Talanta 83 (2011) 1279–1288

Contents lists available at ScienceDirect

Talanta

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

Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGC–HRMS)

Stephen E. Reichenbach[a,*], Xue Tian[a], Qingping Tao[b], Edward B. Ledford Jr.[c], Zhanpin Wu[c], Oliver Fiehn[d]

[a] University of Nebraska – Lincoln, USA
[b] GC Image, LLC, USA
[c] Zoex Corporation, USA
[d] University of California, Davis, USA

ARTICLE INFO

Article history:
Available online 7 October 2010

Keywords:
Cheminformatics
Comprehensive two-dimensional gas chromatography
High-resolution mass spectrometry
Biomarker discovery
Sample classification
Metabolomics

ABSTRACT

This paper describes informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography (GCxGC) and high-resolution mass spectrometry (HRMS). GCxGC–HRMS analysis produces large data sets that are rich with information, but highly complex. The size of the data and volume of information requires automated processing for comprehensive cross-sample analysis, but the complexity poses a challenge for developing robust methods. The approach developed here analyzes GCxGC–HRMS data from multiple samples to extract a feature template that comprehensively captures the pattern of peaks detected in the retention-times plane. Then, for each sample chromatogram, the template is geometrically transformed to align with the detected peak pattern and generate a set of feature measurements for cross-sample analyses such as sample classification and biomarker discovery. The approach avoids the intractable problem of comprehensive peak matching by using a few reliable peaks for alignment and peak-based retention-plane windows to define comprehensive features that can be reliably matched for cross-sample analysis. The informatics are demonstrated with a set of 18 samples from breast-cancer tumors, each from different individuals, six each for Grades 1–3. The features allow classification that matches grading by a cancer pathologist with 78% success in leave-one-out cross-validation experiments. The HRMS signatures of the features of interest can be examined for determining elemental compositions and identifying compounds.

1. Introduction

Advanced instruments for molecular analysis open unprecedented vistas for biological research and are promising tools for discovering biochemical characteristics such as metabolites in tissue, blood, urine, or other fluids, that are indicative of disease, environmental exposure, metabonomics, or other health-related conditions. The pairing of comprehensive two-dimensional gas chromatography (GCxGC) and high-resolution mass spectrometry (HRMS) combines highly effective separations with precise elemental analysis. A critical challenge for effective utilization of GCxGC–HRMS for cross-sample analyses such as sample classification and biomarker discovery is the difficulty of analyzing and interpreting the massive, complex data from many samples for relevant biochemical features. The quantity and complexity of the data, as well as the large dimensionality of the metabolome and the possibility that significant chemical characteristics across many samples may be subtle and involve patterns of variations in multiple constituents, necessitate the investigation and development of new bioinformatics.

GCxGC is an advanced chemical separation technology that provides significant improvements over traditional one-dimensional GC, including an order-of-magnitude increase in chemical separation capacity, multidimensional ordering by chemical properties, and a significant increase in signal-to-noise ratio [1,2]. GCxGC separates chemical species with two capillary columns interfaced by a modulator that traps and concentrates eluents from the first column and then introduces them into the second column, producing a full secondary chromatogram for each single data point of a traditional one-dimensional separation [3,4]. Fig. 1 illustrates GCxGC system components with Zoex Corporation's dual-jet, two-stage loop, thermal modulator [5].

GCxGC–HRMS combines two powerful analytical technologies with complementary attributes: GCxGC separates chemicals in time and HRMS provides mass precision that is fine enough to

* Corresponding author at: 260 Avery Hall, University Nebraska – Lincoln, Lincoln, NE 68588-0115, USA. Tel.: +1 402 472 5007.
E-mail address: reich@cse.unl.edu (S.E. Reichenbach).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.09.057

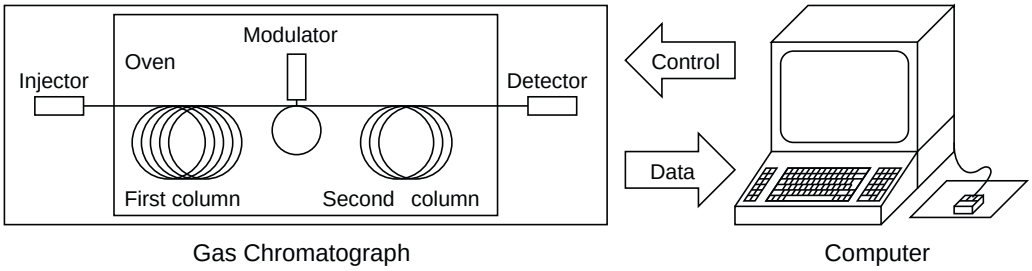

Fig. 1. Instrumentation for comprehensive two-dimensional gas chromatography (GCxGC).

distinguish elemental compositions, providing a more definitive basis for molecular identification. GCxGC is important for HRMS because the better separations significantly reduce co-elution and the problems of mass-spectral mixing. And, HRMS is important for GCxGC because the structural and compositional information available with HRMS aids interpretation of the rich, complex data from GCxGC separations.

GCxGC data can be represented in a two-dimensional array and visualized as an image with data values represented as pixels arranged so that the abscissa (*X*-axis, left-to-right) is the elapsed time for the first-column separation and the ordinate (*Y*-axis, bottom-to-top) is the elapsed time for the second-column separation. Each pixel is colorized to indicate the rate at which molecules are detected at a specific time. For HRMS data, the pixels can indicate total intensity count (TIC) or selected intensity count (SIC) in a mass interval. Each resolved chemical substance in a sample produces a small two-dimensional peak with values that are larger than the background values. In complex samples, GCxGC separates thousands of different chemicals, so the data is rich with information, but analysis of the large and complex data to detect, quantify, and identify chemical constituents is challenging [6].

Analyzing GCxGC data from multiple samples adds another level of complexity. The most important current challenge for GCxGC–HRMS informatics is to comprehensively compare features across many samples. Current state-of-the-art software supports detailed analyses of individual samples for a variety of applications. For example, automated group-type analysis can successfully characterize important characteristics of petroleum feedstocks and products and targeted analysis can accurately detect and quantify dangerous compounds in environmental samples. However, group-type and targeted analyses do not require comprehensive comparisons of every compound, whether known or unknown, across many samples.

Cross-sample analyses require correspondences of *features* (such as peak intensities) across samples and comprehensive analyses require correspondences for all features of every sample—even those for unknown compounds, trace compounds, and compounds present in some samples and not present in others. If chromatographic peaks in different samples are determined to be the result of the same compound, then the measured features of that compound can be statistically characterized and compared. The statistical characteristics of measured features of each compound in each class of samples can be used for classification and the distinctive features of each class can be investigated as a biomarker. The process of determining those features in different samples that correspond, e.g., are the result of the same compound, is *feature matching*. Feature matching is the basis for uniformly labeling structures so that similarities and differences can be documented.

