak focusing and increased noise associated with faster sampling rates. Other researchers reported detectability improvements of two- to fivefold for GC×GC–MS [53] and GC×GC–ECD [54]. Of course, the greatest benefit of GC×GC for quantification frequently is greater selectivity, which allows quantification of compounds that otherwise would be co-eluted and difficult to quantify accurately.

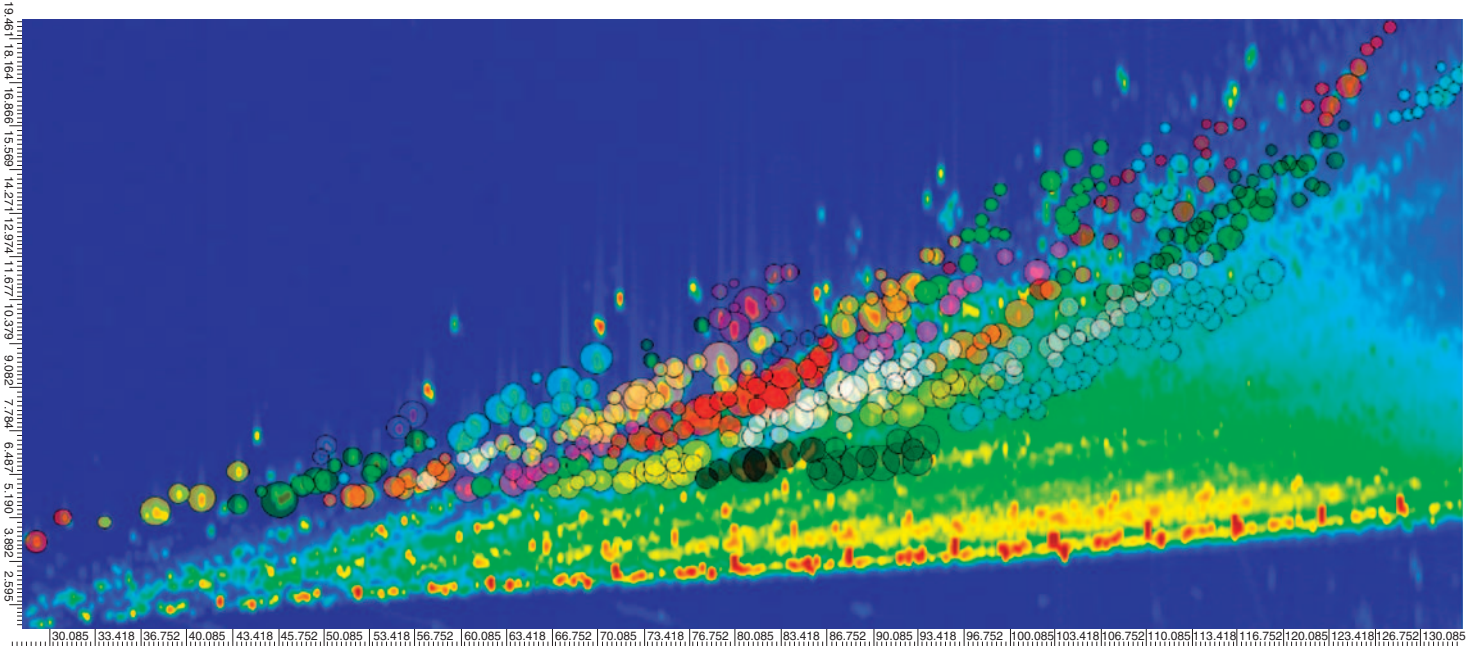

Figure 11 Group analysis of diesel aromatics by a Smart Template [47].

As of 2008, despite more than a decade of research demonstrating the increased selectivity and sensitivity of GC×GC, there are no standard GC×GC methods. One reason may be that GC×GC can be applied to standard GC methods to provide improved performance in a wide range of analyses, as described in the applications-oriented chapters of this book. Another possible reason is that prior to the availability of commercial GC×GC software in 2003, quantitative analysis was laborious and time consuming. Perhaps another reason is that GC×GC opens so many options for new method development that settling on specifics is more difficult and has required a period of research and development of technologies and methods for standardization.

