ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/5341324

Wavelet-Based Method for Noise Characterization and Rejection in High-Performance Liquid Chromatography Coupled to Mass Spectrometry

ARTICLE *in* ANALYTICAL CHEMISTRY · AUGUST 2008

Impact Factor: 5.64 · DOI: 10.1021/ac800166w · Source: PubMed

CITATIONS
30

READS
34

6 AUTHORS, INCLUDING:

Salvatore Cappadona
Centre for Genomic Regulation
14 PUBLICATIONS 214 CITATIONS
SEE PROFILE

Peter James
University of Queensland
120 PUBLICATIONS 4,250 CITATIONS
SEE PROFILE

Sergio Cerutti
Politecnico di Milano
602 PUBLICATIONS 12,500 CITATIONS
SEE PROFILE

Linda Pattini
Politecnico di Milano
34 PUBLICATIONS 433 CITATIONS
SEE PROFILE

Available from: Linda Pattini
Retrieved on: 31 January 2016

Anal. Chem. **2008,** *80,* 4960–4968

Wavelet-Based Method for Noise Characterization and Rejection in High-Performance Liquid Chromatography Coupled to Mass Spectrometry

Salvatore Cappadona,*[†,‡] Fredrik Levander,[‡] Maria Jansson,[‡] Peter James,[‡] Sergio Cerutti,[†] and Linda Pattini[†]

Department of Bioengineering, IIT Unit, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy, and Department of Protein Technology, Lund University, BMC D13, SE-22184 Lund, Sweden

We present a new method for rejecting noise from HPLC–MS data sets. The algorithm reveals peptides at low concentrations by minimizing both the chemical and the random noise. The goal is reached through a systematic approach to characterize and remove the background. The data are represented as two-dimensional maps, in order to optimally exploit the complementary dimensions of separation of the peptides offered by the LC−MS technique. The virtual chromatograms, reconstructed from the spectrographic data, have proved to be more suitable to characterize the noise than the raw mass spectra. By means of wavelet analysis, it was possible to access both the chemical and the random noise, at different scales of the decomposition. The novel approach has proved to efficiently distinguish signal from noise and to selectively reject the background while preserving low-abundance peptides.

High-performance liquid chromatography coupled to mass spectrometry (HPLC−MS) and database searching is increasingly becoming the method of choice for large-scale proteomic analysis of tissue samples and biological fluids. HPLC−MS provides a complementary level of separation of the peptides, which allows identification by their LC retention times in addition to their m/z values.[1] The extra dimension of separation in the chromatographic dimension adds several challenging analytical issues that should be considered when processing data. An increasing number of researchers have begun to manage the data as two-dimensional maps to accomplish differential analyses. The typical sequence of operations needed for data manipulation consists of three steps: preprocessing of the MS spectra for noise rejection, peak detection and quantification, and finally quantitative analysis in order to define expression changes. In the last years, though, the first step has been often neglected and research has focused principally on the second and the third steps, in order to identify potential biomarkers.[2] These approaches may be enhanced if signal preprocessing is applied. In their review on processing and classification methods in mass spectrometry, for instance, Hilario et al.[3] explicitly called for performing meticulous and customized cleaning methods before any further data analysis. Similarly, in their comprehensive review on comparative proteomic profiling, Listgarten and Emili[4] explained how LC−MS data sets are subject to considerable noise that is not fully characterized and for which it is essential to devise statistically sound methods to distinguish signal from background.

While peak detection is often preceded by the characterization of the peak morphology,[5,6] noise rejection is seldom preceded by the characterization of the noise features. The model typically used or understood in the literature[2,3,7] to describe the signal is the following:

$$F(x) = B(x) + NS(x) + \varepsilon(x) \qquad (1)$$

where $F(x)$ is the observed signal, x can be either m/z or retention time t_R, $B(x)$ is the baseline, often described as a slowly varying trend under the spectra, $S(x)$ is the signal produced by the peptides present in the sample, N is a normalization factor, and $\epsilon(x)$ is a general term for the noise, which encompasses both the chemical background and the stochastic noise and which is usually modeled as a zero-mean Gaussian white noise.

