Comprehensive feature analysis for sample classification with comrehensive two-dimensional LC

Research Article

Stephen E. Reichenbach[1]
Xue Tian[1]
Qingping Tao[2]
Dwight R. Stoll[3]
Peter W. Carr[4]

[1]Computer Science and Engineering Department, University of Nebraska – Lincoln, Lincoln, NE, USA
[2]GC Image, LLC, Lincoln, NE, USA
[3]Department of Chemistry, Gustavus Adolphus College, Saint Peter, MN, USA
[4]Department of Chemistry, University of Minnesota, Minneapolis, MN, USA

Received December 23, 2009
Revised January 25, 2010
Accepted January 26, 2010

Comprehensive two-dimensional LC (LC × LC) is a powerful tool for analysis of complex biological samples. With its multidimensional separation power and increased peak capacity, LC × LC generates information-rich, but complex, chromatograms, which require advanced data analysis to produce useful information. An important analytical challenge is to classify samples on the basis of chromatographic features, *e.g.*, to extract and utilize biomarkers indicative of health conditions, such as disease or response to therapy. This study presents a new approach to extract comprehensive non-target chromatographic features from a set of LC × LC chromatograms for sample classification. Experimental results with urine samples indicate that the chromatographic features generated by this approach can be used to effectively classify samples. Based on the extracted features, a support vector machine successfully classified urine samples by individual, before/after procedure, and concentration with leave-one-out and replicate *K*-fold cross-validation. The new method for comprehensive chromatographic feature analysis of LC × LC separations provides a potentially powerful tool for classifying complex biological samples.

Keywords: Classification / LC / Two-dimensional chromatography
DOI 10.1002/jssc.200900859

1 Introduction

Biochemical characteristics or biomarkers, such as metabolites in tissue, blood, urine, and other fluids, that are indicative of disease, environmental exposure, response to treatment, or other health-related conditions, have tremendous potential for improving public health [1]. For example, biochemical characteristics of urine may reflect the majority of pathological changes in human organs and urine can be collected non-invasively in large quantity [2], hence comprehensive separation and analysis of urine could provide an accessible wealth of information for healthcare. However, comprehensive separation and analysis of biological samples is a difficult challenge because of the presence of thousands of constituent compounds with highly variable concentrations and ranges and diverse physicochemical properties and detectability [3].

Correspondence: Dr. Stephen E. Reichenbach, 260 Avery Hall, Computer Science and Engineering Department, University of Nebraska – Lincoln, Lincoln, NE 68588-0115, USA
E-mail: reich@cse.unl.edu
Fax: +1-402-472-7767

Abbreviations: k-NN, *k* nearest neighbors; **LC × LC,** comprehensive two-dimensional LC; **PCA,** principal component analysis; **SVM,** support vector machine; **TIC,** total intensity count

Advanced instruments for biological separations are opening unprecedented vistas for biochemical analyses to discover and use biomarkers. In particular, comprehensive two-dimensional LC (LC × LC) [4] offers multidimensional separation power and increased peak capacity over one-dimensional HPLC [5]. Compared with data from one-dimensional HPLC, LC × LC provides more data points, an order-of-magnitude greater peak capacity, and added data dimensionality. LC × LC is a powerful tool for the separation of biological samples, but transforming the large and complex data generated by LC × LC into useful information is challenging. In a recent survey of fast LC × LC, Stoll *et al.* concluded that "[T]he paucity of effi-cient, convenient and sufficiently powerful data analysis tools [is] the greatest impediment to wide application of 2DLC" [6].

An especially important analytical challenge is classification of samples on the basis of non-target chromatographic features, *e.g.*, to extract and utilize biomarkers indicative of health conditions such as disease or response to therapy. Highly effective LC × LC separations are espe-cially well suited for this challenge because "combinations of biomarkers promise improved diagnostic performance over single markers, which may be lacking in sensitivity and/or specificity" [7]. However, the development of bioinformatic methods that use pattern-based approaches to simultaneously analyze the many chemical constituents of complex samples is a critical unsolved need for

metabolomics, metabonomics, proteomics, and other biological research [8–12].

Classification of samples requires corresponding "features" (such as peaks) across samples. If chromatographic peaks in different samples are determined to result from the same compound, then the measured responses or amounts for those peaks can be statistically characterized, compared, and used for the classification. A comprehensive feature set includes all sample-induced features of every sample – even those for unknown compounds and compounds present in some samples and not present in others. (Some chromatographic artifacts, such as column bleed, are not related to the sample and can be excluded from a comprehensive feature set.) The process of determining that features in different samples correspond, *e.g.*, are the result of the same compound, is feature matching. Automated feature matching of well-separated, well-formed peaks is relatively straightforward, but comprehensive feature matching of chemically complex samples is an extremely challenging problem.

A few research efforts have been made to develop chemometric methods and multivariate analysis techniques, aimed at determining features of two-dimensional chromatograms to quantitatively compare and classify complex homogenous samples. Johnson and Synovec [13] utilized analysis-of-variance-based feature selection to identify chromatographic features and principal component analysis (PCA) to classify jet fuel mixtures. The features were generated by point-by-point analysis of variance calculations, which provided an *f*-ratio for each data point. The data points with an *f*-ratio greater than a selected threshold were used as features. However, point-by-point feature analysis requires precise chromatographic alignment, which is difficult over large sample sets. Mispelaar *et al.* [14] used GC × GC peaks as features with principal component discriminant analysis to discriminate crude oils from different reservoirs, but reported that the results using all the peaks were highly unsatisfactory. Noting that the peak integration and matching errors were problematic, they used the average relative standard deviation between duplicate measurements to eliminate 90% of the peaks. However, doing so results in a non-comprehensive analysis that could miss useful information in the discarded peaks. (Using LC-MS data to classify urine samples, Kemperman *et al.* [2] similarly found that using PCA with all detected peaks was unsuccessful and hence used a nearest shrunken centroid algorithm to select a small number of peaks as features.) Aligning two-dimensional chromatograms for multivariate data or peak analysis is challenging [15–18].