6.2 Sample comparison, classification, and recognition

The first level of intersample comparison is qualitative visualization and tabular comparison of sample constituents. Hollingsworth et al. [55] described various approaches for such comparisons. The visualizations begin with registering (aligning) the chromatograms to minimize the mean-square difference between peak retention times and normalizing the intensities with respect to a standard peak or set of peaks. Methods for visualization include flicker between images (i.e., cycle from one image to another) and display combination images (subtraction, ratio, addition) with grayscale or pseudocolorization. A method for ''fuzzy differences'' adjusts the difference image for residual differences due to peak shape and/or misregistration. Tables can be used to report quantitative differences. Frysinger and Gaines [56] used flicker visualization to find differences between regular and super gasoline for forensic analysis of fire debris. To track an oil spill, Nelson et al. [57] used difference, ratio, and addition images to show chemical changes over time. In Figure 12, the upper visualization shows the arithmetic difference (after registration and normalization) for samples in May and November 2003, and the lower visualization shows the color addition with the May data in green and the November data in red. The color addition image shows not only the magnitudes of the peaks (with intensity), but also the degree of change — from near complete weathering (indicated by the color green) of the n-alkane peaks along the bottom and the more volatile aromatics in the left half of the image to almost no weathering (indicated by the color yellow) of the less volatile aromatics. Their qualitative and quantitative analyses of peak intensity differences showed the differing effects of evaporation, water washing, and biodegradation on different compounds over time.

The classification of samples is another important analytical problem. For example, the search for biomarkers in metabolomic and proteomic research has the goal of finding sample characteristics indicative of a disease state or other biological condition. When samples are reduced to peak sets, the GC×GC classification problem is not significantly different from classification with GC, but the selectivity of GC×GC can be critical for classification accuracy. Frysinger and Gaines [40] demonstrated the utility of GC×GC for separating known biomarkers in crude oil. Shellie et al. [58] analyzed derivatized tissue samples from two classes of mice, obese and lean, and identified the 10 most likely biomarkers in the data

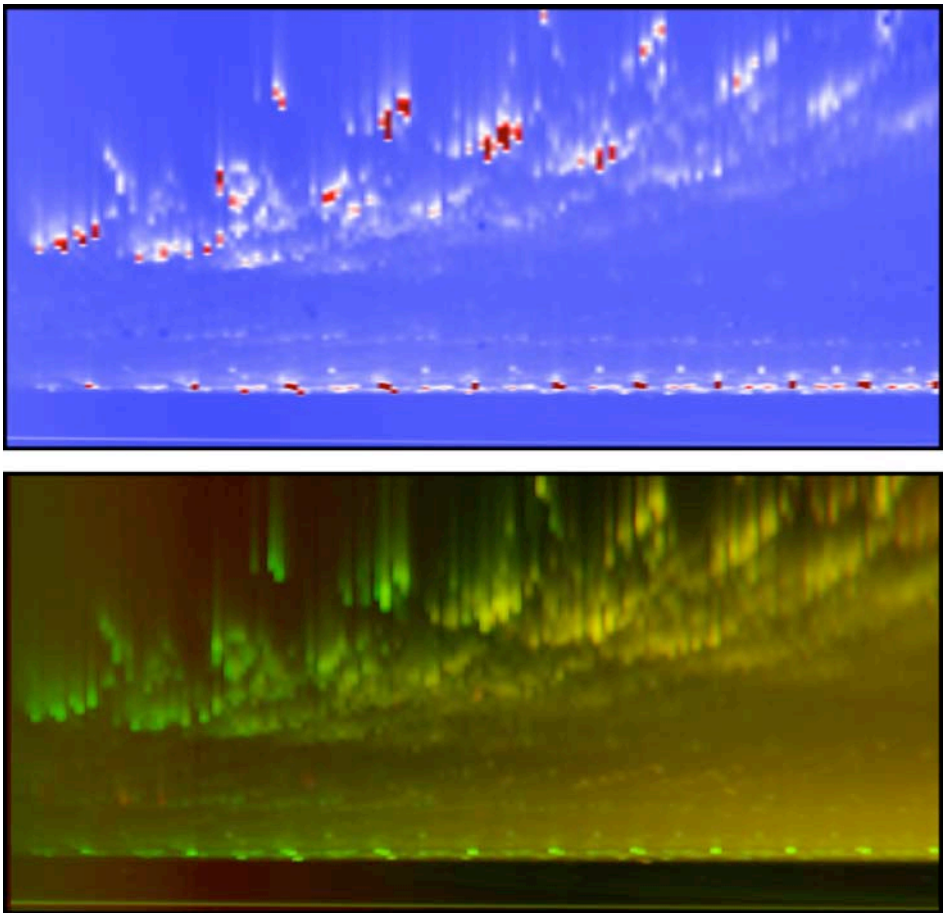

Figure 12 Comparison of oil spill samples in a difference image (between samples in May and November 2003) and a color addition image with the May sample in green and the November sample in red [57].

