Talanta 115 (2013) 887–895

Contents lists available at SciVerse ScienceDirect

Talanta

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

Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry data

Luke C. Marney [a], W. Christopher Siegler [a,1], Brendon A. Parsons [a], Jamin C. Hoggard [a], Bob W. Wright [b], Robert E. Synovec [a,*]

[a] Department of Chemistry, University of Washington, P.O. Box 351700, Seattle 98198, WA, USA
[b] Pacific Northwest National Laboratory, Battelle Boulevard, P.O. Box 999, Richland 99352, WA, USA

ARTICLE INFO

Article history:
Received 13 March 2013
Received in revised form
18 June 2013
Accepted 21 June 2013
Available online 28 June 2013

Keywords:
Fisher ratio
Comprehensive two-dimensional gas chromatography
Time-of-flight mass spectrometry
Chemometrics
Feature selection

ABSTRACT

Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC–TOFMS) is a highly capable instrumental platform that produces complex and information-rich multi-dimensional chemical data. The data can be initially overwhelming, especially when many samples (of various sample classes) are analyzed with multiple injections for each sample. Thus, the data must be analyzed in such a way as to extract the most meaningful information. The pixel-based and peak table-based Fisher ratio algorithmic approaches have been used successfully in the past to reduce the multi-dimensional data down to those chemical compounds that are changing between the sample classes relative to those that are not changing (i.e., chemical feature selection). We report on the initial development of a computationally fast novel tile-based Fisher-ratio software that addresses the challenges due to 2D retention time misalignment without explicitly aligning the data, which is often a shortcoming for both pixel-based and peak table-based algorithmic approaches. Concurrently, the tile-based Fisher-ratio algorithm significantly improves the sensitivity contrast of true positives against a background of potential false positives and noise. In this study, eight compounds, plus one internal standard, were spiked into diesel at various concentrations. The tile-based F-ratio algorithmic approach was able to “discover” all spiked analytes, within the complex diesel sample matrix with thousands of potential false positives, in each possible concentration comparison, even at the lowest absolute spiked analyte concentration ratio of 1.06, the ratio between the concentrations in the spiked diesel sample to the native concentration in diesel.

1. Introduction

Multi-dimensional chromatographic instrumentation produces information-rich, and chemically complex multi-dimensional data containing meaningful chemical signals, often buried in a background of less meaningful chemical signal and noise. Experiments can be designed to analyze the similarities and differences between multiple injections of different samples, producing a data set of even higher dimensionality. With the aid of computer software, scientists need to be able to quickly, easily and comprehensively analyze multi-dimensional data sets, so important analytes and/or chemical fingerprints can be gleaned during discovery-based experimentation. Comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC × GC–TOFMS) is a prominent multi-dimensional separation technique that has been used extensively for discovery-based experimentation, especially when chemical species of interest are sufficiently volatile or amenable to derivatization [1–8]. To address the challenges, chemometric software for analyzing GC × GC–TOFMS data, as well as other multi-dimensional separation techniques, are available and continue to be developed [9,10].

For discovery-based experimentation with GC × GC–TOFMS, the 2D misalignment of peaks across different samples makes non-targeted analysis difficult [11]. Some alignment algorithms have been developed for “point by point” pixel-level data [11–14], while others have been developed for peak table-based data. Current use of these algorithms has been recently reviewed [10]. Briefly, data warping and interpolation are used to stretch and compress data in order to objectively optimize the match between analyte peaks in a “target” GC × GC–TOFMS separation and the analyte peaks in a “sample” separation. The application of 2D

[*] Corresponding author. Tel.: +1 206 685 2328; fax: +1 206 685 8665.
E-mail address: synovec@chem.washington.edu (R.E. Synovec).

[1] Current address: Dow Chemical Corporation, 2301 North Brazosport Boulevard, Freeport 77541, TX, USA.

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.06.038

alignment strategies for comprehensive non-targeted GC × GC–TOFMS is computationally expensive and preserving the peak signal "volumes" of every analyte during interpolation and warping is difficult and often exhibits shortcomings. In general, it is nearly impossible to preserve both signal intensity and signal area (or volume) while applying such alignment. Additionally, for non-targeted approaches, it is highly beneficial to maximize the peak capacity of the 2D separations while maintaining a tri-linear data structure for subsequent deconvolution and quantification [15,16]. With GC × GC–TOFMS, because of the cryogenically focused injection onto the secondary column with thermal modulators, increasing the modulation period (separation run time in the second dimension) generally results in higher peak capacities as well as better chemical selectivity. Optimization for higher peak capacities preferred for non-targeted discovery-based experimentation often leads to a small number of modulations acquired per first dimension peak (~2–4 modulations commonly applied), and thus makes efforts to apply 2D alignment problematic due to a limited data density in the first separation dimension.

Non-targeted discovery-based experimentation generally aims to comprehensively analyze complex chromatograms to bring to light important analytes and/or chemical fingerprints, representing the chemistry that is significantly changing in the context of the experimental design. Non-targeted approaches can be either supervised or unsupervised, where supervision refers to either external calibration or prior classification of chromatograms as they relate to the experimental design. Popular supervised methods are Fisher ratio (referred to herein as F-ratio) and partial least squares discriminant analysis (PLS-DA). F-ratio methods have been applied to GC × GC data [17] and GC × GC–TOFMS data at the pixel-level [18–20] and at the peak table-level (LECO Fisher Ratio ChromaTOF 2009). PLS-DA has also been recently applied successfully to peak table-based GC × GC–TOFMS data [21–24].

Peak table-based data is acquired by peak finding, deconvolution, and alignment across peak tables. The steps for peak table preparation can be computationally demanding, especially deconvolution, and the majority of the signals processed are often not of interest to the experimental design being implemented. Application of a pixel-level based F-ratio approach, performed prior to any peak detection or deconvolution has been proposed previously as an approach to reduce the initial GC × GC–TOFMS data set down to only those 2D separation locations which change significantly between sample classes per the experimental design. Following the pixel-level F-ratio determinations, only those 2D locations undergo deconvolution and peak quantification, improving the efficiency of the discovery-based analysis [17–20].

While the previously reported pixel-level based F-ratio algorithm (referred to herein as software) has the aforementioned distinct intrinsic advantages over a peak table-based approach for discovery-based studies, the prior F-ratio software for application to GC × GC–TOFMS data has some shortcomings that need to be addressed in a sufficiently peak-based way. In order for this powerful algorithmic approach to be effectively implemented, it is essential to (1) reduce the number of false positives that are generated, while (2) significantly improving the sensitivity "contrast" of finding true positives (e.g., spiked compounds). First dimension misalignment can severely impact the sensitivity contrast for the F-ratio determination of true positives with pixel-level data. First dimension misalignment in GC × GC (i.e., retention time variation) is a result of desynchronized modulation of a given peak from one sample run to another, referred to herein as "phasing". Phasing can produce widely differing peak heights (for a given modulated peak on the second separation dimension) from one injection replicate to the next, even when the total peak areas from the sum of all modulations of a given analyte are, in principle, identical. Phasing causes bias of the F-ratio statistical analysis, as well as causes supervised non-target algorithms based in the pixel-level data analysis mode to yield high false-positive rates while concurrently reducing the sensitivity contrast.

