Anal. Chem. **2006,** *78,* 975−983

Articles

Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data

Ragnar Stolt,[†] Ralf J. O. Torgrip,*[,†,‡] Johan Lindberg,[‡] Leonard Csenki,[†] Johan Kolmert,[‡] Ina Schuppe-Koistinen,[‡] and Sven P. Jacobsson[†,§]

Department of Analytical Chemistry, BioSysteMetrics Group, Stockholm University, SE-106 91, Stockholm, Sweden, and Safety Assessment, Molecular Toxicology, and PAR&D, AstraZeneca R&D Södertälje, SE-151 85, Södertälje, Sweden

The first step when analyzing multicomponent LC/MS data from complex samples such as biofluid metabolic profiles is to separate the data into information and noise via, for example, peak detection. Due to the complex nature of this type of data, with problems such as alternating backgrounds and differing peak shapes, this can be a very complex task. This paper presents and evaluates a two-dimensional peak detection algorithm based on raw vector-represented LC/MS data. The algorithm exploits the fact that in high-resolution centroid data chromatographic peaks emerge flanked with data voids in the corresponding mass axis. According to the proposed method, only 4‰ of the total amount of data from a urine sample is defined as chromatographic peaks; however, 94% of the raw data variance is captured within these peaks. Compared to bucketed data, results show that essentially the same features that an experienced analyst would define as peaks can automatically be extracted with a minimum of noise and background. The method is simple and requires a priori knowledge of only the minimum chromatographic peak width—a system-dependent parameter that is easily assessed. Additional meta parameters are estimated from the data themselves. The result is well-defined chromatographic peaks that are consistently arranged in a matrix at their corresponding *m/z* values. In the context of automated analysis, the method thus provides an alternative to the traditional approach of bucketing the data followed by denoising and/or one-dimensional peak detection. The software implementation of the proposed algorithm is available at http://www.anchem.su.se/peakd as compiled code for Matlab.

Today's industrial drug development is strictly regulated, and only a very small fraction of the investigated pool of drug candidates will ever reach the market. The increasing development costs make it necessary to focus on the early stages of drug discovery including biomarker and metabolism studies. Metabolic fingerprints can be gathered in vivo from biofluids such as urine[1] and plasma.[2] Although ^{1}H NMR is perhaps the most widely used biofluid screening technique,[3] liquid chromatography mass spectrometry (LC/MS) is currently emerging as an alternative, mainly because of its high sensitivity[4] and relatively low costs.

The most common way of representing LC/MS data when developing algorithms for data analysis is as a matrix, where columns represent different *m/z* values and rows represent chromatographic time points. The transformation from instrument-specific formats, where spectra are arranged in a vector representation, to a matrix is often achieved by some kind of bucketing of the *m/z* axis. Bucketing in this context could be described as treating all *m/z* values found along the time axis within a certain *m/z* range as identical in order to compensate for small *m/z* differences between separate time points. The bucketing procedure leads to an optimization problem. If the bucket size is too small, a chromatographic peak can be split into several subpeaks,[5] while if it is too large, noise will be added to the peaks, the signal-

* To whom correspondence should be addressed. E-mail: ralf.torgrip@anchem.su.se.

[†] Stockholm University.

[‡] Molecular Toxicology, AstraZeneca R&D Södertälje.

[§] PAR&D, AstraZeneca R&D Södertälje.

(1) Idborg, H.; Edlund, P.-O.; Jacobsson, S. P. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 944−954.

(2) Nicholson, J. K.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Lindon, J. C. *Anal. Chem.* **1995**, *67*, 793−811.

(3) Nicholls, A. W.; Nicholson, J. K.; Haselden, J. N.; Waterfield, C. J. *Biomarkers* **2000**, *5*, 410−423.

(4) Wilson, I. D.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. M. *J. Chromatogr., B* **2005**, *817*, 67−76.

(5) Smedsgaard, J.; Nielsen, J. *J. Exp. Bot.* **2004**, *56*, 273−286.

10.1021/ac050980b CCC: $33.50
Published on Web 01/14/2006

to-noise ratio will decrease, and small peaks will be difficult to detect. Also, peaks emerging on the border between two buckets will inevitably be split, and as a consequence, automated multivariate comparison between different LC/MS matrices may not be successful.

Furthermore, the interpretation of LC/MS data is often not trivial because of the vast amount of noise present, especially when complex samples such as biofluids are analyzed. Generally, some sort of denoising procedure is applied, and today, numerous denoising techniques are available. High-frequency random noise is traditionally removed by filtering the Fourier space, although in the field of chemometrics wavelet analysis is gaining in popularity.[6] The removal of chemical noise is a more hazardous task because of its similarities with possible analytes.[7]

Several papers[8−10] have introduced approaches to denoising LC/MS data. One popular method is CODA,[11] which operates by estimating the fraction of mass chromatograms where the relevant information lies. Noisy m/z chromatograms with low chromatographic quality will be removed. This approach has proved to be successful in terms of finding a few m/z chromatograms of interest among a vast number. However, in the context of biofluid analysis, the majority of m/z chromatograms will most probably contain at least one potential analytical peak, thus undermining the theoretical and practical use of techniques such as CODA.

One way of separating information from data is by peak detection. Most peak detection algorithms are based on either statistical distributions or smoothing functions combined with different orders of derivatives.[12] For several reasons, this might not always be a successful approach. The sampling frequency along the time axis is often relatively low; that is, each chromatographic peak is only represented by three to five time points, and so numerical derivatives will obviously be unsuccessful. Denoising via Fourier transformation suffers from the same limitation. However, it is not always advisable to increase sampling frequency since having more data points per chromatogram results in larger data files, leading to increased data analytical cycle time. Higher sampling frequency will also affect the quality of the data, with altered ion statistics and peak shapes. Using matched filtering[10] and other related techniques when comparing the characteristics of a peak with a known distribution, e.g., a Gaussian, will be successful if again the sampling frequency is fairly high and the peak shapes have a reasonable conformity compared to an idealized target distribution.