using *t*-test values for the spectra of deconvolved peaks. To classify yeast samples grown under either fermenting or respiring conditions, Mohler et al. [59,60] used multivariate methods to identify chromatographic regions with significant interclass differences prior to peak detection. In [59], principal component analysis (PCA) was applied to normalized selected-ion chromatograms to identify regions for peak detection with deconvolution. In [60], they identified chromatographic regions of interest by totaling the mean-signal-weighted Fisher ratio at each point in each spectral channel. Regions of interest were deconvolved, and the detected peaks were evaluated by the *t*-test. Others have used analysis of variance (ANOVA) methods to select chromatographic features for GC×GC classification [61,62]. These methods are discussed in detail in Chapter 5.

Fingerprinting focuses the classification problem to recognize one of multiple individuals (i.e., classes of size one). Gaines et al. [63] used GC×GC FID

fingerprints to identify which of two potential sources was responsible for an oil spill. They used fingerprint features from four chemical groups: naphthalenes, anthracenes/phenanthrenes, alkenes and cycloalkenes, and aliphatics. In each group, they utilized three minutiae (points of interest), each computed as the response ratio of a specific analyte peak within the chemical group to a fourth peak in the chemical group. The fingerprints provided convincing evidence for source identification. Investigating the problem of classifying crude oils by their source reservoir, van Mispelaar et al. [64] did not find any individual chemical markers sufficient for classification, but successfully classified samples based on small differences in many peaks using principal-component discriminant analysis (PCDA).

6.3 Databases and information systems

Software for higher-level database and information queries for GC×GC datasets would be highly useful but have not yet been fully developed. Database systems could support content-based data and information retrieval, for example, list the datasets for which the ratio of Chemical A to Chemical B is greater than x. Such queries could support fingerprint identification [63] on large databases. Information systems could support higher-level queries, for example, to support automated classification based on statistically significant peak-to-peak variations between two groups of datasets. Such queries could support the type of classification Shellie et al. [58] used to chemically distinguish obese and lean mice from tissue samples. Such systems would be useful not only for applications but for quality control, for example, finding differences in datasets of standard runs over time.

7. CONCLUSION

Many of the initial challenges for GC×GC data acquisition, visualization, and analysis have been surmounted, and solutions are available in commercial GC×GC software. Available software supports the following basic operations:

- Reading data from file formats produced by chromatographic systems.
- Displaying data in various modes, for example, as two-dimensional images, as projections of three-dimensional surfaces, as one-dimensional profiles, and so on.
- Preprocessing data to remove acquisition artifacts, such as modulation phase shift and signal baseline.
- Peak detection, including deconvolution/unmixing co-eluted peaks.
- Chemical identification using both retention-time and spectral data.
- Chemical quantification using the same approaches as for GC analysis.
- Multi-dataset analyses such as qualitative and quantitative comparisons.

Some problems require further research and development, notably:

- A standard file format for GC×GC data.
- More effective tools for chromatographic-spectral visualizations and multi-dataset visualization.

- Deconvolution/unmixing of difficult co-elutions.
- Multi-dataset analyses for classification and fingerprinting.
- Advanced queries for GC×GC databases.

These and other challenges are the subjects of ongoing research and development.

ACKNOWLEDGMENT

This chapter is based upon work supported by the National Science Foundation Division of Information and Intelligent Systems under Grant No. IIS-0431119. Zoex Corporation supported this work with example data and GC Image supported this work with GC×GC software.