In this model, the observed signal is simply decomposed as the sum of three independent components, but this assumption has proved too naive for a complete description of the data accounting for the possible correlations between the terms.

The lack of efforts to disentangle noise from signal is also reflected in the current software tools for data analysis. The most typical approach for noise reduction, in fact, consists of using digital filters to smooth or enhance the signal, and the rationale by which many authors motivate this choice is that the specific morphology of a peptide peak can be used to distinguish it and to separate it from the noisy background. Unfortunately, though, the pattern of a noise peak can be very similar to that of a peptide

* To whom correspondence should be addressed. E-mail: salvatore.cappadona@biomed.polimi.it. Phone: +39 02 2399 3322. Fax: +39 02 2399 3360.

[†] Politecnico di Milano.

[‡] Lund University.

(1) Liu, T.; Belov, M. E.; Jaitly, N.; Qian, W. J.; Smith, R. D. *Chem. Rev.* **2007**, *107*, 3621–3653.

(2) Morris, J. S.; Coombes, K. R.; Koomen, J.; Baggerly, K. A.; Kobayashi, R. *Bioinformatics* **2005**, *21*, 1764–1775.

(3) Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M. *Mass Spectrom. Rev.* **2006**, *25*, 409–449.

(4) Listgarten, J.; Emili, A. *Mol. Cell. Proteomics* **2005**, *4*, 419–434.

(5) Windig, W.; Phalp, J. M.; Payne, A. W. *Anal. Chem.* **1996**, *68*, 3602–3606.

(6) Andreev, V. P.; Rejtar, T.; Chen, H. S.; Moskovets, E. V.; Ivanov, A. R.; Karger, B. L. *Anal. Chem.* **2003**, *75*, 6314–6326.

(7) Coombes, K. R.; Tsavachidis, S.; Morris, J. S.; Baggerly, K. A.; Hung, M. C.; Kuerer, H. M. *Proteomics* **2005**, *5*, 4107–4117.

10.1021/ac800166w CCC: $40.75 © 2008 American Chemical Society
Published on Web 05/30/2008

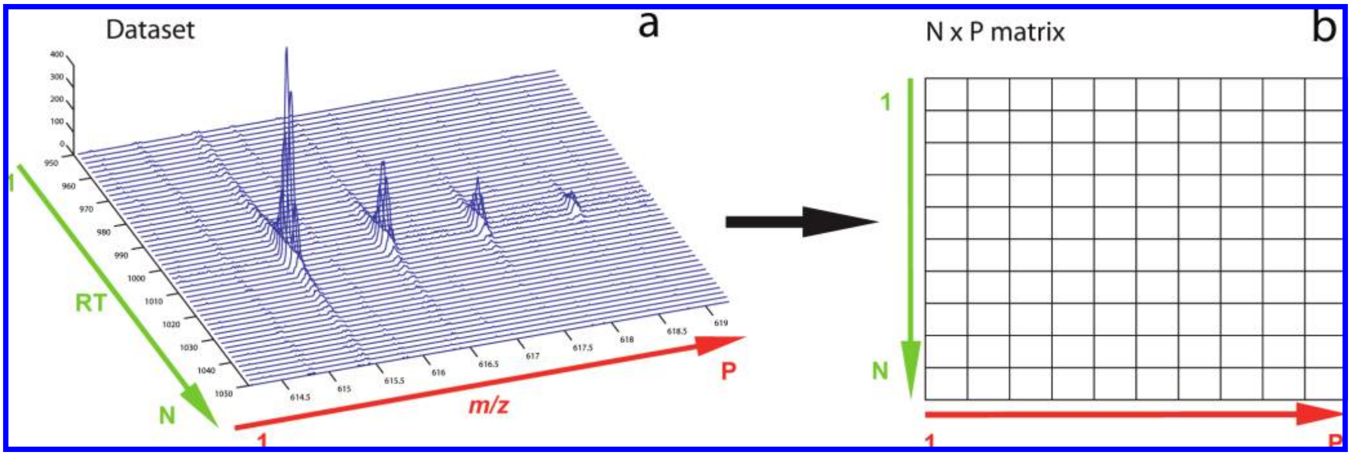


Figure 1. (a) Three-dimensional map of an experiment. (b) Matrix representation of the map: each row represents a specific scan and each column represents the elution trace at a specific *m/z,* also known as a single ion chromatogram (SIC).