In a biofluid-screening context, it is clear that far from all possible chromatographic peaks can be found using the above-mentioned criteria. Peaks with low intensities and differing shapes make it difficult to decide whether a deviation from background noise is a true peak corresponding to an analyte. Furthermore, most denoising and peak detection algorithms significantly alter peak shapes[8,9] and rely on a set of parameters with low generality.[7]

This paper presents a peak detection algorithm with only one user-defined parameter that exploits the two-dimensional nature of LC/MS data. The procedure relies on the fact that in high-resolution centroid data true chromatographic peaks emerge, flanked with data voids in the corresponding m/z axis. These subspaces with zero data density are a consequence of the peak detection algorithm that converts peaks in the continuous spectra to centroid representation, combined with the fact that the probability of finding two or more components with a similar m/z located at the same retention time is low. The proposed approach can be viewed as denoising in the sense that only LC/MS subspaces that make chromatographic sense, i.e., exhibit consistently detected m/z's, are defined as peaks. As a result, vector-represented LC/MS data are transformed into a matrix with peaks assigned to specific m/z values. All other entries will be zero. Furthermore, peak shapes will not be altered, and the only required a priori information is the expected minimum chromatographic peak width.

The suggested procedure is evaluated using LC/MS data from a urine metabolic profiling study.

THEORY

Nature of LC/MS Data. When using electrospray ionization-mass spectrometry (ESI-MS), a mass spectrum contains not only peaks corresponding to the analytes but also other contributions of great significance,[7] especially from the mobile phase. Furthermore, features from limited ion statistics, instability of the ion source, and spikes will be detected as peaks when transforming the data from continuous to centroid mode. An important difference between analytes and nonchemical noise is their time extension. Analytes elute over several consecutive time points in a consistent manner, while noise appears randomly in the LC/MS space. Because of limited m/z resolution, analytical peaks will also have a certain extension along the m/z axis. Due to this extension, in centroid representation, there will be a specific "data void" flanking an emerging analyte, where the probability of detecting noise is low.

Figure 1 illustrates a typical m/z spectrum in continuous representation, with its corresponding centroids. Two major peaks with a certain m/z width dominate the spectrum. Within each peak width it is unlikely that more than one centroid will be found, since the probability of detecting several analytes with very similar m/z values and retention times is low in LC/MS analysis.

Peak detection algorithms, enclosed in MS instrumental software, which transform continuous data to centroid representation in real time, can be adjusted by a user-defined parameter set. Not all parameter configurations will result in data voids described in this paper; for example, a high-intensity threshold will suppress the amount of noise defining the boundaries of the data voids. However, by using continuous data, it is always possible to produce data voids with an appropriate peak detection algorithm. An example of such an algorithm is given in the Appendix.

Figure 2 depicts a portion of the LC/MS space where such data voids can be found.

Description of the Algorithm. A LC/MS data file from a mass spectrometer in centroid mode (data formats such as ASCII and

(6) Leung, A. K.-M.; Chau, F.-T.; Gao, J.-B. *Chem. Intell. Lab. Syst.* **1998**, *43*, 165−184.
(7) Andreev, V. P.; Rejtar, T.; Chen, H.-S.; Mosovets, E. V.; Ivanov, A. R.; Karger, B. L. *Anal. Chem.* **2003**, *75*, 6314−6326.
(8) Fleming, C. M.; Kowalski, B. R.; Apffel, A.; Hancock, W. S. *J. Chromatogr., A* **1999**, *849*, 71−85.
(9) Muddiman, D. C.; Huang, B. M.; Anderson, G. A.; Rockwood, A.; Hofstadler, S. A.; Weir-Lipton, M. S.; Proctor, A.; Wu, Q.; Smith, R. D. *J. Chromatogr., A* **1997**, *771*, 1−7.
(10) Danielsson, R.; Bylund, D.; Markides, K. E. *Anal. Chim. Acta* **2002**, *454*, 167−184.
(11) Windig, W.; Phalp, J. M.; Payne, A. W. *Anal. Chem.* **1996**, *68*, 3602−3606.
(12) Felinger, A. In *Data Analysis and Signal Processing in Chromatography*; Elsevier Science: New York, 1998; pp 183−190.

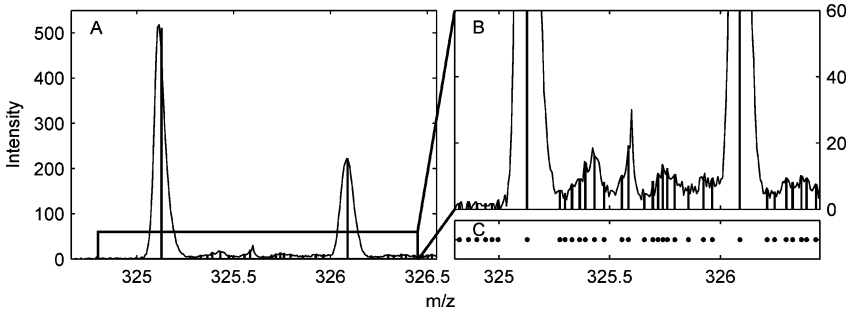

Figure 1. (A) *m*/*z* spectrum in continuous representation. Detected centroids are represented as bars. (B) Zoom of the spectrum. (C) Centroid peaks viewed along the intensity axis. Compared to the surrounding noise, the two major continuous peaks are flanked by data voids in the centroid representation.