This study presents a new approach for extracting comprehensive non-target chromatographic features for a set of two-dimensional chromatograms. The approach is to find enough peaks shared between chromatograms to form a pattern or *template*. Then, features are defined relative to the pattern. Each feature is the chromatographic region of a single peak or of multiple peaks, if the peaks in the region are difficult to unmix (or deconvolve). This approach does not require precise alignment because the features are defined relative to the pattern of peaks detected in each chromatogram and is less susceptible to peak integration and matching errors because peak unmixing is not required. In experiments with urine samples, the approach was used to extract features from a set of LC × LC chromatograms for several classification problems. Section 2 describes the data acquisition, processing, feature analysis, and classification of the urine samples. Section 3 presents the results and discussion of the classification experiments, including the different classification schemes, classification algorithms employed, and the evaluation methods. Section 4 contains concluding remarks, including consideration of the applicability of this approach to other types of detectors and other types of multidimensional chemical separations.

2 Data acquisition, processing and analysis

2.1 Acquisition

The urine data set analyzed in this study was acquired at the University of Minnesota in a series of 65 LC × LC analyses: (i) seven analyses of a standards mixture with potassium nitrate, tryptophan, hydroxytryptophan, indole-3-acetic acid, indole-3-propionic acid, indole-3-ACN, and tyrosine, interspersed in the series; (ii) 14 analyses of a control urine sample, interspersed in the series; (iii) four analyses of water, near the end of the series; and (iv) 40 analyses of experimental urine samples, of which four failed. For the urine analyses, a 460 µL aliquot of each urine sample was transferred to a HPLC vial. To each vial, 40 µL of 70% perchloric acid was added to precipitate proteins and this solution was allowed to stand for 10 min, followed by filtration with a small 0.2-µm PTFE syringe filter. The filtrate was collected in a new vial to which 55 µL of 10 M potassium hydroxide was added. This solution was centrifuged for 5 min to pellet the solid potassium perchlorate. The resulting solution was diluted to 9:10, 1:4, and 1:16 using 20 mM sodium phosphate, 0.1 mM EDTA, pH 6.

Experimental urine samples were provided by Dr. Todd Kellogg at the University of Minnesota. The 36 valid analyses of these samples included nine analyses each for persons A and B before a bariatric surgery and nine analyses each for persons A and B after the procedure. Each set of nine analyses included three analyses each diluted to 9:10, 1:4, and 1:16. For the control urine sample, the solution was diluted only to 9:10. Then, the urine samples were injected without further treatment.

In the dual gradient-elution system developed by Stoll *et al.*, shown in Fig. 1, the first column was a conventional gradient-elution HPLC system with reversed-phase LC column [5]. The effluent from the first column was captured alternately in loop 1 or loop 2 of the ten-port valve shown in the center of the figure. The stored effluent was injected into the second column (the second dimension of the separation)

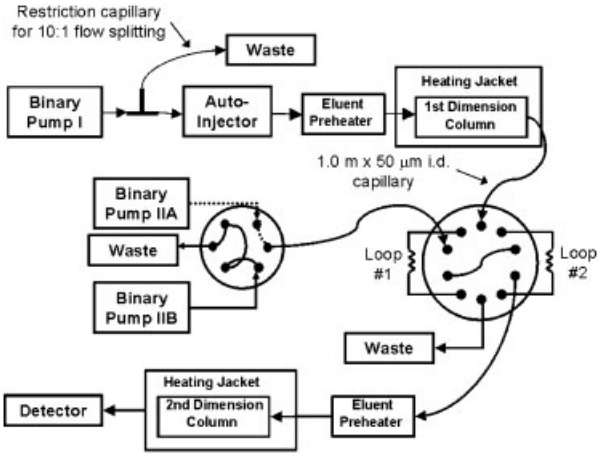

Figure 1. Instrumentation for comprehensive LC × LC [5].

and subjected to gradient elution by the dual gradient pumping system (pumps IIA and IIB). The rapid second separation used a short narrow column with high temperature (>100°C) and high flow rate (3 cc/min) to achieve very high linear velocity, allowing these separations to complete within 21 s. This is extraordinarily fast for LC and the resulting peaks are very narrow (<0.5 s half-height width). The two independent pumps and valve allowed switching between the two systems to minimize the effect of gradient dwell volumes. Otherwise, the chromatography would be slowed substantially and the retention-time reproducibility in the second dimension would be greatly compromised.

Although gradient elution in the second dimension is not as simple as isocratic elution, it is essential for three reasons. First, gradient elution gives higher peak capacity than isocratic elution. Secondly, a strong final eluent ensures that everything elutes before the next separation starts. Thirdly, gradient elution allows the diluted sample from the first dimension to be focused at the start of the second column, thereby improving the second dimension peak width when the first dimension system is delivering the analytes in strong eluent.

In these analyses, the gradient in the first column ran from 0 to 23 min, returned to the initial composition at 23.01 min, and was held until the end of the cycle (29.75 min). The first-column dead time was 1.0 min. The gradient in the second column ran from 0 to 18 s, returned to the initial composition at 18.6 s, and was held until the end of the cycle (21 s). The second-column dead time was 1.3 s.

The data were collected with a PhotoDiode array detector over the wavelength range 200–700 nm sampled in 4 nm intervals at 40 Hz for 29.75 min and written to a file by Agilent ChemStation software. The data for each analysis contains 71 400 data points, each with 126 spectral intensities, for a total of nearly 9 million intensities *per* analysis. As described in the following sections, the data for each analysis was read from the ChemStation UV file, restructured as a series of 65 secondary chromatograms, each 21 s long, and processed for baseline correction, peak detection, and chromatographic feature construction with GC Image® LC × LC software (http://www.gcimage.com) [19].