In this report, we propose and demonstrate a solution that is unbiased to peak shape or specific mass spectral fragmentations, like pixel-level based algorithms, as well as sufficiently peak based to avoid the challenges of GC × GC phasing. Relative to our prior F-ratio software that functioned strictly on the pixel-level data [17–20], the proposed F-ratio algorithm significantly reduces false positives while concurrently improving the sensitivity contrast by which true positives are identified. The proposed algorithm, referred to as the tile-based F-ratio software, works by creating a 2D grid (encompassing the entire GC × GC separation) composed of 2D tiles, whereby each tile is wide enough to capture the retention time variation in both the first and second separation dimensions and sums all signals within the tile as a function of each mass channel (m/z). Thus, the tile-based F-ratio software is designed to eliminate the need to have a rigorous bilinear data structure from the separation [15]. Four adjacent, but overlapping 2D grids, are used so that one grid will optimally capture any given peak in the GC × GC separation. Use of four grids, combined with summing of all chromatographic signal within a tile (at a given m/z), provides data reduction and added sensitivity, similar to peak-based algorithms. Sensitivity is improved because a single grid (of four) that captures a given analyte peak the best in one of its tiles will have the least amount of interference. The signal-to-noise (S/N), i.e., the sensitivity contrast, for a given analyte peak is improved roughly by the square root of the number of points summed, and by minimizing the impact of the GC × GC phasing effect. However, if the tile size is inadvisably too big (as discussed in the data analysis section), summation of noise will adversely impact the S/N benefit achieved through summation. We use the standard addition method by spiking both native and non-native chemicals into a diesel fuel to demonstrate the tile-based F-ratio software in the context of a challenging complex sample matrix. Preparatory and injector variation was normalized by use of the internal standard 1-bromoheptane.

2. Experimental

2.1. Sample preparation

Eight analyte compounds (four non-native and four native to diesel) were spiked at the following nominal concentrations into an ultra-low sulfur diesel (ULSD) fuel: 1000, 750, 500, 250, 100 and 0 parts-per-million by mass (ppm). An internal standard, 1-bromoheptane (also non-native to the diesel fuel), was spiked at a concentration of 1 part-per-thousand (ppt) into each diesel sample. The four, non-native spiked compounds were bromobenzene, 1-chlorohexane, 5-decyne, 3-octanone and the four native spiked compounds were butylcyclohexane, cyclohexylbenzene, heptane and o-xylene. All diesel samples were prepared gravimetrically using a 5-place analytical balance, and the actual (not nominal) concentration for each analyte at each spike level is provided in Table 1. The actual concentration of each spiked analyte is used as much as possible in this report. However, for clarity in certain instances, we use the nominal spike concentrations when referring to a particular concentration comparison.

2.2. Instrumental parameters

The GC × GC–TOFMS instrumental platform consisted of an Agilent 6890N gas chromatograph equipped with an Agilent 7683 autoinjector (Agilent Technologies, Palo Alto, CA, USA) coupled with a Leco Pegasus III time-of-flight mass spectrometer,

Table 1
The actual concentrations are shown in ppm for each nominal spike concentration (first column) for all spiked analytes. The nominal concentrations are used occasionally in the text for brevity and clarity in discussing the data set.

Nominal spike (ppm)	1-Chlorohexane (ppm)	Bromobenzene (ppm)	3-Octanone (ppm)	5-Decyne (ppm)	Heptane (ppm)	o-Xylene (ppm)	Butylcyclohexane (ppm)	Cyclohexylbenzene (ppm)
1000	1172	938	933	1210	1562	3680	2517	1830
750	889	712	708	918	1332	3393	2288	1579
500	608	487	484	628	1104	3107	2060	1329
250	308	247	245	318	860	2802	1818	1063
100	141	113	112	146	725	2632	1683	914
0	0	0	0	0	610	2489	1569	789

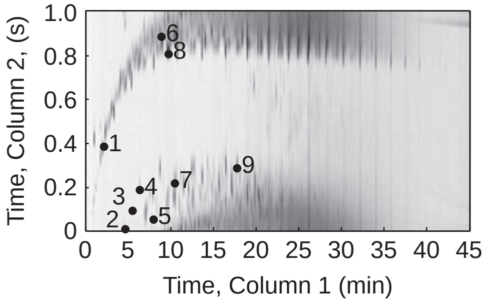

Fig. 1. The GC × GC–TOFMS base ten logarithm (log10) contour plot of the TIC of a nominal 100 ppm spiked diesel sample and the locations of all nine spiked compounds (eight analytes and the internal standard). The first dimension elution order and column 1 retention time of the nine spiked analytes is as follows: (1) heptane at 2.2 min; (2) 1-chlorohexane at 4.6 min; (3) o-xylene at 5.5 min; (4) bromobenzene at 6.4 min; (5) 3-octanone at 7.8 min; (6) butylcyclohexane at 9.0 min; (7) 1-bromoheptane at 9.2 min; (8) 5-decyne at 9.3 min and (9) cyclohexylbenzene at 17.2 min.

and equipped with the 4D thermal modulator upgrade (Leco, St. Joseph, MI, USA). The GC × GC–TOFMS instrument was used to analyze the diesel fuel samples at all spike levels. The primary column of the GC × GC (column 1) was a 20 m × 250 μm inside diameter × 0.5 μm DB-5 film (J&W Scientific/Agilent Technologies, Santa Clara, CA, USA), producing the first dimension separation. The secondary column (column 2) was a 2 m × 180 μm inside diameter × 0.2 μm RTX-200MS film (Restek, Bellefonte, PA, USA), producing the second dimension separation. The GC instrument inlet was set at 275 °C and the transfer line was set at 305 °C. Column 1 was held at 50 °C for 0.25 min and then increased at 5 °C/min to 300 °C, where it was held for 5 min. Column 2 was initially set at 55 °C and followed the same temperature program as column 1 giving a total run time of 55.25 min. The modulator was kept 20 °C higher than column 1, and the modulation period was 1 s. The GC instrument was set to maintain a constant (ambient temperature and pressure corrected) flow rate of 2 ml/min at the outlet of column 2, with helium used as the carrier gas. The ion source was set to 300 °C and the detector voltage was set to 1600 V. Mass channels, m/z 41–340, were collected at 100 spectra/s after a 6 s solvent delay. A 1 μl injection of each diesel sample was made in split mode with a split ratio of 200:1. All diesel samples were injected in quadruplicate, producing a total of 24 runs (6 spike levels injected in quadruplicate). An example separation run is presented in Fig. 1, with the GC × GC–TOFMS base ten logarithm contour plot of the TIC of a nominal 100 ppm spiked diesel sample. The locations of all nine spiked compounds are indicated in the figure (eight analytes and the internal standard). Note that the GC × GC separation conditions were selected to allow some wrap-around to more fully utilize the 2D peak capacity, without having compounds wrap-around too much. Therefore, compounds in a given second dimension separation are not allowed to wrap-around physically into compounds eluting in a subsequent second dimension separation. Fig. 1 was not adjusted to hide the wrap around (when the range of secondary retention times exceeds the modulation period), which is commonly practiced but not necessary.