Automated feature matching and comparative analysis of well-separated, well-formed peaks is typically straightforward, but comprehensive feature matching of chemically complex samples is intractable. With current methods and commercially available software, comprehensive feature matching requires semi-automated or manual processing to deal with trace compounds, coelutions, variable peak shapes, and compositional differences. These issues create uncertainties with respect to feature matching. For example, chromatographic tails of large peaks may perturb the detection of smaller peaks, e.g., such that changes in the amount of the compound causing a large peak may interfere with detection of smaller trace peaks. Likewise, if one sample has two of three compounds with similar retention times and mass spectra that are present in another sample, peak matches may be ambiguous. For pairwise comparisons or small sample sets, it can be practical (albeit tedious) to examine and match features manually, but manual processing is impractical for large sample sets. New informatics are required to successfully automate comprehensive feature matching across many samples.

The informatics developed in this paper allow comprehensive comparative analyses without comprehensive peak matching. As detailed in Section 2, the approach analyzes GCxGC–HRMS data from multiple samples to extract a feature template that comprehensively captures the pattern of peaks detected in the retention-times plane. Then, for each sample chromatogram, the template is geometrically transformed to align with the detected peak pattern and generate a set of feature measurements for cross-sample analyses. The approach avoids the intractable problem of comprehensive peak matching by using a few reliably matched peaks for alignment and peak-based retention-plane windows to define comprehensive features that can be reliably matched for cross-sample analysis.

In Section 3, the informatics are demonstrated with a set of 18 samples from breast-cancer tumors, each from different individuals, six each for Grades 1–3. The features allow classification that matches grading by a cancer pathologist with 78% success in leave-one-out cross-validation experiments. The HRMS signatures of the features of interest can be examined for determining elemental compositions and identifying compounds.

2. Material and methods

2.1. Experimental data

The new informatics are demonstrated with an experimental data set from breast-cancer tumor samples provided by Dr. Oliver Fiehn, UC-Davis. Samples were obtained from tumors from 18 individuals, six each for Grades 1–3, as determined by a cancer pathologist. Extraction protocols followed Fiehn et al. [7]. For these samples, after the −20 °C isopropanol/water/methanol extraction step, the clean-up step was critical to remove most of the triglycerides which otherwise compromise quality. Sample preparation was performed at Zoex Corporation (Houston TX, USA): 10 μL methoxyamine HCl (20 mg/mL in pyridine) were added to each sample, followed by incubation and shaking for 90 min at 30 °C, then 45 μL MSTFA were added, followed by incubation and shaking

for 30 min at 37 °C. GCxGC separations were performed by Tofwerk AG (Thun, Switzerland) on an Agilent 7890 GC and 7693 autosampler with: 1 μL splitless injection; column one HP-1MS (Agilent), 10 m × 0.25 mm, 1 μm film thickness; column two BPX 50 (SGE), 1 m × 0.1 mm, 0.1 μm film thickness; oven temperature from 40 to 310 °C (15 min) at 3.1 °C/min ramp; inlet pressure from 45 PSI to 75 PSI at 0.35 PSI/min; injection temperature 300 °C; transfer line temperature 300 °C; Zoex ZX2 thermal modulator with 6 s modulation period, 260 ms modulation duration, 375 °C hot jet temperature, 18 L/min cold jet nitrogen flow rate, and 40 PSI hot jet nitrogen pressure; and run time 100 min. The Zoex FasTOF™ time-of-flight (TOF)–HRMS system used 70 eV EI ion source, 300 °C ion source temperature, mass range to 600 Th with 6000 FWHM resolution, and 100 spectra/s acquisition rate.

The resulting data for each chromatogram is an array of 1000 × 600 data points, each data point with a HRMS vector of 40 K intensities. Thus, each chromatogram has 24 billion values requiring 96 gigabytes for representation with single-precision floating point numbers without compression. The set of 18 chromatograms have more than 1.7 terabytes of uncompressed data. The data were compressed and stored by the Zoex FasTOF system to HDF5-format files and processed with GC Image GCxGC Software R2.1®. In order to process such large files on computers with limited random access memory (RAM), GC Image Software maintains a chromatogram with integer-mass or centroid resampled spectra in RAM and accesses the HRMS data from disk as needed. GC Image can export raw data and computed results to non-proprietary file formats for processing with external software.

Fig. 2 pictures the 18 chromatograms of the breast-cancer tumor samples. Column 1 (left) shows chromatograms of Grade 1 tumors, Column 2 (center) shows chromatograms of Grade 2 tumors, and Column 3 (right) shows chromatograms of Grade 3 tumors. The visualization uses pseudocolorization (with a cold-to-hot color scale) of the TIC. Chromatographic variations are visible (e.g., the larger detections in Column 1, Row 1, and Column 3, Row 4, and the larger late-time bleed in Column 1, Row 3, and Column 2, Row 1). In the data for each sample, thousands of compounds are separated by GCxGC and characterized by HRMS, providing a rich source of chemical information. Comprehensive analyses of large collections of such samples may yield biochemical features that are indicative of health conditions. Such biomarkers could indicate potential bases for diagnostic tests, provide insights into disease processes, and help researchers to develop treatments. New informatics are required for comprehensively analyzing such large collections of chemically complex samples for important and useful patterns.

2.2. Informatic methods

The informatic methods generate and apply a feature template that comprehensively captures the pattern of GCxGC–HRMS peaks for a set of samples and generates features that correspond (i.e., are matched) across chromatograms. The feature template consists of a few registration peaks used for chromatographic alignment and a set of retention-plane regions that are used to generate a vector of feature measurements for each chromatogram.

In broad terms, the method generates the feature template as follows.

I. Find peaks that are reliably matched across all samples, then create the registration template that records the pattern of those peaks.
II. Use the registration template to align each of the chromatograms, then sum the registered chromatograms to create a cumulative chromatogram.
III. Detect peaks in the cumulative chromatogram, then create a feature template that records both the registration peaks and the retention-times regions (footprints) of all peaks detected in the cumulative chromatogram.

Then, to analyze a chromatogram, the registration peaks in the feature template are matched to the detected peaks in the subject chromatogram. That matching defines a geometric transform in the retention-times plane that is applied to the feature-template regions, thereby maintaining the positions of the regions relative to the positions of the registration peaks matched in the subject chromatogram. Transforming the regions relative to the matching leaves the subject chromatogram unchanged. Then, the detector response within each region of the subject chromatogram is characterized, providing a vector of values for features that are matched across chromatograms.

The feature vectors for a set of chromatograms can be used to perform comparisons (e.g., reporting the absolute and/or relative differences between pairs or groups of chromatograms), cluster analysis (i.e., grouping relatively similar samples into the same group and relatively dissimilar samples into different groups), classifier training (i.e., building a classifier that identifies a category label for a given chromatogram based on information derived from a set of labeled examples), and biomarker identification (e.g., identifying features that are significant for cluster analysis or classification).

Fig. 3 pictures, in more detail, the operational flow of the informatic methods culminating in classification experiments. Steps 1–4 create the feature template for analyzing chromatograms. Step 5 analyzes chromatograms using the feature template. Steps 6 and 7 use the feature vectors for leave-one-out cross-validation classification experiments, but other cross-sample analyses (such as comparing two samples or clustering unlabeled samples) would use different operations following Step 5.