2.2 Visualization and processing

2.2.1 Visualization

The output of each LC × LC analysis can be displayed as a two-dimensional image (the LC × LC chromatogram). Figure 2 illustrates an image of one of seven analyses of the standards mixture. The *x*-axis (left-to-right) represents the retention time in minutes for the first column separation. The *y*-axis (bottom-to-top) represents the retention time in seconds for the second column separation. In Fig. 2, the pixels are pseudocolorized automatically with gradient-based value mapping to a cold–hot color scale commonly used for topographic mapping [20, 21]. The color of each pixel in Fig. 2 is determined by the value of the total intensity count (TIC) of the UV spectral absorbance at the indicated first and second dimension retention times. The UV TIC is the sum of the responses, measured in milli-absorbance units, in all spectral channels (just as the total ion count is the sum of the intensities for MS). Alternatively, the selected intensity count could be computed for a subrange of the spectrum and used for pseudocolorization.

2.2.2 Baseline correction

Each individual chemical compound forms a two-dimensional cluster of pixels (*i.e.*, a peak) with values larger than the background values (*i.e.*, the data values in which no chemical peak is present). Figure 2 shows that the

background values vary greatly across the second column separations and to a lesser extent across the first column separation. The dynamic range of the background obscures peaks, hence the baseline values (*i.e.*, the slowly varying mean background level) must be removed for accurate peak detection (described next) and possibly quantification (not performed with this UV data).

Baseline correction is performed with the LC × LC baseline correction algorithm developed by Reichenbach *et al.* [21, 22]. The approach estimates the baseline values across the chromatogram based on structural and statistical properties of data and then subtracts the baseline estimate from the data at each point. First, background regions are located. Then, background statistics are computed in each background region in every spectral interval. Next, local filters are applied to the estimated statistics to reject outliers. Then, the baseline values across all data points at each spectral interval are interpolated from the filtered, local background statistics. Finally, at each data point and each spectral interval, the estimated baseline value is subtracted from the data. Figure 3 illustrates the image after baseline correction. The resulting background values across the center of the LC × LC chromatogram are near zero and the chemical peaks are clearer against the more uniform background.

2.2.3 Peak detection

The chemical peaks are detected in two dimensions using the drain algorithm [22], a modified and inverted version of the watershed algorithm [23], on the LC × LC TIC. The drain algorithm detects peaks from the top, down to the surrounding valleys, in two dimensions, with user-defined minimum thresholds on the chromatographic footprint (*i.e.*, the temporal area, which is the two-dimensional analog of peak width), apex value (the largest TIC in the peak), and total TIC (summed intensities for all data points in the peak). Figure 3 illustrates the chemical peaks detected in the standards mixture. The footprint regions of the detected peaks are outlined. The peaks of interest are outlined in black. Other peaks caused by artifacts and which are not in the region of analytical interest are outlined in gray. Figure 4 illustrates a three-dimensional perspective view of the detected chemical peaks in the center of the image. The detected peaks rise clearly above the background.

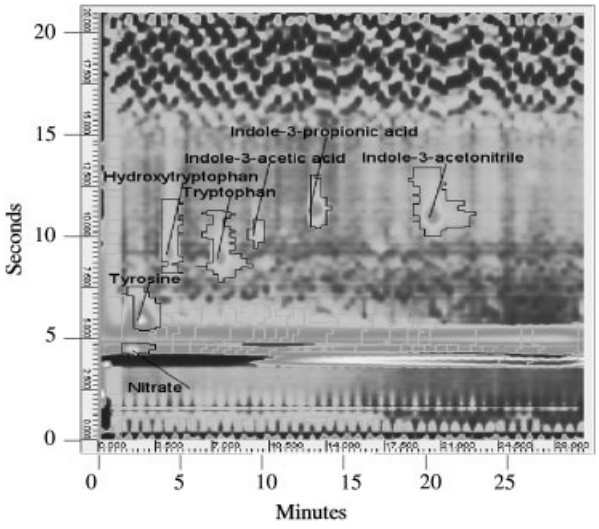

Figure 3. After baseline correction, the resulting background values across the center of the image are near zero and the chemical peaks are clearer against the more uniform background.

2.2.4 Chromatographic feature construction

The goal of the new method is to determine a set of well-defined LC × LC features that (i) account for all the sample-induced aspects of every chromatogram in the data set and (ii) provide corresponding measurements across all samples. Then, the set of feature values computed for each chromatogram provides a representation of that chromatogram which can be used for classification analysis (described in the next section).

In a single chromatogram, the detected peaks can account for all aspects of the sample, but making correspondences between all the peaks across many samples is fraught with problems. For example, slightly different chromatographic data may be detected as one peak in one chromatogram and as multiple peaks in another. Then, it often is difficult to correctly determine whether one chromatogram lacks compound(s) present in the other or whether peak detection fails to unmix co-eluted peaks in one of the chromatograms and succeeds in the other. Incorrect correspondences for features in different chromatograms can obscure true comparisons and lead to incorrect classifications. Manual processing may be able to correct some such errors, but is tedious for even a few chromatograms and is impractical for large sample sets.