2.3. Data analysis methodology

GC × GC–TOFMS data from all of the 24 runs were imported from LECO's ChromaTOF software v 3.32 (LECO, St. Joseph, MI) to Matlab v 7.0.4 via in-house written software that utilized a peg2mat function, which converts the instrumental '.peg' files directly into a MATLAB workspace. The imported GC × GC–TOFMS data was then analyzed with the in-house developed tile-based F-ratio software. For algorithm validation, an in-house developed target-PARAFAC GUI was used to target and extract the quantitative signal "volume" at each of the spiked analyte chromatographic locations and to obtain the normalization values for the internal standard [16]. We refer to each analyte signal as the signal volume because it represents the signal area for a given analyte along the column 2 time axis summed over all column 1 modulations in the GC × GC separation. The determination of absolute concentrations of each nominal spike level was determined via the standard addition method (SAM) and linear regression, with examples presented in Fig. 2 for bromobenzene in Fig. 2(A) and butylcyclohexane in Fig. 2(B). For bromobenzene the y-intercept is essentially zero, indicating it is not present in the unspiked fuel. In contrast, the y-intercept for butylcyclohexane is reasonably large, and by application of the SAM, the concentration in the unspiked fuel was determined to be 1569 ppm. Accordingly, a complete list of concentrations is presented in Table 1 with the corresponding nominal spike level. The concentration ratios comparing all nominal spike concentration levels are summarized in Table 2 for all analytes.

At the pixel-level, an F-ratio is calculated at every point in the 2D separation, as a function of m/z. An F-ratio is the class-to-class variation of the detected signal divided by the sum of the within-class variations of the signal [17,25,26]. The class-to-class variation is calculated as

$$\sigma_{\mathrm{cl}}^2 = \sum \frac{(\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})^2 \boldsymbol{n}_{\boldsymbol{i}}}{(\boldsymbol{k} - 1)} \tag{1}$$

where $\boldsymbol{n}_{\boldsymbol{i}}$ is the number of measurements in the ith class, $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ is the mean of the ith class, $\bar{\boldsymbol{x}}$ is the overall mean, and $\boldsymbol{k}$ is the number of classes. The within-class variation is calculated as

$$\sigma_{\mathrm{err}}^2 = \frac{\sum(\sum(\bar{\boldsymbol{x}}_{\boldsymbol{ij}} - \bar{\boldsymbol{x}})^2) - (\sum(\bar{\boldsymbol{x}}_{\boldsymbol{i}} - \bar{\boldsymbol{x}})^2 \boldsymbol{n}_{\boldsymbol{i}})}{(\boldsymbol{N} - \boldsymbol{k})} \tag{2}$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{ij}}$ is the ith measurement of the jth class, and $\boldsymbol{N}$ is the total number of sample profiles. An F-ratio is then calculated as the ratio between the two variances,

$$\text{Fisher ratio} = \frac{\sigma_{\mathrm{cl}}^2}{\sigma_{\mathrm{err}}^2} \tag{3}$$

For the pixel-level F-ratio software, the F-ratio in Eq. (3) is initially calculated as a function of the 2D time location of the pixel-level data, at each m/z. Finally, with the pixel-level F-ratio software, the F-ratios at each m/z are summed together to give a single sum of F-ratio at each 2D pixel location.

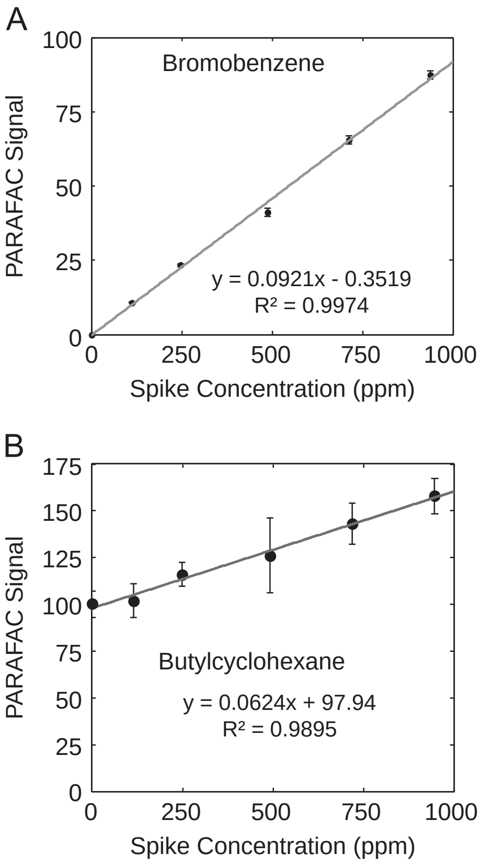

Fig. 2. (A) A calibration curve is plotted using PARAFAC signal volumes as a function of the spike concentration for a non-native analyte, bromobenzene. (B) A calibration curve is plotted using PARAFAC signal volumes as a function of the spike concentration for a native analyte, butylcyclohexane. The PARAFAC signal volumes correlate linearly with the standard concentration spike levels. From a linear fit of the data, based on the standard addition method, the native concentration of butylcyclohexane in the diesel was calculated, and found to be 1569 ppm. This SAM approach was performed to determine the concentration for all native analytes in the unspiked diesel.

In contrast, for the tile-based F-ratio software, the GC × GC pixel-level data is summed using a 2D grid of 2D tiles, as a function of m/z, which provides data reduction in the 2D time domain. Eqs. (1)–(3) are then applied as a function of 2D tile location, instead of the 2D pixel location. Then, like the pixel-level algorithm, the F-ratios at each m/z are summed to give a single sum of F-ratio at each 2D tile location. Four 2D grids of tiles are applied to optimally capture each peak. For added information, F-ratio "spectra" are also reported, whereby the F-ratios as a function of m/z are provided for tiles of interest.