REFERENCES

1 GC Image, LLC, GC Image® software, http://www.gcimage.com (2008).
2 J.V. Seeley, *J. Chromatogr. A*, 962 (2002) 21.
3 P.C. Kelly and G. Horlick, *Anal. Chem.*, 45(3) (1973) 518.
4 K.A. Duell, J.P. Avery, K.L. Rowlen and J.W. Birks, *Anal. Chem.*, 63(1) (1991) 73.
5 R.E. Murphy, M.R. Schure and J.P. Foley, *Anal. Chem.*, 70(8) (1998) 1585.
6 Agilent, Mini-IRS for: IQ Data File Format, Agilent Technologies, revised July 7, 1995.
7 B. Allen, M. Allen, R. Bingham, G. Carpanese, D. Gedcke, S. Haywood, G. Jackson, and J. Peck, FASTFLIGHT™, a Digital Signal Averager for Continuous High-Speed Data Acquisition with Electrospray Time-of-Flight Mass Spectrometers Coupled to Chromatographs, ORTEC, 1998.
8 ASTM International, Standard specification for analytical data interchange protocol for chromatographic data, Tech. Rep. E1947-98, ASTM, West Conshohocken, PA (1998).
9 ASTM International, Standard specification for analytical data interchange protocol for mass spectrometric data, Tech. Rep. E2077-00, ASTM, West Conshohocken, PA (2000).
10 Unidata Program Center, University Corporation for Atmospheric Research, NetCDF (network Common Data Form), http://www.unidata.ucar.edu/software/netcdf/ (2008).
11 P. J. Linstrom, A proposed data model for ASTM E13.15, http://animl.sourceforge.net/Linstrom-Data-Model-old.pdf (2003).
12 M. Fiege, T. Davies, T. Fröhlich, and P. Lampen, The AnIML core, sample, and technique shells: Proposal for an AnIML schema, http://animl.sourceforge.net/CLC Waters AnIML Proposal.pdf (2004).
13 T. Bray, J. Paoli, C. Sperberg-McQueen, and E. Maler, Extensible Markup Language (XML) 1.0, World Wide Web Consortium (2000).
14 R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice-Hall, Englewood, Cliffs, NJ, 2008.
15 A. Visvanathan, S.E. Reichenbach and Q. Tao, *J. Electron. Imaging*, 16(3) (2007) 033004.
16 ASTM, Standard test method for boiling range distribution of petroleum fractions by gas chromatography, Tech. Rep. D5580-02, ASTM, West Conshohocken, PA (2007).
17 S.E. Reichenbach, M. Ni, D. Zhang and E.B. Ledford, Jr., *J. Chromatogr. A*, 985(1) (2003) 47.
18 J. Beens, H. Boelens, R. Tijssen and J. Blomberg, *J. High Resolut. Chromatogr.*, 21(1) (1998) 47.
19 S. Peters, G. Vivó-Truyols, P. Marriott and P. Schoenmakers, *J. Chromatogr. A*, 1156(1–2) (2007) 14.
20 S.E. Reichenbach, M. Ni, V. Kottapalli and A. Visvanathan, *Chemom. Intell. Lab. Syst.*, 71(2) (2004) 107.
21 S. Beucher and C. Lantuejoul, In: *International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation*, 1979, pp. 17–21.
22 GC Image, LLC, GC Image™ Users' Guide, http://www.gcimage.com/usersguide (2008).
23 J.L. Snyder, In: R.L. Grob and E.F. Barry (Eds.), *Modern Practice of Gas Chromatography*, John Wiley and Sons, New York, 2004, pp. 769–882.
24 IUPAC, Compendium of Chemical Terminology, http://goldbook.iupac.org (2007).
25 R. Shellie, L.-L. Xie and P. Marriott, *J. Chromatogr. A*, 968 (2002) 161.