The solution embodied in the method described here is to segment chromatograms into features for which more reliable automated correspondences can be made. In this approach, aspects of the data that cannot be disambiguated

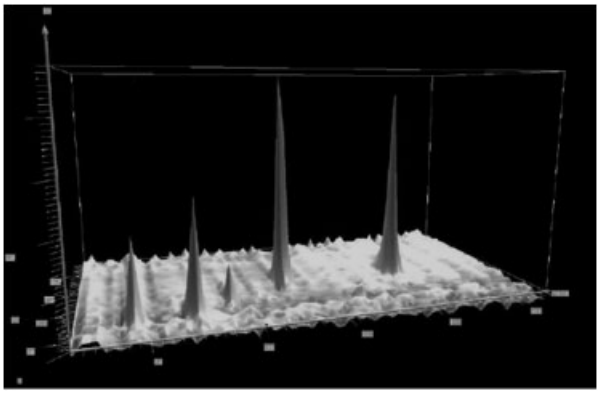

Figure 4. A three-dimensional perspective view of the detected peaks in the center of the LC × LC chromatogram.

reliably are treated as single features. The approach can be successful because the multidimensional separation power of LC × LC typically provides a rich source of reliable characteristics.

The steps for constructing a comprehensive set of reliable chromatographic features for a set of chromatograms are:

(i) *Consensus peaks template.* Identify *consensus peaks*, which are corresponding peaks present in all (or most) chromatograms. The set of consensus peaks need not be comprehensive, but should include peaks across the retention-time plane. Create a *template* that records the average retention time of each consensus peak along with (optionally) a rule to help distinguish that peak from other peaks. GC Image software includes a tool for building templates with such rules [19]. For these experiments, peaks identified by visual inspection were used as the consensus peaks [21]. The 98 consensus peaks in the template for the urine samples are illustrated in Fig. 5, overlaid on one of the chromatograms (shown with a different color scale).

(ii) *Cumulative chromatogram.* Match the template of consensus peaks (from step i) to the detected peaks in each chromatogram. Template matching [24] is based on the presumption that the chromatographic peaks form a pattern (template) that can be recognized from one chromatogram to the next even if not all the peaks in the pattern are detected in each chromatogram. In template matching, the peaks in the template are matched to (paired with) detected peaks in the chromatogram. Then, use the matching to align each chromatogram to the template. The alignment is performed by the global affine transformation that minimizes the residual squared distance from the template peaks to the matched detected peaks, followed by nearest-neighbor interpolation. Sum the aligned chromatograms to form a *cumulative chromatogram.* Figure 6 illustrates the cumulative chromatogram of the experimental urine sample analyses. Note that alignment need not be perfect – the cumulative chromatogram is intended to account for chromatographic variability relative to the consensus peaks pattern.

(iii) *Feature template.* Perform peak detection for the cumulative chromatogram TIC. For each detected peak, use its footprint in the retention-time plane to define a feature for the analysis. The retention-time footprint is the two-dimensional analog of a retention-time window and hence the feature object is an irregularly shaped region (or retention-time area) rather than a one-dimensional retention-time range. Then, add these feature objects (regions) to the consensus peak template to form a *feature template.* Figure 7 illustrates the feature template with consensus peaks and 98 feature objects for the urine samples overlaid on the cumulative chromatogram.

(iv) *Feature computations.* Match the peaks in the feature template to the detected peaks in each chromatogram. Then, for each chromatogram, align the template to the chromatogram using the global affine transforma-

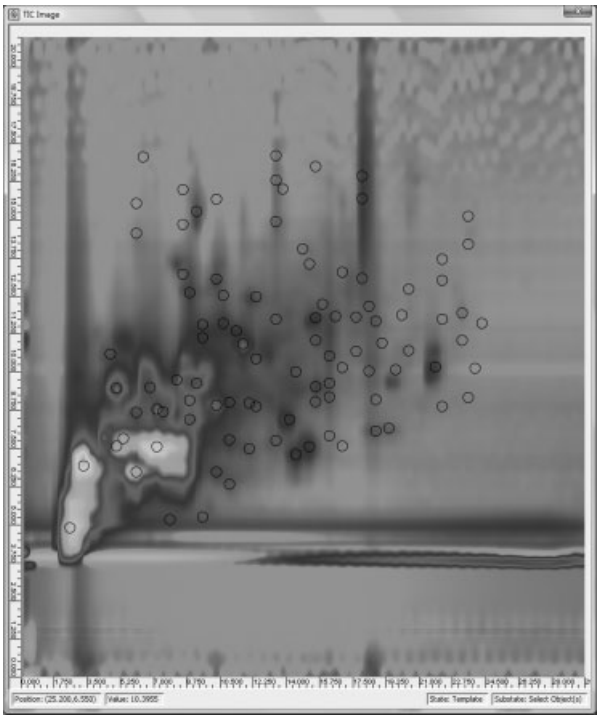

Figure 5. The template with consensus peaks overlaid on one of the LC × LC chromatograms of the urine samples.

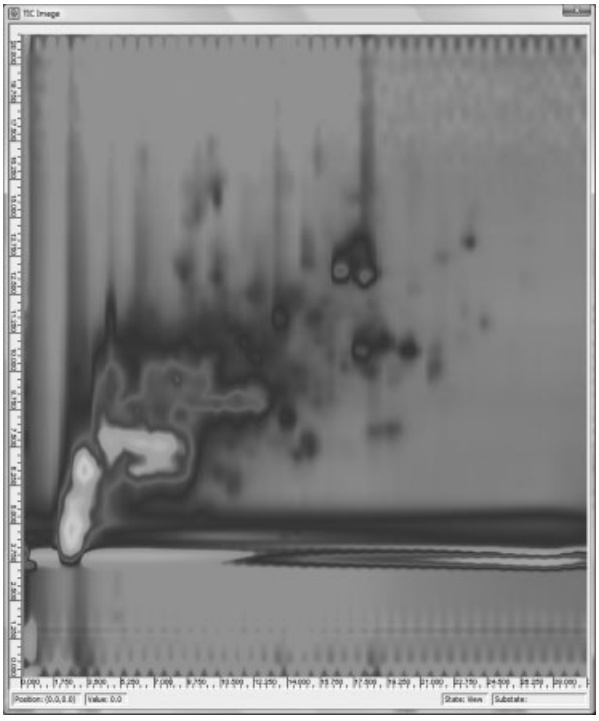

Figure 6. The cumulative chromatogram for the urine samples.