1. For each chromatogram, process the chromatogram to detect all peaks and represent those peaks in that chromatogram's template.
   (a) Correct any slowly varying, non-zero offset in the baseline signal. This operation models the baseline as a function of time based on data in chromatographic regions devoid of peaks and then subtracts the baseline function from the signal at each data point [8].
   (b) Detect blobs. This operation detects, delineates, and characterizes two-dimensional peaked clusters of data points in a chromatogram that are indicative of eluted compounds [6].
   (c) Create a Smart Template™. The template for a chromatogram records the pattern of its individual peaks, capturing information for identifying the same compounds in other chromatograms. For each peak, the template records the 2D retention times and a rule, expressed in the Computer Language for Identifying Chemicals (CLIC)™, that specifies the expected mass spectrum and the required NIST match factor [9]. The match factor required for identification is determined by analyzing the match factors with neighboring peaks (so that, for example, a peak among other peaks with similar mass spectra may require a higher match factor than a peak among other peaks with dissimilar mass spectra) [10].
2. Create the registration template with reliable peaks (peaks that can be detected and matched in all or most samples) and associated CLIC rules (constraints to promote correct matching).
   (a) This step begins by pattern matching the template from each chromatogram (resulting from Step 1) to all of the other processed chromatograms (also resulting from Step 1). In the matching (which uses retention times and mass-spectral matching rules), a template peak from one chromatogram matches at most one detected peak in each other chro-

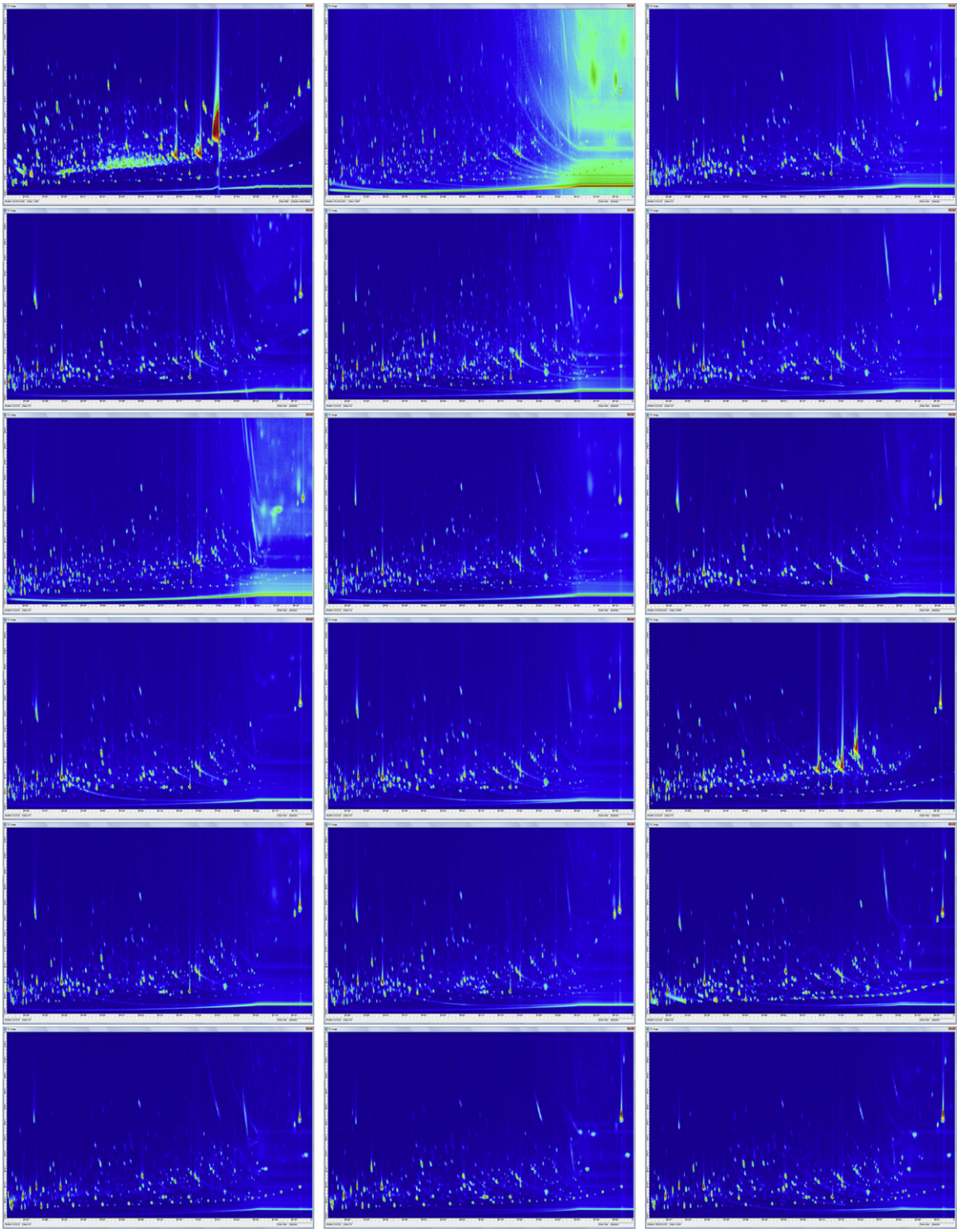

Fig. 2. GCxGC–HRMS chromatograms of breast-cancer tumors, six of Grade 1 (left), six of Grade 2 (center), and six of Grade 3 (right). *Samples courtesy of Dr. Oliver Fiehn, University of California, Davis, prepared by Zoex Corporation, and analyzed by Tofwerk AG.*

matogram. The matching is performed in each direction, so if there are N chromatograms, then the number of template matching operations is $N(N-1)$.

In a graph, the peaks can be represented as vertices and the peak matches can be represented as directed edges (i.e., a vector from one vertex to another, indicating a matching of a template peak from one chromatogram to a detected blob in another chromatogram). Each peak can have at most $N-1$ outgoing edges, with at most one edge to each of the other chromatograms, and at most $N-1$ incoming edges, with at most one edge from each other of the other chromatograms. Fig. 4, discussed subsequently in more detail, illustrates example pattern matchings between a few peaks in three chromatograms (A, B, and C).