All tile-based F-ratio analyses were conducted with the entire m/z range (41–340). Prior to tile-based F-ratio analysis, the data was baseline corrected and normalized to the internal standard signal for each sample (normalization values provided by quantification using target-PARAFAC, [16]). A S/N threshold was applied by computationally ignoring detector signal that was less than three times the standard deviation of a tiled noise region. The noise used for the calculation of an S/N threshold for this study was the first 3 s of detector signal during a representative chromatographic run. The first three seconds of data is binned into 160 chromatographic points in each bin (to represent the 8 s by 200 ms tile used for the tile grids) and each bin is summed. The

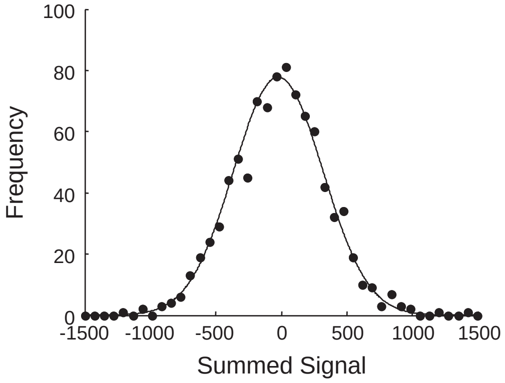

Fig. 3. A representative histogram, frequency of occurrence versus the summed signal, from the distribution of a binned noise region (160 data pixels, the size of a tile) for the first 3 s of data collect for one GC × GC–TOFMS chromatogram. The distribution of summed signal is randomly distributed (suitably Gaussian) with the mean of the distribution essentially at zero. In the tile-based algorithmic approach, this distribution is created for each sample and three times the standard deviation is used as a S/N threshold for each sample. The standard deviation for this particular sample is 366, which would then be used to remove any tiles with summed signal below 1098 for this sample.

Table 2
The concentration ratios for each spiked analyte for each paired sample class are reported. The nominal 0 ppm versus 100 ppm comparison is the most challenging comparison.

Concentration comparison	1-Chlorohexane	Bromobenzene	3-Octanone	5-Decyne	Heptane	o-Xylene	Butylcyclohexane	Cyclohexylbenzene
1000–0	–	–	–	–	2.56	1.48	1.60	2.32
1000–100	8.30	8.30	8.30	8.30	2.15	1.40	1.50	2.00
750–0	–	–	–	–	2.18	1.36	1.46	2.00
1000–250	3.80	3.80	3.80	3.80	1.82	1.31	1.38	1.72
750–100	6.29	6.29	6.29	6.29	1.84	1.29	1.36	1.73
500–0	–	–	–	–	1.81	1.25	1.31	1.69
750–250	2.89	2.89	2.89	2.89	1.55	1.21	1.26	1.49
500–100	4.30	4.30	4.30	4.30	1.52	1.18	1.22	1.45
1000–500	1.93	1.93	1.93	1.93	1.42	1.18	1.22	1.38
250–0	–	–	–	–	1.41	1.13	1.16	1.35
500–250	1.97	1.97	1.97	1.97	1.28	1.11	1.13	1.25
750–500	1.46	1.46	1.46	1.46	1.21	1.09	1.11	1.19
1000–750	1.32	1.32	1.32	1.32	1.17	1.08	1.10	1.16
250–100	2.18	2.18	2.18	2.18	1.19	1.06	1.08	1.16
100–0	–	–	–	–	1.19	1.06	1.07	1.16

"–" means the concentration ratio is not defined since the denominator is essentially zero.

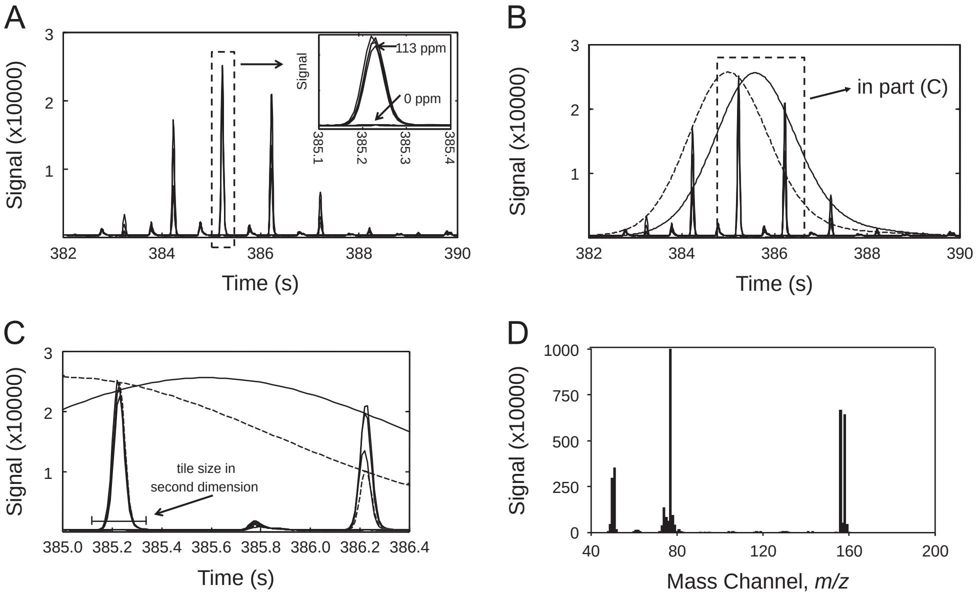

Fig. 4. (A) The unfolded raw 2D chromatogram at *m*/*z* 77, zoomed in on the bromobenzene peak shows four separate injections of a 113 ppm bromobenzene spiked diesel and four separate injections of a non-spike diesel as labeled in the zoomed in box. The further zoomed in second dimension peak in the dashed box, produced the highest pixel-based F-ratio value for bromobenzene (max F-ratio equal to 216). (B) Two Gaussian peaks have been superimposed, showing the first dimension bromobenzene peak sampled by the GC × GC modulator, for the two sample injections with the most retention time shifting. (C) A zoom in view of (B) shows the effect of phasing on the within-class variation of peak height (e.g., the modulated peak at 386.2 s). The black bracket indicates the size (200 ms) of the tile in the second dimension (subsequently used with the tile-based F-ratio software). (D) The average chemical spectra for the bromobenzene peak.

distribution of summed signal is randomly distributed and sufficiently Gaussian (Fig. 3), with the mean of the distribution essentially at zero. This distribution is created for each sample and three times the standard deviation is used as a *S*/*N* threshold for each sample. After the *S*/*N* threshold is applied the F-ratio values for each *m*/*z* within a given tile were summed, giving a singular sum of F-ratio value that is used to rank the values in a hit list. Each row in the hit list is called an individual hit. For clarity, in some figures, a smaller *m*/*z* range may be plotted (less than 41–340), if there was not signal above the *S*/*N* threshold at higher *m*/*z*.

3. Results and discussion

An advantage of pixel-based feature selection and classification algorithms is that no assumptions are made regarding peak shape or spectral features and do not rely on computationally expensive deconvolution algorithms, as do peak table-based approaches. However, there are two major shortcomings to strictly pixel-based F-ratio algorithms. First, there often are many false positives, and second, the true positives may have low F-ratio values that do not provide sufficient contrast from false positives and true negatives. To address these two shortcomings, we investigated the causes of observing low F-ratio values for true positives (in this study using spiked analytes), and we investigated the causes of observing high values for locations that do not correspond with the spiked analytes (false positives). We developed a new algorithmic approach to implement a software solution to address these shortcomings.