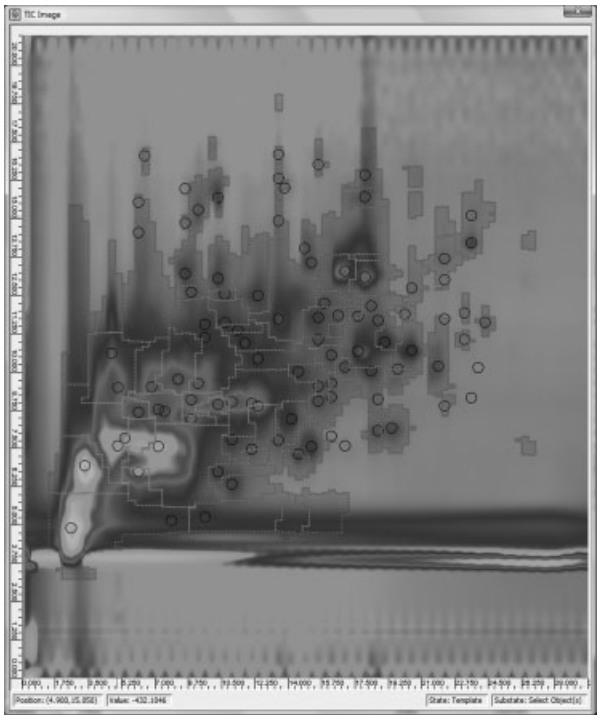

Figure 7. The feature template with consensus peaks and feature objects overlaid on the cumulative chromatogram for the urine samples.

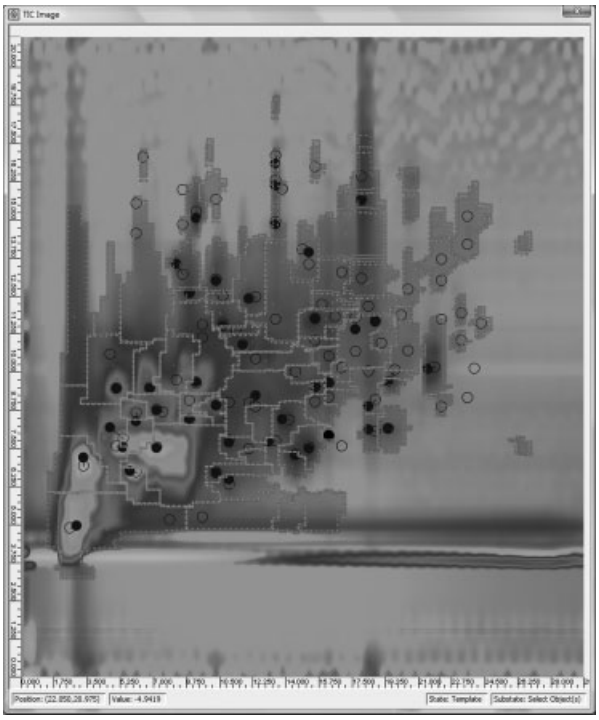

Figure 8. Matching of the feature template to one of the LC × LC chromatograms of the urine samples.

tion that minimizes the residual squared distance from the template peaks to the matched detected peaks. Unlike step (ii), this step transforms the template (rather than the chromatogram). This alignment of the feature template geometrically transforms the features to fit the pattern of peaks (and leaves the chromatographic data unchanged). For each chromatogram, with each transformed feature, compute the feature value as the total TIC within the feature region. Then, normalize the feature values for each chromatogram (*e.g.*, relative to an internal standard if one is available). For these urine samples, no internal standard was used, hence the feature values are normalized by dividing by the sum of all the feature values in the chromatogram so that each feature is a fractional response. Figure 8 illustrates a matching of the feature template to a urine sample chromatogram. Matched detected peaks are shown by filled circles and the transformed positions of the feature regions relative to the detected peaks are shown graphically.

2.3 Classification

2.3.1 Experimental data sets

The urine sample chromatograms are organized in two different experimental sets to accommodate different classification experiments. The first experimental set has 36 urine sample chromatograms including nine sample analyses each for persons A and B before bariatric surgery (the procedure), and nine sample analyses each for persons A and B after the procedure. The second experimental set has the 36 chromatograms in the first experimental set plus three chromatograms for the control urine samples.

2.3.2 Class descriptions

The first experimental set is analyzed by three different classification schemes: (i) two classes by individual, (ii) two classes by before/after procedure, and (iii) four classes by individual and before/after procedure. In Scheme 1, the first class includes 18 chromatograms for person A and the second class includes 18 chromatograms for person B. This scheme is illustrated in Fig. 9. In Scheme 2, the first class includes 18 chromatograms for before the procedure and the second class includes 18 chromatograms for after the procedure. This scheme is illustrated in Fig. 10. In Scheme 3, the first class includes nine chromatograms for person A before the procedure, the second class includes nine chromatograms for person B before the procedure, the third class includes nine chromatograms for person A after the procedure, and the fourth class includes nine chromatograms for person B after the procedure. This scheme is illustrated in Fig. 11.

The second experimental set is analyzed by two different classification schemes: (i) three classes by concentration and (ii) thirteen classes by individual, before/after procedure, and concentration. In Scheme 1, the first class includes 15 chro-

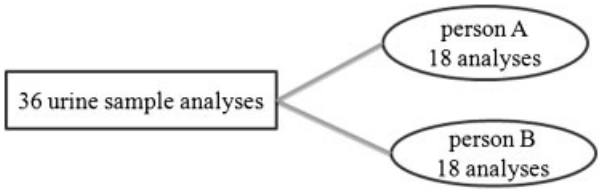

Figure 9. Experimental Set 1, Classification Scheme 1: two classes by individual.