(b) Determine the peaks that are reliably matched across all chromatograms. If Peak i in the template from Chromatogram A matches Peak j detected in Chromatogram B and Peak j in the template from Chromatogram B matches Peak i

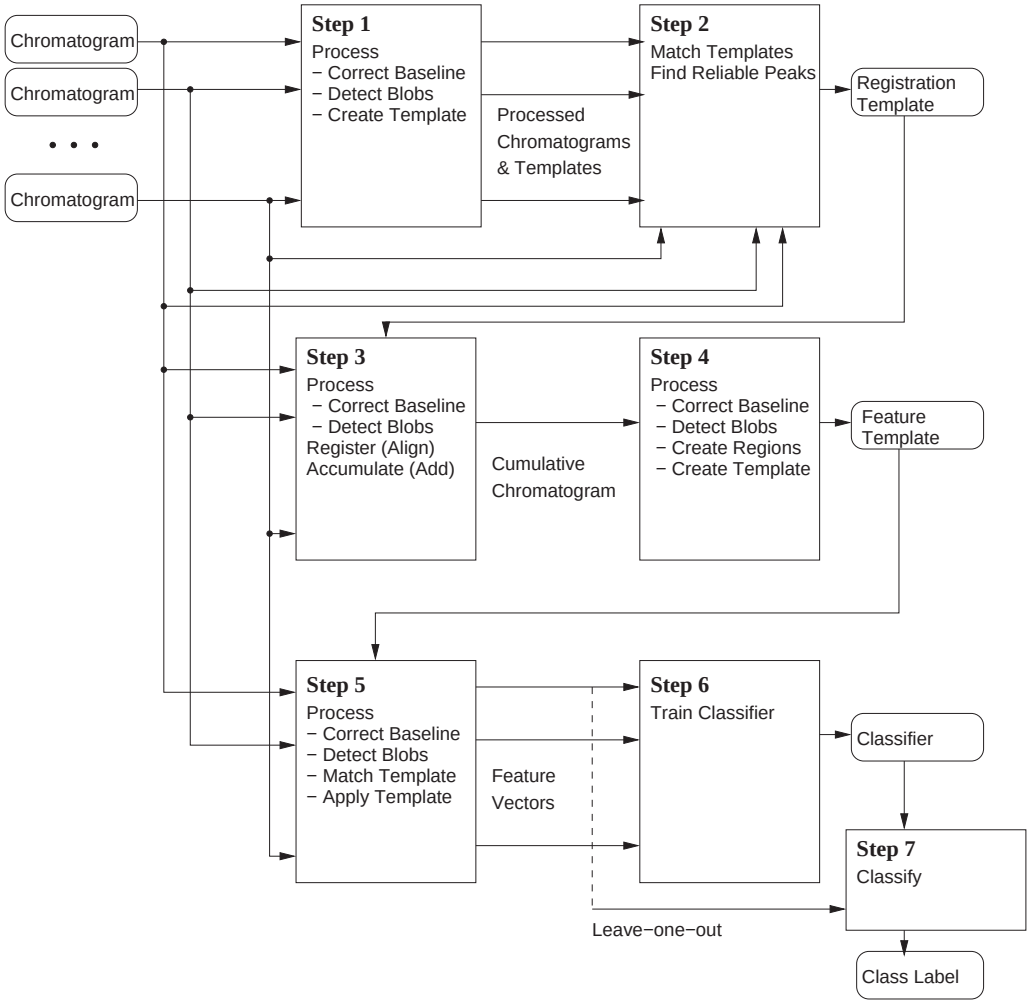

Fig. 3. Operational flowchart for cross-sample analysis (Steps 1–5) and classification experiments (Steps 6–7) with GCxGC–HRMS data.

detected in Chromatogram *A*, then the peaks are said to correspond. In the experiments described in Section 3, for each reliable peak, there must be a set consisting of one peak from each chromatogram such that each pair of peaks corresponds in their respective chromatograms.

In graph theory, this set is a bidirectionally connected clique with N vertices, where N is the number of chromatograms. In Fig. 4, Peak *A*.1 corresponds with Peak *B*.1 and with Peak *C*.1 and Peaks *B*.1 and *C*.1 correspond, so that peak is reliably matched across all three chromatograms. Peak *A*.2 corresponds with Peak *B*.2 and with Peak *C*.2, but Peaks *B*.2 and *C*.2 do not correspond, so these peaks do not reliably match across all three chromatograms. Peaks *A*.3 and *C*.4 correspond, but neither has a corresponding peak in Chromatogram *B*, so these peaks do not reliably match across all three chromatograms.

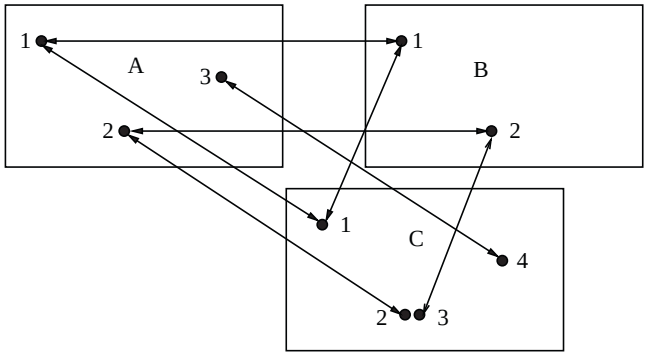

Fig. 4. Graph visualization of example peak matchings across three GCxGC chromatograms.

With the requirement of correspondences across all pairs of chromatograms, the number of matches required for a reliable peak is $N(N-1)$, so the number of reliable peaks tends to diminish as the number of chromatograms increases. Therefore, if the requirement for correspondence across all chromatograms results in two few reliable peaks for a set of chromatograms, it may be necessary to relax the requirement to be something less than complete bidirectional matching. In graph theory, a relaxed requirement might be for a bidirectionally connected clique with at least M vertices, where $M \leq N$.

(c) Create the registration template consisting of the reliable peaks. Here, a CLIC rule is created for each reliable peak by averaging the mass spectra of corresponding peaks and averaging their match-factor thresholds, but other approaches could be used to generate matching rules for the reliable peaks.

3. Create the cumulative chromatogram by aligning (registering) the individual chromatograms using the registration template and summing the aligned chromatograms.

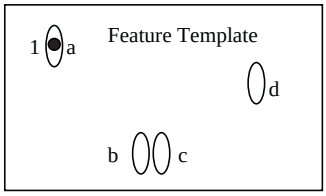

Fig. 5. Template with reliable peak (filled circle) and region features (open ovals) for the GCxGC example in Fig. 4.

(a) If the processed chromatograms are not available, then perform baseline correction and blob detection on the raw chromatograms as in Step 1.
(b) To align each chromatogram, the reliable peaks recorded in the registration template are matched to the detected peaks in the subject chromatogram. Then, the chromatogram is geometrically transformed in the retention-times plane to align with the template. This is a reversal of the usual template matching operation, in which the template is transformed to align with the chromatogram. Transforming the chromatogram to align with the template requires that the data be resampled. This operation modifies the data, which ordinarily is undesirable, but in this case, the transformed data is used only to create the cumulative chromatogram which is used to create the feature template. No other operation is performed with the transformed data.
(c) As each chromatogram is aligned with the registration template, compute the cumulative chromatogram as the pointwise sum of the individual registered chromatograms.

4. Create the feature template by adding retention-plane regions for all peaks detected in the cumulative chromatogram to the registration template.
(a) Perform baseline correction and blob detection on the cumulative chromatogram, as described for Step 1. Baseline correction of the cumulative chromatogram may not be necessary, because the individual contributing chromatograms are baseline corrected. However, if there is an accumulation of residual non-zero bias, this step can reduce or remove it.
(b) For each peak detected in the cumulative chromatogram, create an object that delineates the region in the retention-times plane occupied by the peak.
(c) Create the feature template by adding the region objects for all peaks to the registration template. Fig. 5 pictures the feature template for the example in Fig. 4, with one registration peak (which serves as an example, but more than one registration peak would be desired in practice) and four feature regions designated *a*–*d*. As will be illustrated in Section 3, for complex samples, most of the footprint retention-times regions are contiguous with other regions, providing coverage of much of the GCxGC plane.