Sensitivity contrast is the most important feature that the F-ratio software should provide the analyst. Increasing contrast is achieved by maximizing the chemical selectivity in discovering statistically significant changes between the two (or more) sample classes. Ideally, true positive hits should have vastly higher F-ratio values than false-positives to ensure that small changes in the concentration ratio for a given true positive hit can be detected. To illustrate and address these issues, we begin with the comparison of quadruplicate injections of a 113 ppm bromobenzene spike in diesel versus quadruplicate injections of unspiked diesel, Fig. 4(A), with the bromobenzene peak from a zoomed-in section of the separation also indicated (an eight modulation time window along the column 1 time axis, 8 s, since the modulation period is 1 s). Gaussian peaks have been superimposed showing the nominal peak width for the first separation dimension peak as it is modulated onto the shorter second dimension separation (Fig. 4 (B) and (C)). The misalignment in the first dimension that is observed in the two superimposed Gaussian peaks in Fig. 4 (B) generally causes significant peak height variation (referred to as "phasing") for each modulated second dimension peak across the four separate sample replicate injections of the nominal 100 ppm spiked diesel (actual concentration of bromobenzene is 113 ppm). Phasing can produce widely differing peak heights (for a given modulated peak on the second separation dimension) from one injection replicate to the next, even when the total peak areas from the sum of all modulations of an analyte should be, in principle, identical. In this example in Fig. 4, the highest F-ratio obtained by the pixel-based F-ratio software for bromobenzene was found to be 216 (for the selective *m*/*z* 77). As we develop the tile-based F-ratio software, we shall compare the F-ratio obtained to this pixel-based F-ratio of 216, which serves as a benchmark metric. Fig. 4(D) shows a library mass spectra (National Institute for Standards in Technology 2011) from the bromobenzene peak shown in Fig. 4(A)–(C). Mass channel *m*/*z* 77 was used for each chromatographic data plot in Fig. 4. The mass spectrum for each spiked analyte, in conjunction with the 2D retention time, is used to identify the analyte associated with each F-ratio spectra in subsequent figures.

Another significant shortcoming of applying pixel-based F-ratio analysis is the common occurrence of false positives, with a numerically very substantial false positive hit shown in Fig. 5. This false positive hit was on the shoulder of a large native compound peak (eicosane), with the large peak not changing in concentration from one class to the other class. The actual pixel that resulted in a high F-ratio is shown in the zoomed-in selection at 1404 s along the first dimension and 0.71 s along the second

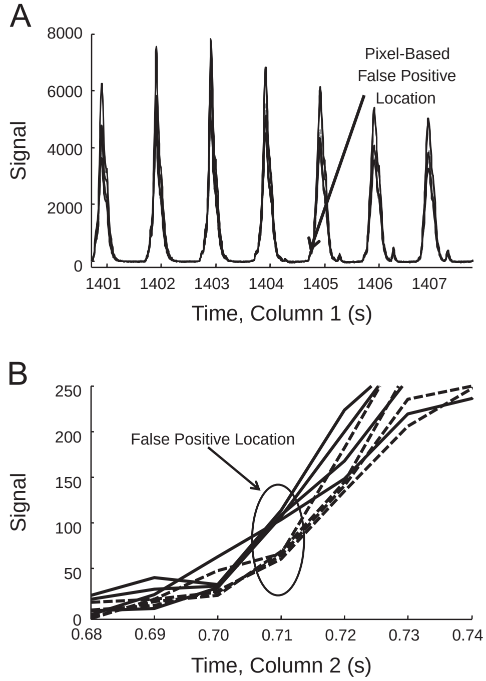

Fig. 5. (A) The chromatographic location in the raw 2D data of a high hit number (from application of previous pixel-based F-ratio software) at the selective *m*/*z* 95, that appears on the shoulder of a peak (native compound in diesel). Four injections of a 113 ppm bromobenzene spiked diesel (solid) and four injections of a unspiked diesel (dashed) are shown. (B) Random fluctuation of the detector coincidently co-vary with the sample classes at 1404.71 s, producing a large false positive hit for this location. The pixel-based F-ratio value at *m*/*z* 95 for this hit no. was 249.

dimension in Fig. 5(B), in which we overlay four runs from each samples class (100 ppm spiked versus unspiked). The raw data in Fig. 5(A) along the first dimension time axis are plotted in Fig. 5(B) and plotted along the second dimension time axis for clarity, with it understood that we are zooming in on the modulation initiated at 1404 s. Randomly occurring covariance of detected signal (above the *S*/*N* threshold) resulted in a pixel-based F-ratio value of 249 at *m*/*z* 95 (other *m*/*z* not shown for brevity also indicated the same result). This result is of great significance because this pixel-based F-ratio value for a single *m*/*z* of a known false positive is higher (i.e., more prominent) than the pixel-based F-ratio obtained for the true positive bromobenzene at a single selective *m*/*z* (Table 3).

To address these shortcomings (low sensitivity contrast and high false positive rate) in the pixel-based F-ratio software performance, we developed the tile-based F-ratio software utilizing the positive attributes of the pixel-based and peak-table based approaches. We have developed an algorithmic approach that functions by creating a 2D tile (i.e., a 2D time window) that is wide enough to capture the retention time variation in both the first and second separation dimensions, and summing all signals at a given *m*/*z* within the tile. The size of the tile is chosen to contain ~99% of a Gaussian peak in both the first and second dimensions, plus two extra modulations in the first dimension, and additional time greater than the peak width in the second dimension, in order to fully contain retention time variation in each dimension (i.e., eliminating the negative impact of phasing from one sample injection to the next). For this data set, the window size was set at eight modulations in the first dimension, the time window of the data in Fig. 4(A), and 0.20 s in the second dimension, which is the inset caliper in Fig. 4(C). These tile dimensions were acceptable sizes for all the spiked analytes in this study. This tile-based signal-summation approach removes false positives caused by random covariance of detector noise within and across sample classes (as illustrated in Fig. 5, because summing all signals within a given tile very effectively averages out false positive pixels). Data processing strategies could be applied to define grids in an automated fashion, but some optimization of the grid size would be required.

Table 3
Results showing the increased sensitivity contrast achieved by using the tile-based F-ratio software versus the previous pixel-based F-ratio software. The known pixel-based false positive F-ratio values (Fig. 4) have been decreased by two orders of magnitude, while the known true positive F-ratio values have increased by two orders of magnitude.

	Pixel-based	Tile algorithm
F-ratio (*m*/*z* 95) false positive	249	1.1
F-ratio (*m*/*z* 77) bromobenzene	216	9103
Sum of F-ratio false positive	1068	88
Sum of F-ratio bromobenzene	2799	213,712

One set of tiles, referred to as a tile grid, covering the entire 2D chromatogram is not sufficient, because with only one tile grid some peaks will be split into two or more tiles. Thus, four overlapping rectangular tile grids are required. Tile grids are shifted by half a tile length in the first separation dimension, or shifted half a tile length in the second separation dimension, or shifted half a tile length in both dimensions (illustrated in Fig. 6(A)). Naming of each of the tile grids is done in relation to the Cartesian quadrant system. Tile grid 3 (t3) is anchored with the lower left corner at the origin of the 2D chromatogram (lower-left quadrant) while tile grid 1 (t1) is shifted right and up (upper-right quadrant), with the other tile grids named with respect to the direction of the grid shifting. With this tile-based F-ratio software, baseline corrected signal at all of the 2D points (or pixels) in each tile is summed to provide a single signal value per tile, and per each *m*/*z*.