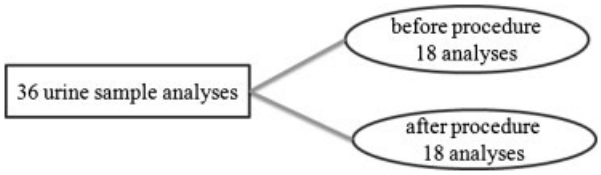

Figure 10. Experimental Set 1, Classification Scheme 2: two classes by before/after procedure.

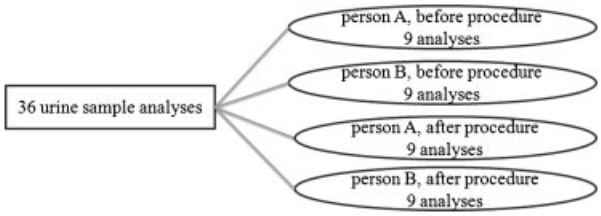

Figure 11. Experimental Set 1, Classification Scheme 3: four classes by individual and before/after procedure.

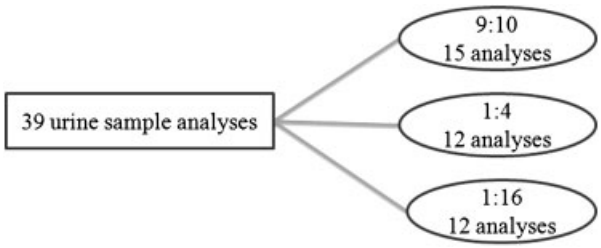

Figure 12. Experimental Set 2, Classification Scheme 1: three classes by concentration.

matograms for samples diluted 9:10, the second class 12 chromatograms for samples diluted 1:4, and the third class 12 chromatograms for samples diluted 1:16. This scheme is illustrated in Fig. 12. In Scheme 2, each class includes three chromatograms: Classes 1–3 contain chromatograms for person A before the procedure with samples diluted 9:10, 1:4, and 1:16, respectively; Classes 4–6 contain chromatograms for person B before the procedure with samples diluted 9:10, 1:4, and 1:16, respectively; Classes 7–9 contain chromatograms for person A after the procedure with samples diluted 9:10, 1:4, and 1:16, respectively; Classes 10–12 contain chromatograms for person B after the procedure with samples diluted 9:10, 1:4, and 1:16, respectively; and Class 13 contains chromatograms for the control sample (diluted to 9:10). This scheme is illustrated in Fig. 13.

2.3.3 Classification algorithms

Support vector machines (SVMs) and *k*-nearest neighbors (*k*-NN) are the classification algorithms used in the classification analysis.

SVMs are learning systems that use a hypothesis space of linear functions in a high-dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [25]. SVMs have empirically good performance and have successful applications in many fields (bioinformatics, text recognition, image recognition, *etc.*). The sequential minimal optimization algorithm [26] from the WEKA data mining system [27] is employed to build an SVM with a degree-one polynomial kernel.

The *k*-NN algorithm [28] retrieves a set of *k* training samples closest to a query sample (the *k* nearest neighbors), and classifies the query sample according to a majority vote of the neighbor class labels. It has been used in applications of data mining, statistical pattern recognition, image processing, and many other fields. Some successful applications include recognition of handwriting, satellite images, and electrocardiogram patterns. The nearest-neighbor algorithm (*k*-NN with $k = 1$) from the WEKA data mining system [27] is employed to build a *k*-NN classifier with normalized Euclidean distance.

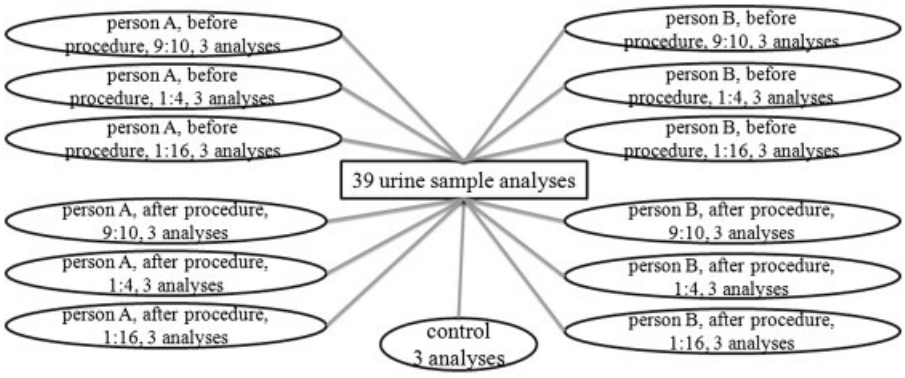

Figure 13. Exprimental Set 2, Classification Scheme 2: 13 classes by individual, before/after procedure, and concentration.

Table 1. Classification performance for leave-one-out cross-validation

	SVM		*k*-NN	
	Accuracy (%)	Confusion matrix	Accuracy (%)	Confusion matrix
Set 1, Scheme 1	100.0	18 0 0 18	100.0	18 0 0 18
Set 1, Scheme 2	100.0	18 0 0 18	100.0	18 0 0 18
Set 1, Scheme 3	100.0	**9** 0 **0** 9 **0** 0 **0** 0	100.0	**9** 0 **0** 9 **0** 0 **0** 0
Set 2, Scheme 1	97.44	15 0 0 0 11 1 0 0 12	100.0	15 0 0 0 12 0 0 0 12
Set 2, Scheme 2	97.44	**3** 0 **0** 0 **0** 0 **0** 2 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **3** 0 **0** 0 **0** 0 **0** 3 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **3** 0 **0** 0 **0** 0 **0** 3 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0 **0** 0	100.0	**3** 0 0 **0** 0 **0** 0 **0** 3 0 **0** 0 **0** 0 **0** 0 0 **0** 0 **0** 0 **0** 0 0 **0** 0 **0** 0 **0** 0 3 **0** 0 **0** 0 **0** 0 0 **3** 0 **0** 0 **0** 0 0 **0** 3 **0** 0 **0** 0 0 **0** 0 **0** 0 **0** 0 0 **0** 0 **0** 0 **0** 0 0 **0** 0 **0** 0 **0** 0 0 **0** 0 **3** 0 **0** 0 0 **0** 0 **0** 3 **0** 0 0 **0** 0 **0** 0 **0** 0 0 **0** 0 **0** 0