At this point in the operational sequence, the informatics has determined a feature template that can be used to analyze sample chromatograms that are similar to those used to generate the feature template.

5. For each chromatogram to be analyzed, create a feature vector that characterizes the signal in the regions of the feature template.
(a) If the processed chromatogram is not available, then perform baseline correction and blob detection on the raw chromatogram as in Step 1.
(b) Match the registration peaks of the feature template to the detected peaks in the subject chromatogram. The matching uses both the retention-plane pattern and the mass-spectral matching rules of the template's registration peaks. Then, a geometric transform is applied to the template to align the matched peaks. Here, the geometric transform is a scale and translation that gives the least-squares difference between the retention times of the matched peaks after alignment, but more complex transforms, such as general affine or non-linear warping, could be used. Applying the geometric transform to both the registration peaks and feature regions in the feature template maintains the geometries of the regions relative to the registration peaks and brings them into proper alignment with the detected peaks in the subject chromatogram.
(c) Within each region, compute the characteristics of the subject chromatogram. In Section 3, each region is characterized by the total TIC summed over all data points in the region, but other characteristics, such as total SIC values, could be used.

Here, the process continues for leave-one-out cross-validation experiments which are reported in Section 3. Other cross-sample analyses would proceed differently from this point.

6. This step builds a classifier which, given a feature vector (such as is generated in Step 5), predicts the class label. For example, in Section 3, the class labels are Grade 1, Grade 2, and Grade 3, indicating the degree of cellular abnormality and predicting how quickly the tumor is likely to grow. Cancer pathologists grade samples from microscopic examination. The classifier predicts a label based on the feature vector.

The classifier is built from a set of labeled feature vectors, i.e., a class label is given for each vector. The set of labeled vectors used to build the classifier is called the training set. The process of building the classifier attempts to determine which features are indicative of the class label and the manner in which they are indicative so that the class label of an unlabeled sample can be predicted.

In leave-one-out cross-validation, the classification experiment is conducted once for each sample chromatogram. In each experiment, the data set is partitioned into a test set with just the subject chromatogram without its class label and a training set with all of the other chromatograms with their class labels. Then, a classifier is constructed based on the training set, according to whichever classification method is used.

7. A sample is classified by inputting its feature vector into the classifier which predicts its class label. In leave-one-out cross-validation, the class label of each test sample is known but not provided to the classifier (i.e., the classifier has only the feature vector). If the predicted class label is the same as the known class label, then the classifier is credited with a correct classification. The accuracy of the classifier is defined as the number of samples that are classified correctly divided by the number of samples that are classified.

Most of the operations for Steps 1–5 are available in the current version of GC Image GCxGC and LC×LC Software (R2.1, 2010), but executing the sequence of operations requires numerous user interactions and three operations – 2.b Find reliable Peaks, 3.b Register, and 4.b Create Regions – are not supported. The next version of the GC Image Software (R2.2, 2011) will fully support all operations for Steps 1–5, will provide a convenient interface for executing the sequence of operations, and implement classification for Steps 6–7.

3. Results

3.1. Features for cross-sample analysis

The experiments reported in this section demonstrate the informatics with the intention to present the process and methods of

cross-sample data analysis with GCxGC–HRMS. The breast-cancer tumor data set analyzed in these experiments is large enough for this demonstration, but is not large enough to provide a sufficient basis for conclusions about the metabolomics of cancer. Much larger sets of samples would be required to firmly establish biochemical characteristics of samples from breast-cancer tumors and to develop accurate and reliable methods for such cross-sample analyses as classifying samples and discovering biomarkers. Similarly, although results are reported for several classification algorithms with this data, the experiments are not sufficient to infer that their relative performance would be the same for larger studies. The process and methods illustrated here do provide a roadmap and tools for undertaking larger and more comprehensive investigations.

Fig. 6 visualizes the cumulative chromatogram for the 18 samples in the breast-cancer tumor data set, created as described in Section 2.2. Only the datapoints between 20 and 100 min for the first-column separation and between 2 and 6 s for the second-column separation are pictured. Each data point is pseudocolorized according to its TIC using a highly logarithmic value-mapping onto a cold-to-hot color scale. The color scale, shown to the right of the chromatographic image, illustrates the value-to-color mapping over the range of −0.01 to 7751.21. This color scale makes trace peaks visible, but obscures variations among data points with large TICs.

For the breast-cancer tumor data set, 13 registration peaks were identified by the method described in Section 2.2. The positions of the registration peaks in the retention-times plane are highlighted in Fig. 6 by dark ovals. As can be seen, the ranges of the registration peaks nicely cover the chromatographic region in which most peaks appear. As expected, most of the registration peaks are well-separated from neighboring peaks and so can be reliably detected and recognized across chromatograms.

In the cumulative chromatogram for the breast-cancer tumor data set, more than 3300 blobs were detected. The feature regions delineated by the footprints (retention-times regions) of those blobs are shown with red outlines in Fig. 6. Relatively low thresholds for blob detection were used, which results in extensive, but incomplete, coverage of the retention-times plane. Even if features are defined for false detections in noise regions, such features are extremely unlikely to be identified as candidate biomarkers and, if so identified, can be easily rejected. Inevitably, there are some blobs that appear to result from two analyte peaks and some analyte peaks that appear to be split into two blobs. These cases illustrate the unavoidable problems that make comprehensive peak matching across many chromatograms so intractable. Because the feature template uses regions extracted from the cumulative chromatogram, comprehensive cross-sample analysis can be performed consistently without comprehensive peak matching. After the template is aligned to match the registration peaks in the feature template to blobs detected in a subject chromatogram and the feature regions are transformed accordingly, the region-based features are consistently evaluated across samples, regardless of such peak detection issues.

The feature template illustrated in Fig. 6 was applied to each of the 18 sample chromatograms in the breast-cancer tumor data set. The result is a set of 18 feature vectors, one for each sample, each vector with more than 3300 feature values, each feature value characterizing the total response in that feature region of the individual chromatogram (where total response is the sum of the TIC at each data point in the retention-times region of the feature). For comparisons across chromatograms, the total response for each feature in a chromatogram was expressed as a percentage of the sum of the total responses for all features in the chromatogram. These feature vectors can be used for such cross-sample analyses as classification, discriminant analysis, clustering, etc. Here, the features of the breast-cancer tumor data set are used for classification experiments and Fisher ratio analysis to identify features of interest for further investigation by HRMS.

Table 1
Results of the most successful WEKA classification methods for the breast-cancer tumor samples.