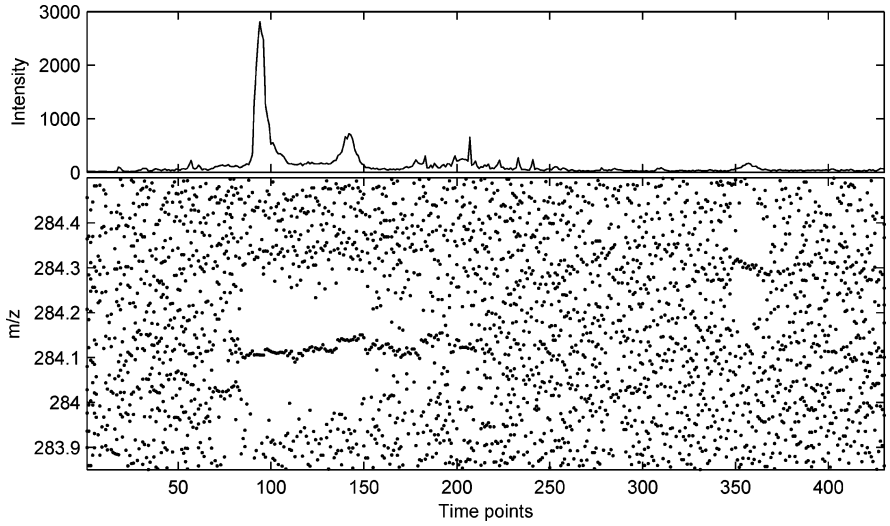

Figure 2. Small portion of the LC/MS space (rat 1). The upper part of the figure shows the corresponding TIC.

NetCDF are often used) contains primarily three vectors—one representing the detected intensity values $\mathbf{x}_{i,i=1...N}$ (where N is the total number of accumulated data points) arranged in ascending time order, one with the corresponding m/z values $\mathbf{y}_{j,j-1...N}$, and one vector $\mathbf{z}_{k,k=1...T}$ (where T is the total number of mass spectra) containing $\mathbf{x}$ (and $\mathbf{y}$) positions where each mass spectrum starts.

The main feature of the proposed procedure is to localize the data voids within these arrays. Data points flanked by a data void can be discriminated from noise by two features: (1) Each point has a large m/z distance to its nearest neighbor along the m/z axis. (2) Each point has a small m/z distance to its nearest neighbor within the following spectrum.

A potential surface can be constructed by combining these two properties using the following procedure.

Let $\mathbf{l}$ be the indices of the kth mass spectrum, $\mathbf{l} = z_k...(z_{k+1} - 1)$ and $\mathbf{k} = 1...T - 1$. The m/z distance for each centroid within the same mass spectrum is defined as

$$w_l = \begin{cases} y_{l+1} - y_l & \text{if } l = z_k \\ \min(y_l - y_{l-1}, y_{l+1} - y_l) & \text{if } z_k < 1 < z_{k+1} - 1 \\ y_l - y_{l-1} & \text{if } l = z_{k+1} - 1 \end{cases} \quad (1)$$

With $\mathbf{l}$ as above and $\mathbf{k} = 1...T - 1$ (where $z_{T+1} - 1$ corresponds to N), the m/z distances to the nearest neighbor in the following spectrum are defined as

$$p_l = \min(|y_l - \mathbf{y}_{z_{k+1}...z_{k+2}-1}|) \quad (2)$$

A data void is characterized as a maximum in $\mathbf{w}$ and a minimum in $\mathbf{p}$. The total potential field $\Phi_{j,j=1...(z_T-1)}$ is defined as

$$\Omega_j = w_j - \frac{\sum_i w_i}{\sum_i p_i} p_j \quad (3)$$

$$\Phi_j = \begin{cases} \Omega_j & \text{if } \Omega_j > 0 \\ 0 & \text{if } \Omega_j < 0 \end{cases} \quad (4)$$

To minimize the influence of a drastic drop of the potential field within a chromatographic peak, a smoothing procedure is applied to $\mathbf{\Phi}$. This situation is unlikely to occur, but might do so if, for example, a peak in a specific mass spectrum in continuous mode is of a thorny nature and interpreted as two different peaks in centroid mode. Given a specific peak width, t, in time points, characteristic of the data set, the smoothing procedure averages the potential field around the maximum intensity value of

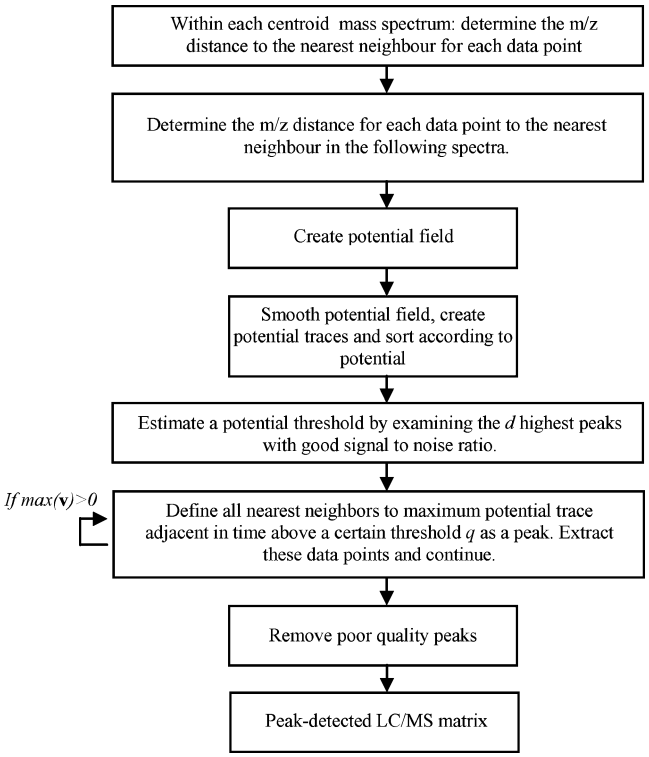

Figure 3. Flowchart of the second-order peak detection algorithm.

$\mathbf{x}_{i,i=1\ldots(z_T-1)}$. The modified potential surface at this location is set to the mean value of the t most intensive data points in $\mathbf{x}$ and adjacent in time that meet criterion 2, that is, are the nearest neighbors in time. The smoothing function repeats the procedure with the subsequent maximum intensity in $\mathbf{x}$ corresponding to nonsmoothed potential in $\mathbf{\Phi}$ until a certain intensity threshold r is reached.