peak, especially in the mass domain, and this issue becomes very critical at low concentrations, where noise peaks can cause either false positive or false negative identifications by mimicking or masking the signal.[6]

For these reasons, this study was focused on the preprocessing step of noise rejection and aimed at improving peak detection beyond the noise threshold, so as to unveil the so-called "hidden proteome".[8] The novel strategy devised to reach this goal took advantage of the two-dimensional nature of the data and minimized the noise in the mass spectra by processing the signal in the chromatographic domain, where morphological features allowed us to distinguish signal from noise. The suggested procedure has proved to efficiently distinguish signal from noise and to selectively reject background noise while preserving the signal from peptides found at low concentrations.

EXPERIMENTAL SECTION

Test Data Set. All samples analyzed as test data set contained a varying concentration of a tryptic digest of alcohol dehydrogenase I from *Saccharomyces cerevisae* (ADH; Sigma, Buchs, Switzerland). HPLC separation of peptides was performed on a Waters CapLC system (Waters, Manchester, UK). All data sets were produced using a 40-min gradient, from 0.1% formic acid in water to 70% acetonitrile, 0.1% formic acid in water. The HPLC was interfaced online with a Q-TOF Ultima API mass spectrometer (Waters), by means of an electrospray ion source. The MS scan range was m/z 400−1600, and the scan time was 1.9 s with 0.1-s interscan delay. The MS data were collected in continuum mode.

The test data set was purposely devised to tune the proposed denoising algorithm at decreasing concentrations of ADH. The first experiment of the data set was an LC−MS/MS run containing 1 pmol (total injected) of ADH. This run was meant to locate the m/z ratios and the retention times of a set of peptides univocally recognized as ADH fragments by a Mascot[9] search. The identified peptides were only a subset of the peptides actually present in the experiment, but were identified with a very good level of confidence and were thus chosen as our standard. The other experiments of the data set were a set of LC−MS runs containing the enzymatic digestion of ADH at decreasing concentrations (1 pmol and 300, 100, 30, 10, 3, and 1 fmol) and a blank run. The underlying hypothesis in devising the data set was that each experiment could be described as a blank run with the addition of the protein peptides. To satisfy this hypothesis, three strict constraints were imposed to the experimental setup. The first was to maintain exactly the same chromatographic state all over the runs, with a special attention to preserve the same gradient of acetonitrile, the same chromatographic column, and the same acquisition program. The second was to minimize the risk of physical noise, resulting from thermal fluctuations and instability of the ion source. The third was to use a sample of ADH as pure as possible, in order to avoid any contamination of the data.

Complex Data Set. The membrane fraction of *Bacillus subtilis* 1A1 was analyzed to validate the applicability of the proposed method on a real biological sample. The specimen was treated in the same way as described in Janson et al.[10] with the exception that 150 μg of sample was loaded on the SDS-PAGE gel and the digestion with proteinase K was carried out at pH 12.

After the in-gel digestion and extraction of peptides, the samples were analyzed on a Q-TOF Ultima API coupled to a Waters CapLC HPLC. The autosampler injected 6 μL of sample, and the peptides were trapped on a precolumn (C18, 300 μm × 5 mm, 5 μm, 100 Å, LC-Packings), and separated on a reversed-phase analytical column (Atlantis, C18, 75 μm × 150 mm, 3 μm, 100 Å, Waters) with the flow rate set to 200 nL/min. Solvent A consisted of 2% acetonitrile, 98% water with 0.1% formic acid. Solvent B consisted of 90% acetonitrile, 10% water and 0.1% formic acid. The HPLC method started at 5% B for 18 min, was then raised from 5 to 80% B over 57 min, from 80 to 100% B over 1 min, and held at 100% B for 25 min before reducing from 80 to 5% B in 1 min and re-equilibrating at 5% B for 15 min. The total run time was 115 min, the MS scan range was m/z 400−1600 and the scan time was 1.9 s with 0.1-s interscan delay.