Method	# Correct (of 18)	Accuracy	Confusion matrix			
			Class	Predicted		
Decision Table	14	77.78%		1	2	3
			1	5	1	0
			2	1	5	0
			3	1	1	4
Ordinal Class	12	66.67%		1	2	3
			1	3	2	1
			2	0	4	2
			3	0	1	5
Nested Dichotomies	11	61.11%		1	2	3
			1	3	2	1
			2	1	3	2
			3	0	1	5

3.2. Classification

The feature vectors for the 18 samples were used for leave-one-out cross-validation experiments. In leave-one-out cross-validation, one chromatogram (without its class label) constitutes the validation or test set and the other 17 chromatograms (with their class labels) comprise the training set. From the training set, the classification method builds a classifier to predict the class label for the unlabeled chromatogram. The experiment is repeated such that each chromatogram is used once for the validation set, so that there are 18 classifications in total. Overall classification accuracy is used to quantitatively measure the performance of the classification. Overall classification accuracy is defined as the number of correct classifications divided by the number of attempted classifications (here, 18). This is an ordinal classification problem, there are multiple classes with ordered labels – Grade 1, Grade 2, and Grade 3 – and the data set has an equal number of chromatograms for each class. So, a classifier that guessed randomly has an expected classification accuracy of 33%.

Several classification methods available in the WEKA collection of machine learning algorithms [11,12] were evaluated. A number of algorithms achieved classification accuracy of greater than 50%, i.e., 10 or more correct classifications of the 18 attempted. An algorithm that achieves 10/18 correct classifications for this problem performs at a level that would be achieved by random guessing with less than 5% (0.05) probability. Even better accuracy is even less likely: 11 or more, 1.44% (0.0144), 12 or more 0.39% (0.0039), 13 or more 0.09% (0.0009), and 14 or more 0.01% (0.0001). The performance of the most successful classification algorithms from the WEKA suite is shown in Table 1.

The WEKA Decision Table algorithm [13] builds a table of rules as a classification model based on a subset of the features with wrapper-based feature selection. The WEKA Ordinal Class Classifier [14] converts a k-class problem to $k-1$ binary class problems based on ordering information in class labels, then builds decision trees (here, C4.5, the WEKA J48 algorithm [15]) as the classification models. The WEKA Nested Dichotomies algorithm [16,17] is a meta classifier for handling multi-class datasets with 2-class classifiers (here, C4.5, the WEKA J48 algorithm [15]) by building a random tree structure.

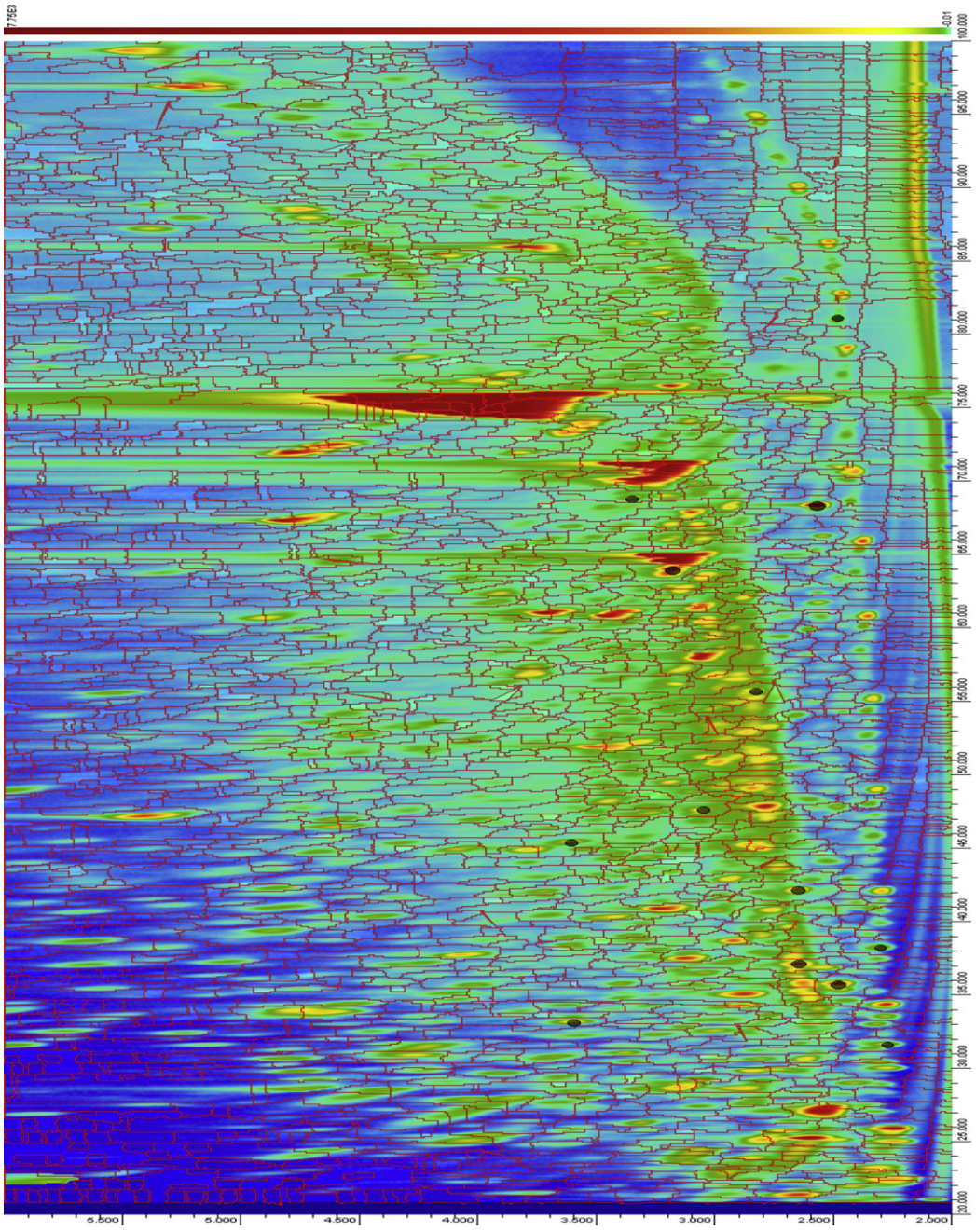

Fig. 6. Cumulative chromatogram for the 18 breast-cancer tumor samples overlaid with the feature template (registration peaks shown with dark ovals and region features shown with red outlines). The color bar shows the logarithmic pseudocolorization mapping.

3.3. Fisher ratio analysis

The Fisher ratio is a measure for linear discriminant analysis (LDA) and can be used to assess the ability of a feature to discriminate between classes:

$$S_{i,j} = \frac{\left(\mu_i - \mu_j\right)^2}{\sigma_i^2 + \sigma_j^2} \quad (1)$$

where μ_i and σ_i are the feature's mean and standard deviation for samples from class i. If the square of the difference between the class means for a feature is large relative to the sum of the class variances, then the feature is useful for discriminating between classes i and j.