2.3.4 Evaluation

Leave-one-out cross-validation and replicate *K*-fold cross-validation are used for testing. In leave-one-out cross-validation, one urine sample analysis from the data set is used as the validation data and the remaining sample analyses are used as the training data. This is repeated such that each sample analysis in the data set is used once as the validation data. In replicate *K*-fold cross-validation, the data set is divided into *K* partitions, here one partition for each set of replicates. Of the *K* partitions, a single partition is retained as the validation data and the remaining *K*-1 partitions are used as training data. This cross-validation process is then repeated *K* times (the folds), with each of the *K* partitions used exactly once as the validation data.

Overall classification accuracy is used to quantitatively measure the performance of the classification. Overall classification accuracy is defined as:

$$\text{Accuracy} = \frac{\sharp \text{ of sample analyses classified correctly}}{\sharp \text{ of sample analyses in the data set}} \quad (1)$$

3 Results

The two experimental sets are classified by SVM and *k*-NN for each of the different classification schemes.

Table 1 illustrates the overall classification accuracy and the confusion matrices of the two classifiers with leave-one-out cross-validation. For the first experimental set (without the control urine samples), SVM and *k*-NN classify with 100% accuracy by: (i) individual, (ii) before/after procedure, and (iii) individual and before/after procedure. For the second experimental set (with control urine samples), *k*-NN classifies with 100% accuracy by: (i) concentration and (ii) individual, before/after procedure, and concentration. For both classification schemes for the second experimental set, SVM misclassified one chromatogram, slightly lower accuracy but comparable with *k*-NN and the difference is not statistically significant. Each type of urine sample has three replicate analyses and leave-one-out cross-validation includes two replicate analyses of the validation sample in the training set, hence the classification success is not surprising, especially for *k*-NN. Replicate *K*-fold cross-validation is used to address this issue.

Table 2 illustrates the overall classification accuracy and the confusion matrices of the two classifiers with replicate *K*-fold cross-validation. The classifications for the first experimental set use 12-fold cross-validation. Each of the 12 folds includes three replicate analyses of a urine sample for a specific individual, before/after procedure, and concentration. The classifications for the second experimental set use 13-fold cross-validation, with the 12 folds in the first experimental set plus a fold for the replicate analysis of the

Table 2. Classification performance for *K*-fold cross-validation

	SVM		*k*-NN	
	Accuracy (%)	Confusion matrix	Accuracy (%)	Confusion matrix
Set 1, Scheme 1	88.89	17 1 3 15	61.11	9 9 5 13
Set 1, Scheme 2	97.22	17 1 0 18	94.44	16 2 0 18
Set 1, Scheme 3	97.22	**9** 0 **0** 9 **0** 0 **0** 0	58.33	**9** 0 **0** 8 **0** 0 **0** 0
Set 2, Scheme 1	87.18	15 0 0 **0** 9 0 2 10	41.03	4 11 0 **4** 5 **0** 5

Table 3. Classification performance of SVM with PCA for *K*-fold cross-validation

	Accuracy (%) for SVM		
	3 principal components	10 principal components	All features
Set 1, Scheme 1	50.00	86.11	88.89
Set 1, Scheme 2	100.00	97.22	97.22
Set 1, Scheme 3	50.00	83.33	97.22
Set 2, Scheme 1	17.95	23.08	87.18

control urine sample (with 9:10 concentration). Scheme 2 (by individual, before/after procedure, and sample concentration) for the second experimental set (with the control urine sample) cannot be evaluated because all members of a class would be excluded with a fold, thereby rendering training impossible.

As shown in Table 2, for the replicate *K*-fold cross-validation experiments, SVM outperforms *k*-NN. Although *k*-NN has 94.44% accuracy for Set 1, Scheme 2 (by before/after procedure), the performance of *k*-NN for the other classification schemes is much lower. For example, for Scheme 1 (by individual) for the first experimental set, the accuracy for *k*-NN was only 61.11%, which is not statistically significant (at 95% confidence level) relative to the null hypothesis of random guessing, whereas the accuracy for SVM was 88.89%, which is statistically significant (at 99.999999% confidence level) relative to the null hypothesis of random guessing. The differences for Set 1 (without control urine sample), Scheme 2 (by before/after procedure) are not statistically significant, but the differences between SVM and *k*-NN for the other classification schemes are statistically significant. The accuracy for *k*-NN classification suffers without available replicate training analyses for the validation data, whereas SVM is successful for all the classification schemes for both the experimental sets.

The utility of comprehensive feature sets for these experiments is supported by results for PCA with SVM for this feature set. For Set 1 (without control urine samples), Scheme 1 (by individual), and replicate *K*-fold cross-validation, the accuracy of SVM with the first three principal components is only 50.00% compared with 88.89% with all the features. With the first ten principal components, accuracy of SVM is 86.11%. For Set 2 (with control urine samples), Scheme 1, the accuracy of SVM with the first three principal components is 17.95% and with the first ten principal components is only 23.08%, compared with 87.18% with all the features. Results of SVM with PCA for the replicate *K*-fold cross-validation experiments are shown in Table 3.