The smoothed potential field is used to calculate a vector $\mathbf{v}$ of potential traces. A potential trace is defined as the mean of the potential values corresponding to the t data points in $\mathbf{x}$ that are most intensive and adjacent in time that meet criterion 2. Potential traces are calculated for all intensities larger than r, each data point belonging uniquely to only one potential trace. The position in the LC/MS space of a potential trace is identical to the position of the highest intensity value x of the data points belonging to the potential trace.

Peaks are detected and defined via a similar search. In the first step, the maximum potential trace is located and all the nearest neighbors adjacent in time that have a potential value above a threshold q are collected, and together this set defines a true peak. The procedure is repeated for all potential traces in $\mathbf{v}$. The m/z value of the peak is taken as the mean of the detected m/z values. The flowchart of the algorithm is shown in Figure 3.

The threshold parameter q is estimated from the data by sampling the potential values at the starting and end points of the d most intensive chromatographic peaks in the sample that have high S/N ratios. The procedure can be summarized as follows.

Potential Threshold Estimation. In chromatographic peaks with high S/N ratio surrounded by well-behaved baselines, the absolute value of the slope between the starting point and the end point is relatively low; that is, the intensity difference between these points is small compared to the maximum value of the peak.

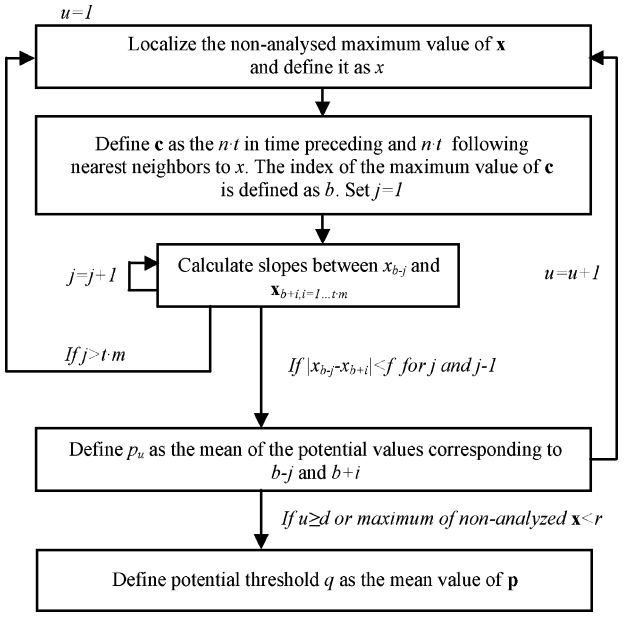

Figure 4. Flowchart showing the estimation of the potential threshold.

This feature can be exploited by means of the following procedure: localize the maximum value in $\mathbf{x}$ and collect the $n{\cdot}t$ neighbors nearest in time preceding that time point and the $n{\cdot}t$ following neighbors, where n is a scalar. The resulting vector is defined as $\mathbf{c}$. The index of the maximum value in $\mathbf{c}$ is defined as b. Next, the slopes between x_{b-1} and $\mathbf{x}_{b+i,i=1\ldots t\cdot m}$, where m is a scalar, are calculated, and the procedure is terminated if $|x_{b-1} - x_{b+i}| < f$, where f is a scalar. If the convergence criterion is not met, the procedure is repeated with x_{b-2} and so on until the stop criterion is reached at two adjacent points, i.e., x_{b-j} and x_{b-j-1}. The starting point of the peak is now defined as $(b - j)$, and the end point is defined as $(b + i)$. If convergence is not reached within $\mathbf{c}$ the peak is discarded. Figure 4 shows the various steps of the algorithm.

The potential threshold q is finally set as the mean of the potential values at the starting and the end points of the d detected peaks.

Removal of Poor-Quality Peaks. Not all features flanked with data voids that are detected in the LC/MS space are of interest. Bleeding columns and solvent ions will have approximately the same potential behavior as analytical peaks, although they are not eluted as normal unimodal peaks but rather like a more or less horizontal baseline. To minimize the number of false positives, each potential peak has to meet the following simple criterion: there must be at least t adjacent points above a straight line drawn between the starting point and end point of the peak. Knowledge about the peak width is thus necessary. However, observe that it is only information about the minimum peak width that is required. Varying peak shapes and peak broadening will not influence peak detection as long as the detected peak width is larger than t.

EXPERIMENTAL SECTION

Sample Preparation. The data in this study are a subset of a larger study, consisting of urine from 35 male rats (Han Wistar, substrain BrlHan:WIST@Mo), ~2 months of age at study onset, weight range 220−320 g. Animals were multiple housed (2 or 3

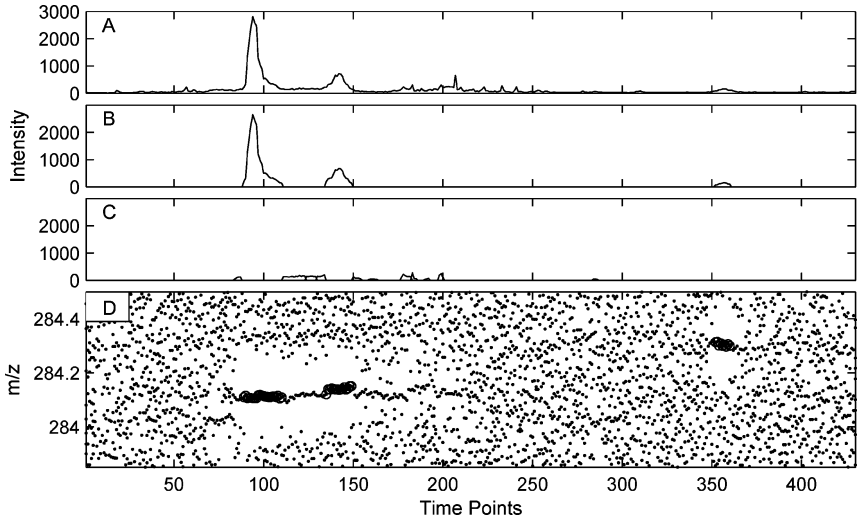

Figure 5. Small portion of the LC/MS space with detected peaks (rat 1). (A) Corresponding TIC. (B) Detected peaks. (C) Rejected poor-quality peaks. (D) Raw data with points corresponding to the detected peaks encircled.