For the breast-cancer tumor data set features, the Fisher ratios ranged up to 4.47 for distinguishing between grades. Fifteen features exhibited Fisher ratios of at least 1.5 for Grades 1 and 3; 14 features had Fisher ratios of at least 1.5 for Grades 2 and 3; and 10 features had Fisher ratios of at least 1.5 for Grades 1 and 2. Six features that had relatively large Fisher ratios (at least 1.3) for two or more pairs of grades are presented in Table 2. Features were numbered in order of peak apex intensities in the cumulative chromatogram and the feature identification numbers of all features with Fisher ratio at least 1.5 were small (less than 352) relative to the total number of features (more than 3300), indicating that they have relatively strong signal intensity. Table 3 shows the individual values and in-class (i.e., by grade) statistics for each of these discriminating features. These values are percent response, i.e., total

Table 2
Characteristics of six top discriminating features as indicated by the Fisher ratios.

	Feature identification number					
	208	297	239	224	351	91
RT1 (min)	76.59	71.65	68.88	62.71	25.99	22.58
RT2 (s)	4.04	3.83	3.21	3.19	2.53	2.30
$S_{1,2}$	4.47	2.44	1.71	1.86	0.00	0.00
$S_{1,3}$	1.31	1.36	2.10	1.66	1.80	1.71
$S_{2,3}$	0.18	0.82	0.14	0.01	1.81	1.53

response for a feature in a chromatogram as a percentage of the summed total responses for all features in the chromatogram.

3.4. HRMS analysis

Although the features are generated on the basis of peaks detected in the TIC of the cumulative chromatogram, the high-resolution mass spectra of features of interest can be examined to identify compounds, substructures, and elemental compositions [18–21]. For example, Fig. 7 shows the blob mass spectrum at (70.70 min, 3.16 s) from one of the samples (top, in blue) head-to-tail with the mass spectrum of stearic acid TMS from the Manchester Metabolomics Database [22] (bottom, in magenta). The two mass spectra are excellent matches: match factor 863, reverse match factor 866, and probability 78.9%. Fig. 8 shows the use of the high-resolution analysis to investigate elemental composition. The computed centroid of the HRMS peak at 341.279 Th (top, in blue) is within 0.1 milli-mass units of that expected for $C_{20}H_{42}O_2Si^+$ (top, overlaid, in green). The isotopic peaks (actual in blue and expected in green) also are excellent matches for mass and proportion.

In complex biological samples, compounds for many of the features will not have been documented in mass-spectral libraries. For such features, especially those that classification algorithms,

Table 3
Individual feature values and in-class (i.e., by grade) statistics for the six features in Table 2. Feature values are percent response, i.e., total response for a feature in a chromatogram as a percentage of the summed total responses for all features in the chromatogram.

Grade	Sample	Feature identification number					
		208	297	239	224	351	91
1	1	0.1168	0.0355	0.0327	0.0767	0.0042	0.0336
1	2	0.1169	0.0292	0.0606	0.0709	0.0130	0.1108
1	3	0.1304	0.0208	0.0567	0.0797	0.0107	0.0124
1	4	0.0498	0.0122	0.0364	0.0479	0.0104	0.0268
1	5	0.1343	0.0203	0.0282	0.0431	0.0103	0.0184
1	6	0.0826	0.0419	0.0092	0.0378	0.0188	0.0160
1	Mean	0.1051	0.0266	0.0373	0.0593	0.0112	0.0363
1	SD	0.0298	0.0100	0.0174	0.0168	0.0043	0.0340
2	1	0.0394	0.0120	0.0082	0.0255	0.0066	0.0078
2	2	0.0548	0.0121	0.0182	0.0475	0.0065	0.0091
2	3	0.0330	0.0110	0.0179	0.0403	0.0118	0.0449
2	4	0.0409	0.0092	0.0179	0.0264	0.0187	0.1352
2	5	0.0407	0.0109	0.0129	0.0273	0.0111	0.0328
2	6	0.0248	0.0103	0.0076	0.0358	0.0129	0.0085
2	Mean	0.0375	0.0109	0.0137	0.0337	0.0112	0.0396
2	SD	0.0115	0.0009	0.0045	0.0081	0.0041	0.0449
3	1	0.0163	0.0179	0.0149	0.0242	0.0285	0.2333
3	2	0.1131	0.0136	0.0101	0.0222	0.0253	0.2042
3	3	0.0839	0.0078	0.0087	0.0190	0.0329	0.3336
3	4	0.0365	0.0184	0.0162	0.0563	0.0369	0.0869
3	5	0.0439	0.0147	0.0120	0.0411	0.0616	0.2296
3	6	0.0163	0.0128	0.0088	0.0289	0.0106	0.0072
3	Mean	0.0530	0.0142	0.0117	0.0319	0.0326	0.1825
3	SD	0.0343	0.0035	0.0028	0.0129	0.0153	0.1064

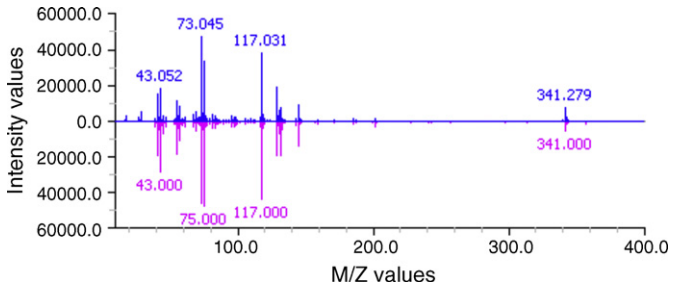

Fig. 7. Mass spectrum for a detected blob (top, in blue) head-to-tail with the Manchester Metabolomics Database mass spectrum for stearic acid [22] (bottom, in magenta). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

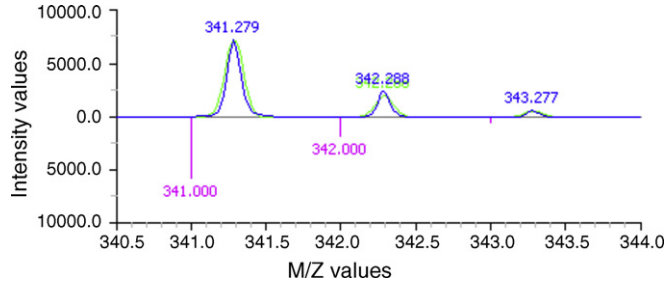

Fig. 8. Selected ion peaks from the blob mass spectrum shown in Fig. 7 (top, in blue) head-to-tail with the Manchester Metabolomics Database mass spectrum for stearic acid [22] (bottom, in magenta) and the expected isotopic peaks for $C_{20}H_{42}O_2Si^+$ (top, overlaid, in green).

Fisher ratios, and/or other methods indicate may be biomarker candidates, HRMS is especially helpful. For example, Fig. 9 shows the high-resolution mass spectrum for one of the features (Number 297, RT1 71.8 min, RT2 3.82 s) that is used in the Decision Table classification. Library searches with this mass spectrum did not yield good matches. Although thorough consideration of the biochemistry is beyond the scope of this paper and is the subject of ongoing work, examination of the HRMS peaks indicates likely elemental compositions of $C_4H_{10}NOSi^+$ for the peak at mass 116 and $C_5H_{11}NOSi^+$ for the peak at mass 129, suggesting the possible arrangement $CONHSi(CH_3)_3$ for the 116 fragment ion and $CHCONHSi(CH_3)_3$ for the 129 fragment ion [23].