26 M. Ni, S.E. Reichenbach, A. Visvanathan, J.R. TerMaat and E.B. Ledford, Jr., *J. Chromatogr. A*, 1086(1–2) (2005) 165.
27 R.J. Western and P.J. Marriott, *J. Sep. Sci.*, 25(13) (2002) 832.
28 R.J. Western and P.J. Marriott, *J. Chromatogr. A*, 1019(1–2) (2003) 3.
29 J. Arey, R. Nelson, L. Xu and C. Reddy, *Anal. Chem.*, 77(3) (2005) 7172.
30 S. Bieri and P.J. Marriott, *Anal. Chem.*, 78(23) (2006) 8089.
31 J.V. Seeley and S.K. Seeley, *J. Chromatogr. A*, 1172 (2007) 72.
32 M. Ni, Point pattern matching and its application in GC×GC, Ph.D. thesis, University of Nebraska (2004).
33 M. Ni and S.E. Reichenbach, In: *IEEE Workshop on Statistical Signal Processing*, 2003, pp. 369–372.
34 M. Ni, Q. Tao and S.E. Reichenbach, In: *IEEE Workshop on Statistical Signal Processing*, 2003, pp. 497–500.
35 M. Ni and S. E. Reichenbach, In: Automatic Target Recognition XIV, Proc. SPIE 5426, 2004, pp. 155–163.
36 M. Ni and S. E. Reichenbach, In: Visual Information Processing, Proc. SPIE 5438, 2004, pp. 101–110.
37 M. Ni and S. E. Reichenbach, In: *Proceedings of the International Conference on Pattern Recognition*, Vol. 2, IAPR/IEEECS, 2004, pp. 145–148.
38 NIST/EPA/NIH Mass Spectral Library with Search Program, NIST Standard Reference Database 1A (2005).
39 F.W. McLafferty, *Interpretation of Mass Spectra*, 4 th Edition, University Science Books, Herndon, VA, 1996.
40 G.S. Frysinger and R.B. Gaines, *J. Sep. Sci.*, 24(2) (2001) 87.
41 W. Welthagen, J. Schnelle-Kreis and R. Zimmermann, *J. Chromatogr. A*, 1019 (2003) 233.
42 S.E. Reichenbach, V. Kottapalli, M. Ni and A. Visvanathan, *J. Chromatogr. A*, 1071(1–2) (2004) 263.
43 H. Lohninger and K. Varmuza, *Anal. Chem.*, 59(2) (1987) 236.
44 K. Varmuza and W. Werther, *J. Chem. Inf. Comput. Sci.*, 36(2) (1996) 323.
45 L. Vogt, T. Gröger and R. Zimmermann, *J. Chromatogr. A*, 1150(1–2) (2007) 2.
46 S.E. Reichenbach, P. Carr, D. Stoll and Q. Tao, *J. Chromatogr. A*, 1216(16) (2009) 3458.
47 S.E. Reichenbach, S.B. Cabanban, E.B. Ledford, H.A. Pham, W.E. Rathbun, Q. Tao, and H. Wang, In: *Pittcon Conference and Expo*, Chicago, IL, 2009, p. 540.
48 J. Dallüge, J. Beens and U.A.Th. Brinkman, *J. Chromatogr. A*, 1000 (2003) 69.
49 M. Adahchour, J. Beens, R. Vreuls and U.A.Th. Brinkman, *Trends Anal. Chem.*, 25(6) (2006) 540.
50 O. Amador-Muñoz and P.J. Marriott, *J. Chromatogr. A*, 1184(1–2) (2007) 323.
51 R.B. Gaines, E.B. Ledford, Jr. and J.D. Stuart, *J. Microcol. Sep.*, 10(7) (1998) 597.
52 A.L. Lee, K.D. Bartle and A.C. Lewis, *Anal. Chem.*, 73(6) (2001) 1330.
53 J. Dallüge, R. Vreuls, J. Beens and U.A.Th. Brinkman, *J. Sep. Sci.*, 25(4) (2002) 201.
54 P. Korytár, P. Leonards, J. de Boer and U.A.Th. Brinkman, *J. Chromatogr. A*, 958(1–2) (2002) 203.
55 B.V. Hollingsworth, S.E. Reichenbach and Q. Tao, *J. Chromatogr. A*, 1105(1–2) (2006) 51.
56 G. Frysinger and R. Gaines, *J. Forensic Sci.*, 47(3) (2002) 471.
57 R.K. Nelson, B.S. Kile, D.L. Plata, S.P. Sylva, L. Xu, C.M. Reddy, R.B. Gaines, G.S. Frysinger and S.E. Reichenbach, *Environ. Forensics*, 7(1) (2005) 33.
58 R.A. Shellie, W. Welthagen, J. Zrostliková, J. Spranger, M. Ristowe, O. Fiehn and R. Zimmermann, *J. Chromatogr. A*, 1086(1–2) (2005) 83.
59 R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young and R.E. Synovec, *Anal. Chem.*, 78(8) (2006) 2700.
60 R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young and R.E. Synovec, *Analyst*, 132 (2007) 756.
61 K.J. Johnson and R.E. Synovec, *Chemom. Intell. Lab. Syst.*, 60(1–2) (2002) 225.
62 M. Kallio, T. Hyötyläinen, M. Lehtonen, M. Jussila, K. Hartonen, M. Shimmo and M. Riekkola, *J. Chromatogr. A*, 1019 (2003) 251.
63 R.B. Gaines, G.S. Frysinger, C.M. Reddy and R.K. Nelson, Identification. In: S.S.Z. Wang (Ed.), *Oil Spill Environmental Forensics: Fingerprinting and Source Identification*, Academic Press, Burlington, MA, 2007, p. 169.
64 V. van Mispelaar, A. Smilde, J. Blomberg and P. Schoenmakers, *J. Chromatogr. A*, 1096 (2005) 156.