The four tile grids are used to ensure that one of the four grids will optimally capture the entire signal for a given analyte peak hit, and thus, that one tile out of the four tile grids will have the highest F-ratio value for a given true positive analyte hit. The other three tile grids inevitably split the analyte peak. Fig. 6 illustrates the tile grid approach with an illustration of the tile grids (Fig. 6(A)) and contour plots of zoomed portions the chromatograms from the 113 ppm bromobenzene quadruplicate injections (Fig. 6(B)). The bromobenzene peak is misaligned slightly in both dimensions (also shown in Fig. 4), but alignment is not needed because summing all data points within the tile obviates the need to align.

As shown in Fig. 6(B), tile grid t4 optimally captures the bromobenzene peak, and minimizes retention time variation of the four injections, resulting in the highest F-ratio of 9103 at *m*/*z* 77 (40-fold greater than by the pixel-based F-ratio software, as per Fig. 4). This example illustrates how the tile-based F-ratio software substantially increases the sensitivity contrast of the hit discovery by decreasing the within-class variance (denominator of the F-ratio per Eq. (3)), while at the same time improving the between-class variance by boosting *S*/*N* with summation (numerator of F-ratio per Eq. (3)). The F-ratio spectra (as a function of *m*/*z*) for adjacent tiles for bromobenzene are shown in Fig. 7. The mass spectrum shown in Fig. 4(D) can be compared with each tile's F-ratio spectrum to assist in the identification of bromobenzene as the analyte changing between the two classes. The sum of F-ratio values for the three most adjacent tiles are included; they are much less than the bromobenzene tile in tile grid t4.

Application of the four tile grids to a larger section of 2D data containing multiple peaks is shown in Fig. 8. Each individual peak is optimally sampled by only one of the four tile grids and that tile grid will have the highest F-ratio value for all selective *m*/*z*,

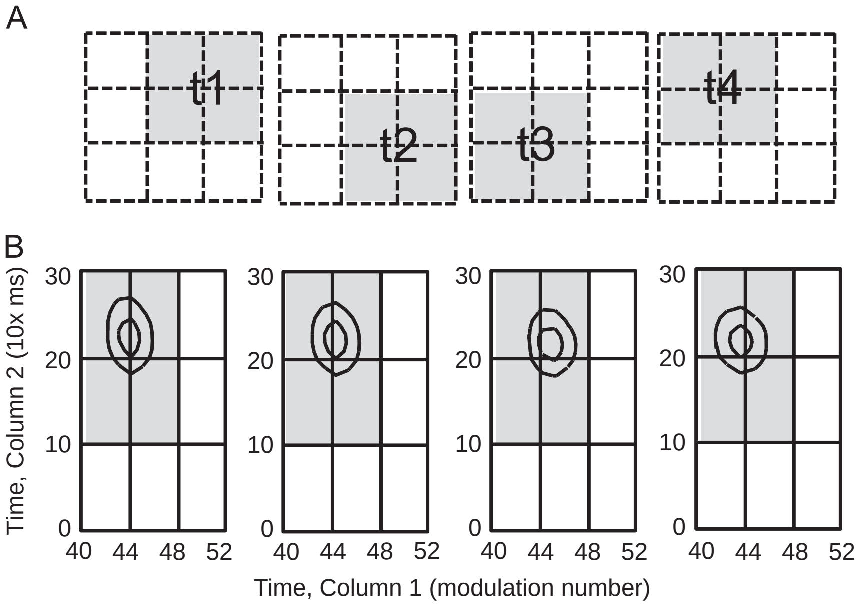

Fig. 6. (A) A graphical representation of the differences in position of the four tile grids used in concert with the tile-based F-ratio software. All the points within each of the gray rectangle tiles (t1, t2, t3, t4) are summed to create the four binned tile grids. Tile grid t3 is centered at the origin of the 2D chromatogram. (B) Four 113 ppm bromobenzene spiked diesel injections are shown, zoomed in upon bromobenzene at m/z 77. The contour lines for the signal are at ± 1 SD (67% of signal intensity, inner ring) and ± 2 SD (95% of signal intensity, outer ring) of the bromobenzene peak are shown, where SD is the standard deviation of the peak based upon a Gaussian peak shape definition. Tile grid t4 (shaded) optimally captures the small retention time misalignment seen for the bromobenzene peak over the four separate injections.

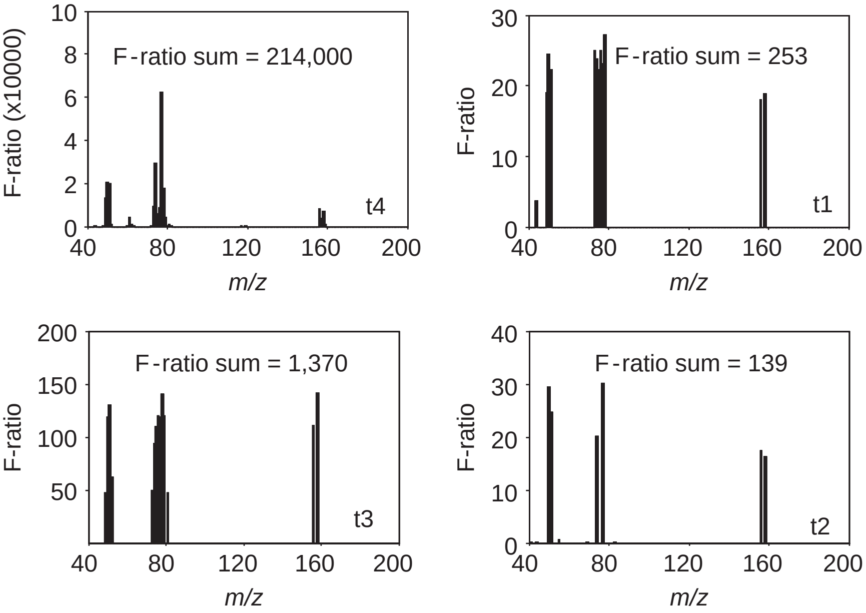

Fig. 7. The F-ratio spectra (value at each m/z) for each tile grid (t1, t2, t3, and t4) for bromobenzene are indicated in each plot. The F-ratios from the comparison of bromobenzene at 113 ppm versus 0 ppm at each m/z are provided, along with the F-ratio sum at all m/z, at the same ^{1}D and ^{2}D tile number for the four tile grid schemes that each optimally contains the bromobenzene peak. Tile grid t4 optimally captures the retention time alignment and phasing effects, and thus, provided the highest F-ratio values, indicating it contains the bromobenzene peak to the greatest extent of the tile grids. A S/N mask has been applied removing any signal from m/z without statistically significant detector signal.

because that is where the within-class variance will be the smallest. However, the software at the stage reported herein, produces some “redundant” hits initially for each analyte peak, because of the different grid schemes, but one grid scheme will be optimal for each analyte. In a future software development stage, it is envisaged that the software will readily allow for removing redundant hits, but this is not discussed here for brevity.