4. Discussion

The primary contributions of this work are: (a) a methodological roadmap for comprehensive cross-sample analysis with GCxGC–HRMS and (b) nascent implementations of tools for

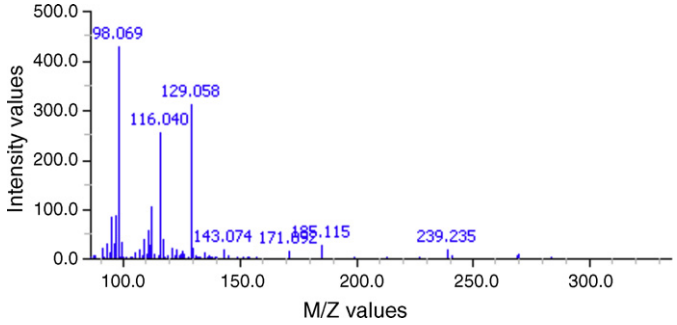

Fig. 9. The high-resolution mass spectrum of feature 297 from one of the samples.

performing such analyses. The approach avoids the intractable problem of comprehensive inter-sample peak matching by: (1) detecting and recording the peak patterns in individual chromatograms, (2) determining a few peaks that can be reliably matched across samples, (3) aligning and summing the sample chromatograms to create a cumulative chromatogram, and (4) defining a pattern of region features from the peaks detected in the cumulative chromatogram. Then, for analysis of a target chromatogram: (5) the registration peaks are matched to detected peaks in a target chromatogram to align the feature regions relative to those peaks and the characteristics of those features are computed to create a feature vector for the target chromatogram, and (6) the feature vector can be used for cross-sample analysis such as classification, discriminant analysis, clustering, etc.

This approach was demonstrated on a data set from GCxGC–HRMS analysis of breast-cancer tumor samples. The results indicate that feature vectors generated by this approach are useful for discriminating between samples of different grades (as labeled by a cancer pathologist) and may provide information that can be used to identify potential biomarkers for closer examination. One classification algorithm demonstrated accuracy that would result from random guessing with probability less than 0.01% (i.e., 99.99% confidence relative to the null hypothesis that the algorithm achieves its result by random guessing). HRMS is especially useful for investigating potentially significant features for chemical information.

Work continues on methods and tools to support this approach for cross-sample analysis. As a practical matter, graphical user interfaces are being designed and developed to make the methods easier to use. Also, using HRMS to further segment features on the basis of mass spectrometry may improve performance and robustness.

The experimental results indicate that the approach has great potential for cross-sample analyses, but the data set used in the experiments was relatively small. The approach should be applied to larger experimental data sets that could provide a solid foundation for more conclusive results.

The approach also would benefit from improvements of related technologies. In particular, the recent development of pulse calibration (the brief introduction of a calibrant during the void time of each secondary separation) will provide highly accurate mass calibration [5]. Soft ionization with GCxGC will provide more direct characterization of molecular composition. Negative chemical ionization (CI), positive CI, and electron ionization (EI) paired with GCxGC would provide highly complementary data. And, more extensive metabolomics databases would support better identification of sample constituents.

Disclosure

Dr. Qingping Tao is an employee of and has ownership in GC Image, LLC. Dr. Edward B. Ledford and Dr. Zhanpin Wu are employees of and have ownership in Zoex Corporation. Dr. Stephen E. Reichenbach has ownership in GC Image, LLC.

Acknowledgements

This work was supported in part by the U.S. National Science Foundation under Award Number IIP-1013180. The authors gratefully acknowledge Tofwerk AG (Thun, Switzerland), especially Marc Gonin, Christian Tanner, and Martin Tanner, for performing the GCxGC–HRMS analyses. The work on breast cancer metabolomics for Oliver Fiehn was funded through the MetaCancer Consortium project (European Union FP7 Health—project 200327) under the lead of Prof. Dr. Carsten Denkert (Charité Hospital, Berlin).

References

[1] W. Bertsch, Journal of High Resolution Chromatography 23 (2000) 167–181.
[2] L. Ramos, Comprehensive Two Dimensional Gas Chromatography, Elsevier, Oxford, UK, 2009.
[3] J. Phillips, et al., Journal of High Resolution Chromatography 22 (1999) 3–10.
[4] E.B. Ledford Jr., C.A. Billesbach, Journal of High Resolution Chromatography 23 (2000) 202–204.
[5] Zoex Corporation, FasTOF GCxGCxHiResTOFMS, http://www.zoex.com/products.html, 2010. U.S Patent (Application No. 61308519) and foreign counterparts pending.
[6] S.E. Reichenbach, Comprehensive Two Dimensional Gas Chromatography, Elsevier, Oxford, UK, 2009, pp. 77–106.
[7] O. Fiehn, G. Wohlgemuth, M. Scholz, T. Kind, D.Y. Lee, Y. Lu, S. Moon, B. Nikolau, Plant Journal 53 (2008) 691–704.
[8] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford Jr., Journal of Chromatography A 985 (2003) 47–56.
[9] S.E. Reichenbach, V. Kottapalli, M. Ni, A. Visvanathan, Journal of Chromatography A 1071 (2004) 263–269.
[10] S.E. Reichenbach, P.W. Carr, D.R. Stoll, Q. Tao, Journal of Chromatography A 1216 (2004) 3458–3466.
[11] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, CA, 2005.
[12] M. Hall, WEKA, http://www.cs.waikato.ac.nz/ml/weka/, 2010.
[13] R. Kohavi, Machine Learning: ECML 1995, vol. 912, Springer, 1995, pp. 174–189.
[14] E. Frank, M. Hall, L.D. Raedt, P. Flach, Machine Learning: ECML 2001, vol. 2167, Springer, New York, NY, 2001, pp. 145–156.
[15] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[16] E. Frank, S. Kramer, Twenty-first International Conference on Machine Learning, ACM, 2004, p. 39.
[17] L. Dong, E. Frank, S. Kramer, PKDD, Springer, 2005, pp. 84–95.
[18] T. Kind, O. Fiehn, BMC Bioinformatics 7 (2006) 234.
[19] T. Kind, O. Fiehn, BMC Bioinformatics 8 (2007) 105.
[20] F. Matsuda, Y. Shinbo, A. Oikawa, M.Y. Hirai, O. Fiehn, S. Kanaya, K. Saito, PLoS ONE 4 (2009) e7490.
[21] S. Abate, Y.G. Ahn, T. Kind, T.R.I. Cataldi, O. Fiehn, Rapid Communications in Mass Spectrometry 24 (2010) 1172–1180.
[22] M. Brown, W.B. Dunn, P. Dobson, Y. Patel, C.L. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst, A. Tseng, N. Swainston, I. Spasic, R. Goodacre, D.B. Kell, Analyst 134 (2009) 1322–1332.
[23] D. Swinton, S. Hattman, P.F. Crain, C.-S. Cheng, D.L. Smith, J.A. McCloskey, Proceedings of the National Academy of Sciences of the United States of America 80 (1983) 7400–7404.