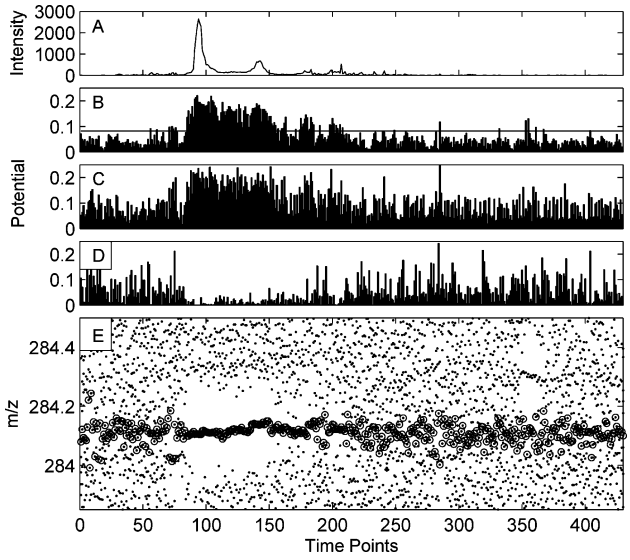

Figure 6. Potential values (rat 1). (A) Chromatogram corresponding to the circled data points in (E). Each circled point fulfills (2). (D) *m*/*z* distance to the nearest neighbor in the following spectrum. (C) *m*/*z* distance to the nearest neighbor in the current spectrum. (B) Combined and smoothed potential values. The solid line represents the estimated threshold.

per cage) during an acclimatization period (10 days) in the animal house. During the course of the study, 5 animals/group were housed in individual metabolism cages to allow continuous collection of urine. Urine samples were collected on ice into 1% (1 mL, w/v) sodium azide. Urine pH was recorded and the samples were centrifuged at 500*g* for 10 min, the supernatant for subsequent analysis being kept at −80 °C.

From this study, a subset of four control samples (rats 1−4) was used. Prior to LC/MS analysis, 30 μL urine was diluted with 90 μL of MilliQ H_2O. The injected volume was 10 μL.

LC/ESI-MS Analysis. Urine mass spectra were obtained at a rate of 30 centroid spectra/min using a Waters Alliance HT 2695 (Manchester, U.K.) coupled to a Waters Micromass QTOF-Micro.

A Symmetry C18 (100 × 2.1 mm, 3.5 μm) column was used for hydrophobic LC separation at 20 °C. Mobile phase A consisted of water with 0.1% (v/v) formic acid, and mobile phase B consisted of acetonitrile, also with 0.1% formic acid. The flow rate was 300 μL/min, and the gradient started at 1% B for 2 min and was then linearly increased to 30% B in 1 min. Next, the gradient was linearly increased to 95% B in 6 min and kept at isocratic conditions for 2 min. Finally, the gradient was allowed to reach the initial conditions in 6 s and then equilibrated for 4.9 min. The total cycle time was thus 16 min.

The capillary voltage of the ESI interface was 3.0 kV, and the cone voltage was 35 V. The source temperature was 120 °C and the desolvation temperature 300 °C. Lockspray was used to ensure mass accuracy and reproducibility. Leucine-enkephalin (M + H m/z 556.2771) was used as the lock mass at a concentration of 1 ng/μL and a flow rate of 9 μL/min. The Lockspray frequency was set to 5 Hz. The resulting mass resolution was ∼5000 at full width at half-maximum (fwhm). Data were obtained between m/z 80 and 800 in positive mode and exported from Masslynx 4.0 (Waters UK Ltd.) raw data files into NetCdf files using Masslynx's enclosed software Databridge.

Data Analysis. All data were analyzed and all algorithms were written in Matlab v6.5 (Mathworks Inc.). The computer configuration was a Pentium 4, 2.8 GHz with 1 GB of RAM.

For these data, the minimum peak width was set to $t = 3$, which corresponds to an elution time of 6 s. The intensity threshold r was set to 50 counts, corresponding to ∼1‰ of the highest intensity. The 20 highest peaks were assumed to have acceptable S/N ratios and could be used to estimate the potential threshold parameter, i.e., $d = 20$, the parameters n and m being set to 10 and 4, respectively. The intensity threshold f used when estimating the potential threshold was set to 1% of the maximum peak height for each analyzed peak.

Bucketed Data. To compare the performance of the proposed algorithm, peak-detected data were compared with data processed with a bucketing algorithm. A simple bucketing method involves splitting the m/z axis into a number of buckets with a specific resolution and taking the resulting m/z chromatograms as the total ion chromatograms (TICs) or base peak chromatograms of the corresponding LC/MS subspaces. However, the risk of splitting chromatographic peaks into different buckets is substan-

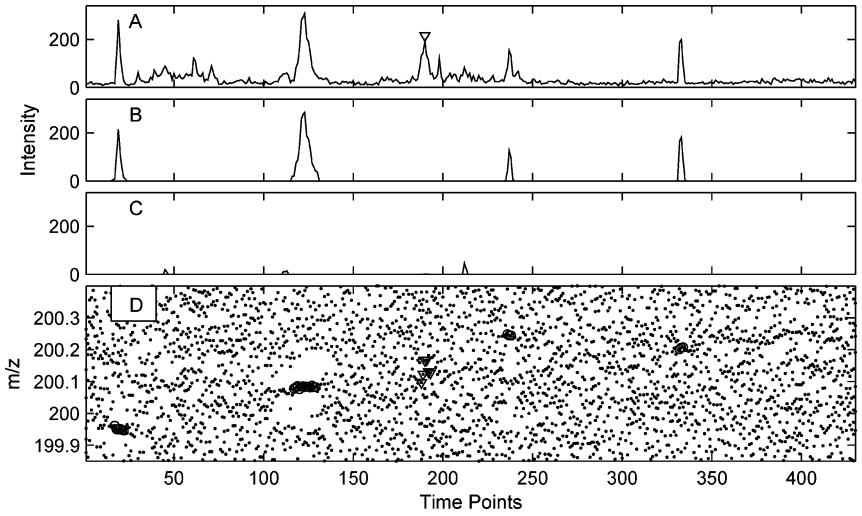

Figure 7. Subdomain of the LC/MS space with detected peaks (rat 1). (A) Corresponding TIC. (B) Detected peaks. (C) Rejected low-quality peaks. (D) Raw data with points corresponding to the detected peaks encircled. A peak marked with a triangle has a large m/z distribution, suggesting that the data points involved are either spikes or correspond to peaks eluting with only one data point above the detection limit. In either case, rejection of the peak can be supported.