4 Concluding remarks

This study develops a new method to extract comprehensive non-target chromatographic features from a set of two-dimensional chromatograms for sample classification. The method defines a set of chromatographic regions relative to a pattern of peaks and hence is relatively robust with respect to compositional differences among samples, chromatographic variations, and co-eluted peaks. The method was demonstrated on a set of LC × LC chromatograms for urine samples. After the features were extracted, two different classification methods, SVMs and *k*-NN, were evaluated for several different classification scenarios using leave-one-out and replicate *K*-fold cross-validation. Experimental results suggest that the method produces not only

feature sets that can be used successfully for classification, but also indicate that performance varies for different classifiers. Performance of SVM classification with PCA suggests that comprehensive feature sets provide more complete information for classification. The experiments involved only a few dozen chromatograms, hence more definitive conclusions require additional research and development. Ongoing research involves classification of cancerous tissue samples analyzed by GC × GC and incorporation of spectral information, including MS. A significant need for such work is the generation of two-dimensional chromatograms for large sets of clinically relevant samples.

The authors gratefully acknowledge the contribution of urine samples from individuals before and after bariatric surgery by Todd Kellogg, MD, at the University of Minnesota Center for Minimally Invasive Surgery. Research at the University of Nebraska was supported by National Science Foundation funding to S. E. Reichenbach (IIS-0431119) and research at GC Image, LLC, was supported by the National Institutes of Health National Center for Research Resources (RR020256). Dr. Stoll acknowledges support from the Camille and Henry Dreyfus Foundation Faculty Start-up Award program.

Dr. Qinping Tao is an employee of GC Image, LLC, and Professor Stephen E. Reichenbach and Dr. Tao have financial interests in GC Image, LLC.

The authors have declared no conflict of interest.

5 References

[1] Qian, W., Jacobs, J., Liu, T., Camp, D., Smith, R., *Mol. Cell. Proteomics* 2006, *5*, 1727–1744.

[2] Kemperman, R. F. J., Horvatovich, P. L., Hoekman, B., Reijmers, T. H., Muskiet, F. A. J., Bischoff, R., *J. Proteome. Res.* 2007, *6*, 194–206.

[3] Eggink, M., Romero, W., Vreuls, R. J., Lingeman, H., Niessen, W. M. A., Irth, H., *J. Chromatogr. A* 2008, *1188*, 216–226.

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

[5] Stoll, D., Cohen, J., Carr, P., *J. Chromatogr. A* 2006, *1122*, 123–137.

[6] Stoll, D., Li, S., Wang, S., Carr, P., Porter, C., Rutan, S., *J. Chromatogr. A* 2007, *1168*, 3–43.

[7] Etzioni, R., Kooperberg, C., Pepe, M., Smith, R., Gann, P., *Biostatistics* 2003, *4*, 523–538.

[8] Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., Kovach, J., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 14666–14671.

[9] Wagner, M., Naik, D., Pothen, A., Kasukurti, S., Ram, R., Bao-Ling, D., Semmes, O., Wright, G., Jr., *BMC Bioinf.* 2004, *5*, 1–9.

[10] Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T., Foss, E., Mao, Y., Emili, A., *Mol. Cell. Proteomics* 2004, *3*, 984–997.

[11] Vlahou, A., Fountoulakisa, M., *J. Chromatogr. B* 2005, *814*, 11–19.

[12] Ulintz, P., Zhu, J., Qin, Z., Andrews, P., *Mol. Cell. Proteomics* 2006, *5*:497–509.

[13] Johnson, K. J., Synovec, R. E., *Chemom. Intell. Lab Syst.* 2002, *60*, 225–237.

[14] Mispelaar, V. G., van Smilde, A. K., Noord, O. E. de Blomberg, J., Schoenmakers, P. J., *J. Chromatogr. A* 2005, *1096*, 156–164.

[15] Pierce, K. M., Wood, L. F., Wright, B. W., Synovec, R. E., *Anal. Chem.* 2005, *77*, 7735–7743.

[16] Porter, S. E. G., Stoll, D. R., Rutan, S. C., Carr, P. W., Cohen, J. D., *Anal. Chem.* 2006, *78*, 5559–5569.

[17] Zhang, D., Huang, X., Regnier, F. E., Zhang, M., *Anal. Chem.* 2008, *80*, 2664–2671.

[18] Pierce, K. M., Hoggard, J. C., Mohler, R. E., Synovec, R. E., *J. Chromatogr. A* 2008, *1184*, 341–352.

[19] GC Image, LLC, GC Image® LCxLC Software, 2008.

[20] Visvanathan, A., Reichenbach, S. E., Tao, Q., *J. Electron. Imaging* 2007, *16*, 033004.

[21] Reichenbach, S. E., Carr, P., Stoll, D., Tao, Q., *J. Chromatogr. A* 2009, *1216*, 3458–3466.

[22] Reichenbach, S. E., Ni, M., Kottapalli, V., Visvanathan, A., *Chemom. Intell. Lab. Syst.* 2004, *71* 107–120.

[23] Beucher, S., Lantuejoul, C., International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation 1979, pp. 17–28.

[24] Ni, M., Reichenbach, S. E., Proc. Int. Conf. Pattern Recognition, Vol. 2, 2004, pp. 145–148.

[25] Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, New York, NY 2000.

[26] Platt, J. C., in: Scholkopf, B., Burges, C. J. C., Smola, A. J. (Eds.), *Advances in Kernel Methods – Support Vector Learning*, MIT Press, Cambridge, MA 1999, pp. 185–208.

[27] Witten, I. H., Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edn, Morgan Kaufmann, San Francisco, CA 2005.

[28] Mitchell, T. M., *Machine Learning*, McGraw–Hill, Boston, MA 1997.