The tile for a given grid that captures a peak in the center of the tile will have the largest F-ratio values for each selective m/z of that chemical component, and is the chromatographic location of most interest. However, tiles from other redundant tile (and associate grid) locations that contain portions of multiple peaks from other analytes not centered in a given tile will have higher within-class variation (due to phasing) and lower sensitivity (the summation of less chemical signal), and thus will have lower F-ratio values per m/z. These locations are less informative than for a tile (for a given grid) that optimally captures a given peak in the center of a tile. In special cases where two chemical components are changing in close proximity, even when overlapped or if an increase in signal for one and a decrease in the other occurred, the

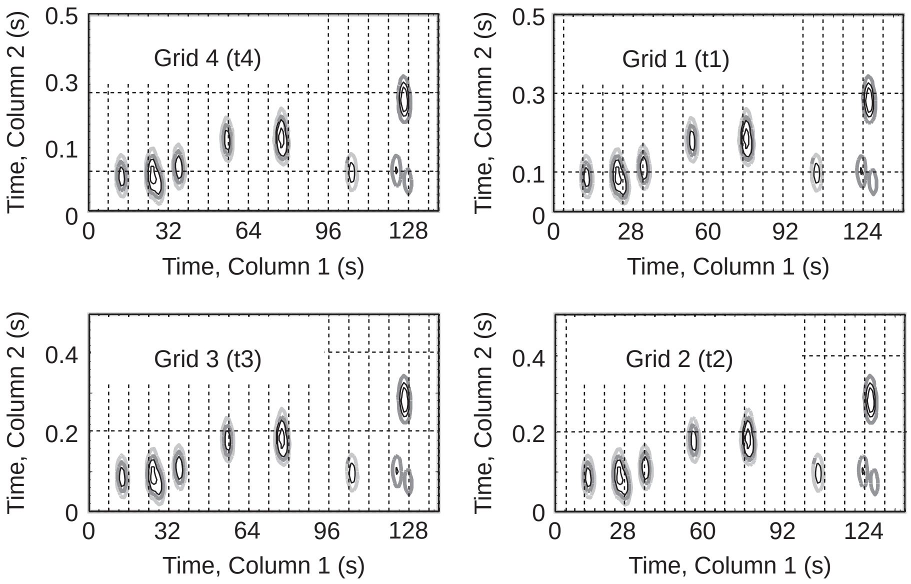

Fig. 8. A contour plot from a subset of a nominal 100 ppm spiked diesel chromatogram is shown to illustrate the four tile grids (t1, t2, t3, t4) with a width of 8 modulations along the first dimension time axis (^{1}D), and a height of 200 ms along the second dimension time axis (^{2}D). Each peak is optimally contained by only one of the tile grids (more than ~99% of the peak area for typical 2D peaks).

F-ratio per m/z allows the discovery of these overlapped chemical components. For initial ease of analysis, the sum of F-ratio value for each tile is calculated to provide a list for the analyst to begin to further identify the changing chemical components. The F-ratio as a function of m/z for a given location gives the user additional identification information. While the F-ratio could also simply be calculated based on the total ion current (TIC) signal, this approach would not take advantage of the chemical selectivity provided by the full m/z range approach implemented.

To summarize (see Table 3) with these two representative hit examples (the true positive for bromobenzene in Fig. 4, and the false positive in Fig. 5), utilizing a novel tile-based software with four tile grids has significantly increased sensitivity contrast for the sum of F-ratio value for the true positive by two orders of magnitude (an increase in chemical selectivity across sample classes), relative to the previously reported pixel-based F-ratio software [17–20]. Concurrently, the sum of F-ratio value for the highest false positive by pixel-based analysis has decreased by two orders of magnitude. Thus, the tile-based F-ratio software provides a substantial improvement in reducing false positives, concurrent with enhancing the sensitivity contrast to discover true positives.

Application of the tile-based F-ratio software to the entire 2D chromatogram for the nominal 100 ppm spiked diesel versus 0 ppm (unspiked) diesel, summarized in Table 4, indicates that the eight spiked analytes are the top eight hits. Note that at this point in the software development, the redundant hits caused by peak splitting in adjacent tile grids have been removed by analyzing the spectrum for each tile. The first false positive seen after the eight spiked compounds has a sum of F-ratio value of 643 (provided in Table 4), with only half the sum of F-ratio value obtained for cyclohexylbenzene, which had an absolute concentration ratio of only 1.16 (since it was present in the unspiked fuel at a concentration of 789 ppm). Of further note, the tile-based F-ratio software detected o-xylene with a sum of F-ratio value of 3464 at an absolute concentration ratio of 1.06, suggesting that very small concentration ratios can likely be found for some compounds of interest in future studies. Thus, the up or down change of analytes of interest that can be confidently discerned, from one sample class to the next, appears to be approaching very small concentration ratios. This is very important for studying complex samples in which only small concentration changes may be occurring. As opposed to our previous studies, no threshold value for the observed F-ratios was used in the present report. However, a threshold value may be of interest, and could be calculated using methods such as uninformative variable elimination PLS (UVE-PLS) [27], iterative variable elimination PLS (IVE-PLS), robust PLS, recursive outlier removal PLS [28], or a variety of statistical methods. Overall, the tile-based F-ratio software now provides outstanding sensitivity contrast to discover true positives relative to false positives.

Table 4
The beginning portion of the hit list is shown, resulting from the tile-based F-ratio software comparison of a nominal 100 ppm spiked diesel versus the 0 ppm unspiked diesel. The first false positive in the hit list is the last entry in the table. At this stage of the software development, redundant hits caused by splitting of each of the eight analytes were removed by analyzing the mass spectrum of each in adjacent grid tiles. Each tile no. on ^{1}D and ^{2}D is defined for each of the four grid schemes, so each of the ^{1}D and ^{2}D locations per each grid scheme is at a slightly different 2D locations in the GC × GC separation (as illustrated in Fig. 7). For example, for hit no. 1 there will be a tile location no. 34 on ^{1}D and tile location no. 5 on ^{2}D for each of the four grids, but the 1-chlorohexane peak is centered optimally in only one grid (t1 in this case).