tial.[5] An alternative approach would be to adjust the bucket positions to the intensities of the data, that is, have the first bucket positioned symmetrically around the highest intensity, remove the bucketed intensity values, and have the second bucket be positioned around the highest remaining intensity, and so on. In this paper, the latter positioning approach was used with a bucket size of 0.1 Th, with the resulting m/z chromatogram represented as the TIC of each bucketed portion of the LC/MS space.

RESULTS AND DISCUSSION

Peak Detection Examples. Figure 5D,B illustrates a small portion of the LC/MS space ($283.85 < m/z < 284.5$) and the corresponding features that were defined as peaks using the suggested approach. The potential values representing this subspace's largest peak, and its surroundings, are visualized in Figure 6B. Three peaks were detected at different retention times and at slightly different m/z values (284.1104, 284.1413, 284.3038). The intensities of the peaks range from approximately 200 to 3000 counts. The noisy fraction of this subspace between 170 and 240 time points contains signal patterns with higher intensities than the smallest peak found, suggesting that intensity is not a valid peak quality criterion. However, in this example, there is a significant difference in peak widths between the chromatographic peaks and what is believed to be noise. One-dimensional peak detection using peak widths as a discriminating criterion should thus be possible. A more complicated situation occurs when noisy patterns and what are believed to be chromatographic peaks have not only a similar intensity distribution but also approximately the same extension along the time axis.

An example of such a situation ($199.85 < m/z < 200.4$) is illustrated in Figure 7. In this case, four peaks were detected and the peak at $r_T \sim 333$ consists of only three points. Peaks with these characteristics are difficult to find with a method based on derivatives without the use of a number of sample specific parameters.

This example is representative of urine metabolic LC/MS data. It is obviously difficult to find an automated procedure for separating possible peaks from noise, since noise is a dominant part of the chromatogram. One conclusion must be that no separation is possible, without a massive loss of information, between m/z chromatograms of high quality and m/z chromatograms of low quality, since most m/z chromatograms contain both information and noise at a similar level. In this context, methods that rank entire m/z chromatograms based on their quality such as CODA will not work efficiently. In Figure 7, a peak (marked with a triangle) that was rejected by the algorithm can be found. The reason may at first seem unclear since the chromatographic profiles of the rejected peak and those accepted are similar. Looking at the two-dimensional plot, where the data points building up the major shape of the specific peak are marked with triangles, it is obvious that the rejected peak has a significantly different appearance. Compared to the accepted peaks, it seems

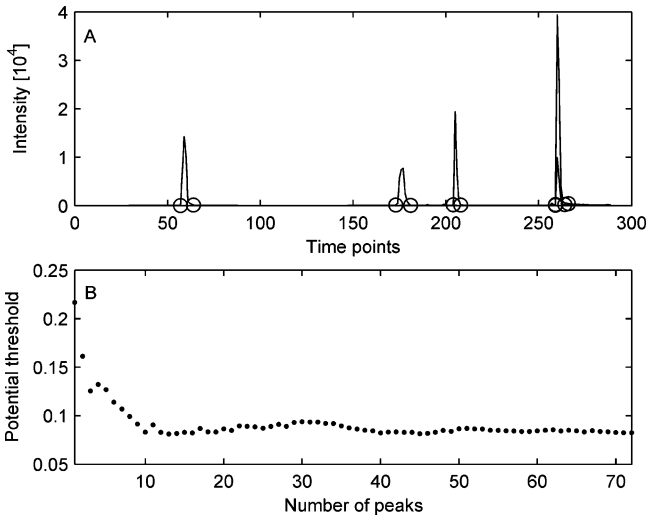

Figure 8. Potential threshold estimation (rat 1). (A) The most intensive peaks detected by the potential threshold estimation approach. Only the five most intensive peaks are shown for clarity. The circled points correspond to the detected starting and end values of each peak. (B) The potential threshold as a function of the number of analyzed peaks.

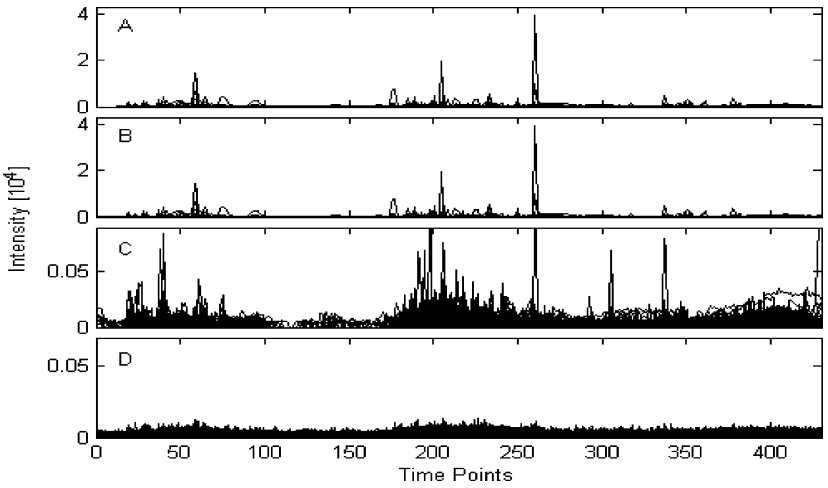

Figure 9. Detected peaks in the entire LS/MS space (rat 1). (A) Peak detected with the proposed method. (B) EIC from raw data using bucketing with a bucket size $R_S = 0.1$ Th. (C) Rejected poor-quality peaks, either background or peaks with widths $< t$ data points. (D) EIC from the remaining data (bucket size $R_S = 0.1$ Th), not defined as peaks or rejected as poor-quality peaks. No chromatographic peak shapes could be found by visual inspection when traversing the *m*/*z* chromatograms.