F-ratio hit no.	Sum of F-ratio	Tile no., ^{1}D	Tile no., ^{2}D	Grid	Compound
1	807,000	34	5	t1	1-Chlorohexane
2	214,000	48	1	t4	Bromobenzene
3	84,700	70	4	t4	5-Decyne
4	67,300	58	5	t4	3-Octanone
5	3460	41	1	t3	o-Xylene
6	3180	16	2	t4	Heptane
7	2070	67	5	t2	Butylcyclohexane
8	1270	128	2	t2	Cyclohexylbenzene
First false positive	643	175	1	4	Eicosane

In comparison to the previous pixel-based software, the tile-based F-ratio software performs much better. Table 5 shows the results of the previous pixel-based software applied to the same 100 ppm spiked diesel versus unspiked diesel as shown in Table 4 for direct comparison. While all eight spiked analytes, both the native and non-native compounds, are the first eight hits using the tile-based F-ratio software (Table 4), only the non-native compounds are readily found via the pixel-based F-ratio software (Table 5). Indeed, the native compounds, which are present at

Table 5
A portion of the hit list resulting from the pixel-based F-ratio software is shown, for the comparison of a nominal 100 ppm spiked diesel versus the 0 ppm unspiked diesel. The retention time (t_r in s) in the first dimension (^{1}D) and the second dimension (^{2}D) is defined for each high F-ratio result. Not shown here for brevity are the many multiple F-ratio values for each analyte, as a function of pixel location, that have been removed by analyzing the spectral similarities between entries in the original hit list table. The sensitivity contrast, i.e., the difference between an experimentally spiked compound's sums of F-ratio value versus a false positives sum of F-ratio value, is poor. By tiling the data we improve this contrast dramatically by removing a large source of systematic bias (phasing) to the statistical analysis of the data.

F-ratio hit no.	Sum of F-ratio	t_r ^{1}D	t_r ^{2}D	Compound
1	2799	360	0.11	Bromobenzene
2	1725	534	0.73	5-Decyne
3	1723	255	0.92	1-Chlorohexane
4	1068	1404	0.71	False positive
5	925	441	0.93	3-Octanone
6	827	1227	0.98	False positive
7	768	1543	0.99	False positive
8	742	1036	0.19	False positive
9	723	1191	0.06	False positive
…	…	…	…	…
41	536	998	0.10	Cyclohexylbenzene
281	211	515	0.77	Butylcyclohexane
367	536	110	0.31	Heptane
453	111	304	0.99	o-Xylene

concentration ratios ranging from 1.06 to 1.19, have hit numbers ranging from 41 to 453. Thus, from a practical perspective, if this analysis were performed blind, the analyst would readily be able to “discover” all eight analytes with the tile-based F-ratio software, while the “discovery” of all eight using the pixel-based F-ratio software would necessitate analyzing an impractical number of false positives. Comparable pixel-based algorithmic approaches other than our own in-house approach [18–20] can be found elsewhere [29,30].

4. Conclusions

A major challenge for non-targeted chemometric methods for 2D separations is inadvertent misalignment of chromatographic peaks from one sample run to the next. We have developed a novel algorithmic approach that addresses this challenge by taking a quick, sufficiently peak-based approach (tiling), while maintaining the advantages of pixel-based algorithmic approaches. The tile-based F-ratio software provides effective data reduction, detection of chemical patterns and is robust to misalignment.

The data reduction strategy of summing signal within a defined tile region is beneficial for GC × GC–TOFMS and is computationally fast. Utilizing four grids ensures that the highest F-ratio value calculated for a given up or down changing concentration ratio, will represent the most selective chromatographic data cube for a given discovered analyte. The further analysis of this data cube by deconvolution software (i.e., PARAFAC) is simple and the m/z associated with separation of sample class via the F-ratio analysis can be used to readily and confidently identify which PARAFAC loading (i.e., peak profile) represents the discovered chemical component.

Acknowledgment

This work was supported by the Internal Revenue Service (IRS) under an Interagency Agreement with the US Department of Energy (DOE) under Contract DE-AC05-76RLO 1830 with the Pacific Northwest National Laboratory. Battelle Memorial Institute operates the Pacific Northwest National Laboratory.

References

[1] Z. Liu, J.B. Phillips, J. Chromatogr. Sci. 29 (1991) 227–231.
[2] M. Adahchour, L.L. van Stee, J. Beens, R.J. Vreuls, M.A. Batenburg, U.A.Th. Brinkman, J. Chromatogr. A 1019 (2003) 157–172.
[3] J. Beens, M. Adahchour, R.J. Vreuls, K. van Altena, U.A.Th. Brinkman, J. Chromatogr. A 919 (2001) 127–132.
[4] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796–2804.
[5] J. Dalluge, M. van Rijn, J. Beens, R.J. Vreuls, U.A.Th. Brinkman, J. Chromatogr. A 965 (2002) 207–217.
[6] R.M. Kinghorn, P.J. Marriott, J. Hum. Rights Commonw. 21 (1998) 620–622.
[7] J.V. Seeley, F. Kramp, C.J. Hicks, Anal. Chem. 72 (2000) 4346–4352.
[8] R. Shellie, L. Mondello, P. Marriott, J. Chromatogr. A 970 (2002) 225–234.
[9] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, J. Chromatogr. A 1184 (2008) 341–352.
[10] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, J. Chromatogr. A 1255 (2012) 3–11.
[11] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 73 (2001) 5833–5840.
[12] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154–4162.
[13] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, Anal. Chem. 77 (2005) 7735–7743.
[14] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, J. Chromatogr. A 1216 (2009) 4020–4029.
[15] K.J. Johnson, B.J. Prazen, R.K. Olund, R.E. Synovec, J. Sep. Sci. 25 (2002) 297–303.
[16] J.C. Hoggard, R.E. Synovec, Anal. Chem. 79 (2007) 1611–1619.
[17] K.J. Johnson, R.E. Synovec, J. Chemometrics Intell. Lab. Syst. 60 (2002) 225–237.
[18] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756.
[19] E.M. Humston, K.M. Dombek, B.P. Tu, E.T. Young, R.E. Synovec, Anal. Bioanal. Chem. 401 (2011) 2387–2402.
[20] A.C. Beckstrom, E.M. Humston, L.R. Snyder, R.E. Synovec, S.E. Juul, J. Chromatogr. A 1218 (2011) 1899–1906.
[21] K.K. Pasikanti, J. Norasmara, S. Cai, R. Mahendran, K. Esuvaranathan, P.C. Ho, E. Chan, Anal. Bioanal. Chem. 398 (2010) 1285–1293.
[22] X. Li, X. Lu, J. Tian, P. Gao, H. Kong, G. Xu, Anal. Chem. 81 (2009) 4468–4475.
[23] C. Ma, H. Wang, X. Lu, H. Wang, G. Xu, B. Liu, Metabolomics 5 (2009) 497–506.
[24] Y. Qiu, X. Lu, T. Pang, C. Ma, X. Li, G. Xu, J. Sep. Sci. 31 (2008) 3451–3457.
[25] D.L. Massart, Chemometrics: A Textbook, Elsevier Sciences Ltd, New York, 1988.
[26] R.O. Duda, P.E. Hart, Pattern Classifications and Scene Analysis, Wiley, New York, 1973.
[27] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851–3858.
[28] R. Grohmann, T.J. Schindler, Comput. Chem. 29 (2008) 847–860.
[29] T. Gröger, M. Schäffer, M. Pütz, B. Ahrens, K. Drew, M. Eschner, R. Zimmerman, J. Chromatogr. A 1200 (2008) 8–16.
[30] J. Vial, B. Pezous, D. Thiébaut, P. Sassiat, B. Teillet, X. Cahours, I. Rivals, Talanta 83 (2011) 1295–1301.