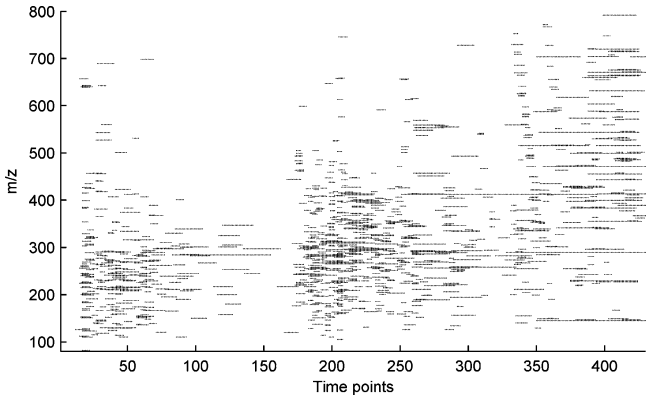

Figure 10. Peak detection in the entire LC/MS space (rat 1).

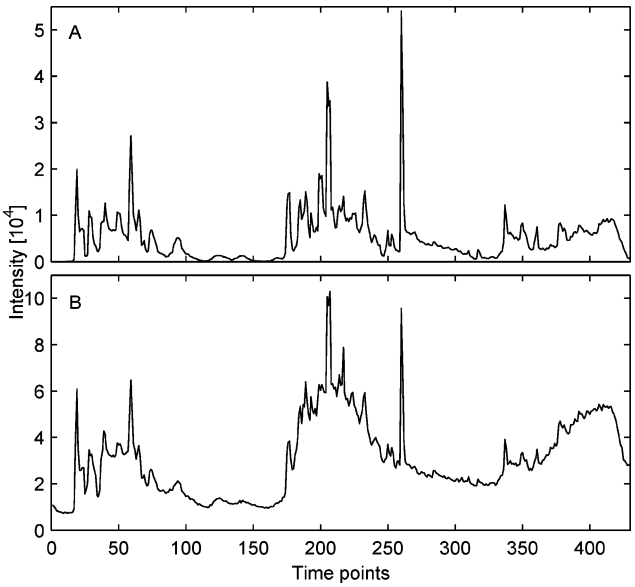

Figure 11. Comparing TICs (rat 1) from (A) peak-detected data and (B) raw data. A significant difference in intensity and baseline can be noted.

as if it consists of spikes widely spread on the m/z axis. This phenomenon could originate from analytes with only one point above the LOD, poor precision, or electronic spikes.

Estimating the Potential Threshold. Complex urine LC/MS data most certainly contain a number of peaks with high signal-to-noise ratios that can be detected using the potential threshold-estimating approach described above. Figure 8 illustrates the potential threshold estimated from 73 peaks that were successfully detected among the 80 most intensive peaks. Due to massive tailing, seven peaks were rejected by the algorithm. The potential threshold is here shown as a function of the number of contributing peaks, and it is clear that ~15 chromatographic peaks are sufficient in order to find a robust potential threshold.

Evaluating the Entire LC/MS Space. The entire LC/MS sample (rat 1) was found to contain 1189 peaks, corresponding to 4‰ of the data and 94% of the raw data variance. Chromatographic peaks with low intensities might be expected not to have the same data void structure as peaks with higher intensities, in view of differences in ion statistics. Continuous mass peaks with insufficient ion statistics can be interpreted as several centroid peaks, resulting in a blurry data structure with no data voids. However, results show that chromatographic peaks with low intensities are flanked by more or less identical data voids to those flanking peaks with high-intensity values. The maximum intensity of the detected peak with the lowest maximum intensity was 57 counts, and the corresponding value of the detected peak with the highest maximum intensity was 4×10^4 counts. Figure 9 illustrates the corresponding extracted ion chromatograms (EICs) from both peak-detected data and bucketed data. No major features that differ between the two methods can be found by visual inspection. The EICs of the residual, i.e., the part of the data not defined as peaks or discarded as no peaks, are also shown in Figure 9D. Consistent features resembling chromatographic peaks can no longer be found, suggesting that the algorithm has retrieved all consistent peaks. As previously mentioned, it is not only analytical peaks that will give rise to data voids; phenomena such as column bleeding and mobile-phase ions will also create a similar pattern, though one of no interest. Removed nonpeak features (corresponding to 3‰ of the data size and 4% of the raw data variance) are illustrated in Figure 9C. Visual inspection revealed only a couple of possible analytical peaks among these rejected features, suggesting that except for a few false negatives

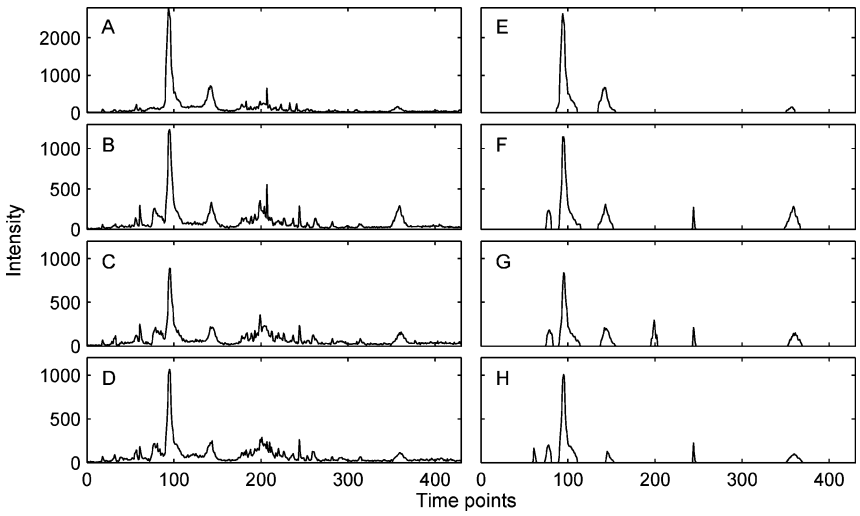

Figure 12. Comparing peak detection (283.85 < m/z < 284.5) from four different samples (rats 1−4). (A−D) TIC from raw data. (E−H) Corresponding peaks detected by the proposed method.

the algorithm truly extracts the kernel of information from the data space. Figure 10 demonstrates the structure of the extracted peaks. A few detected peaks with very long elution times can be found. These features are probably false positives, indicating that the peak quality criterion could be improved. However, a more stringent peak quality criterion will not only decrease the number of false positives but is also bound to increase the number of false negatives.

The effects of the peak detection algorithm can also be seen when studying the TIC. Figure 11 compares the TICs from peak detected data (A) and raw data (B). As expected, the baseline is close to zero and peaks are more resolved in the peak-detected data.

It is remarkable that the detected peaks are represented by only 4‰ of the total amount of data. Assuming that more or less all peaks of interest have been found, the conclusion must be that the vast majority of the collected data are only noise or peaks where most of the data points are below the detection limit. Urine LC/MS samples of this kind contain $\sim 3 \times 10^6$ data points, which corresponds to 20 MB of data arranged in a vector representation. A peak-detected matrix with 430 time points and 1189 m/z values corresponds to 5×10^5 data points or 3.5 MB, although the actual number of data points above zero is only 1×10^4, equivalent to 87 kB. Bucketed data with $R_S = 0.1$ Th corresponds to 1×10^6 data points equivalent to 7 MB.

Peak-detected data in the region (283.85 < m/z < 284.5) from different rats can be found in Figure 12. Major similar peaks are detected in all samples, supporting the validity of the method. Figure 12 also illustrates common problems with this kind of data. When comparing different samples, corresponding peaks have retention time shifts, different peak shapes, and variations in peak intensity.

User-defined meta parameters in the proposed algorithm are mainly a rough estimation of the minimum chromatographic peak width t, which is not an unreasonable parameter to estimate, and the intensity threshold r. It is not necessary to specify r, although a fairly good estimate of the noise level reduces the computational time significantly. With the given parameter set, it takes ∼2 h to analyze a sample using Matlab. The parameters regulating the estimate of the potential threshold were found to be quite robust. Changing f from 1 to 2% of the peak height did not change q significantly. Changing m did not result in any drastic changes either, as long as $m \cdot t$ was larger than the estimated peak width.

One might expect the m/z peak width to increase with increasing m/z and thus also the m/z distance to the nearest neighbor. No such correlation could be found, however, among the detected peaks (results not shown).

To thoroughly evaluate the method and to be able to compare it to alternative software such as MarkerLynx (Waters), in a multivariate sense, intersample comparison has to be performed. To successfully apply multivariate methods such as PCA and PARAFAC, retention time shifts and variations in m/z between different samples have to be compensated for. One approach is a two-dimensional alignment procedure between peak-detected matrices such as the one shown in Figure 10. Future development of the method is also directed toward a sparse representation of the peak-detected data instead of a matrix representation. A sparse vector representation will not only reduce demands on data storage capacity but also increase the speed of the method, which at the present time could be improved.

CONCLUSIONS

A peak detection procedure is presented that extracts consistent chromatographic peaks from centroid high-resolution LC/MS data in a biofluid context. The method simultaneously assesses the quality of chromatographic peaks in both the m/z and the time axis. In other words, the definition of an analytical peak is that the m/z representation is consistent and the peak elutes in a way that conforms to chromatography, that is, is present over a number of sampled time points.

The number of peaks detected from a typical sample is ∼1200, corresponding to ∼4‰ of the total amount of urine LC/MS data. The detected peaks correspond to 94% of the raw data variance. Intersample comparison confirms that the method performs well. What are defined as peaks have large similarities between different samples, and what is defined as noise does not. The method has thus the potential of becoming an alternative to the traditional approach of bucketing the data followed by denoising.

APPENDIX

A simple peak detection algorithm converting MS spectra from continuous to centroid representation can be described as follows:

(1) Let **x** be an arbitrary spectrum. Define **x′** as the Savitzky−Golay[13] first-order derivative with a given window length w and polynomial order p. Define **x″** as the second-order Savitzky−Golay derivative.

(2) Let **c** be all zero crossing points in **x′**, i.e., the indices immediately before the derivative changes sign, and **d** all zero crossing points in **x″**.

(3) For an element in **c**, c_i, let s be the first adjacent element in **d** smaller than c_i and l the first adjacent element in **d** larger than c_i. Define c_i as the position of the maximum intensity in a possible peak if $x_s'' > 0$ and $x_l'' < 0$ and remove all other elements in **c**.

(4) Adjust the position of the possible peaks to the index corresponding to $max(x_{c_i}, x_{c_i+1})$. Let m m/z bins be the minimum peak width. Remove all possible peak positions in **c** that do not satisfy the conditions, $x'_{c_i-m\ldots c_i-1} > 0$ and $x'_{c_i-1\ldots c_i+m} < 0$.

(5) Let the remaining elements in **c** be the positions of the centroid peaks where centroid peak heights correspond to $\mathbf{x_c}$. Centroid peak heights can also be represented by peak areas using more sophisticated algorithms.[14] However, when using the above-described algorithm, data voids will be created.

Received for review June 3, 2005. Accepted November 30, 2005.

AC050980B

(13) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−1639.
(14) Gentzel, M.; Koecher, T.; Ponnusamy, S.; Wilm, M. *Proteomics* **2003**, *3*, 1597−1610.