Algorithm Development. All data were processed and all algorithms were written using MATLAB (v.7.0.4 for Windows; The

[8] Righetti, P. G.; Boschetti, E.; Lomas, L.; Citterio, A. *Proteomics* **2006**, *6*, 3980–3992.

[9] Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.

[10] Janson, M.; Wårell, K.; Levander, F.; James, P. *J. Proteome Res.* **2008**, *7*, 659–665.

Mathworks Inc., Natick, MA) and the MATLAB Wavelet toolbox. The computer configuration was a Pentium IV, 3.00 GHz with 1 GB of RAM.

Data Structure. A direct way to visually reconstruct the chromatographic dimension consisted of drawing each scan next to the previous one, to create a three-dimensional map in which the intensity of a peptide peak was a function of both its mass and its elution time in the liquid column of the chromatographer (Figure 1a).

To facilitate data manipulation, each LC−MS run was then rearranged from the instrument-specific format into a uniformly spaced matrix (Figure 1b), with rows representing the consecutive scans and columns representing all possible single ion chromatograms (SICs) at a specific m/z value. Since the scans were acquired at regular retention time intervals, the chromatographic dimension was uniformly spaced and did not need any processing. Conversely, the mass domain was not equally sampled, because TOF mass spectrometers apply a nonlinear transformation to the acquired data in order to determine the m/z values and since points with zero intensities are even omitted from the raw data in order to save space.

Three different approaches were tested to obtain uniformly spaced data also in the mass domain: bucketing of the m/z axis,[11] resampling of the mass spectra through a linear interpolation at constant increments of m/z, and converting the raw spectra from their original m/z domain back to their native time-of-flight domain[12] (where they are originally equally sampled), by manipulating the equation that describes the motions of the ionized peptides to reach the detector of the TOF analyzer:[13]

$$t^2 = \underbrace{\frac{m}{z}\left(\frac{d^2}{2V_s e}\right)}_{const} \Rightarrow \frac{m}{z} \propto t^2 \qquad (2)$$

The latter option was chosen as the only one that permitted us to retain all original values. It also permitted us to easily infer all missing null values from the TOF spectra, and this led to an equal number of samples in each scan, called number of clock ticks.

The final matrix used for all subsequent analysis was thus the matrix M_{NxP}, with N, total number of scans; P, total number of clock ticks; i, scan number; j, clock tick; $m_{i,j}$, intensity value for scan i at clock tick j; $m_{\bar{i},j}$, row vector, corresponding to the ith spectrum; and $m_{i,\bar{j}}$, column vector, corresponding to the jth SIC.

Wavelet Decomposition Analysis. Signal processing of the SICs was performed by means of wavelet decomposition. Wavelets are functions that are both localized in the time and the frequency domain. Their basic advantage over the classical signal processing transformations is their ability to process both time and frequency at once. A thorough presentation on the application of the wavelet transform (WT) in analytical chemistry can be found elsewhere.[14–17] WT analysis exploits the property of any signal S (of length N) to be decomposed into a sum of constituent functions, D_j, each representing the events in S that occur at a particular scale: $S = A_j + \sum_{j=1}^{J} D_j (J < \log_2 N)$. A_j represents the residual trend in S at the scale j. This hierarchical representation of the signal (Figure 2) allows rapid access to both the low-frequency and the high-frequency components of a chromatogram, respectively, at the high scales of the decomposition approximations and at the low scales of the decomposition details.

Here, the undecimated discrete wavelet transform was preferred to the discrete wavelet transform because it is translation invariant and thus practically unaffected by small shifts of the original signal.[18] Several families of wavelets were tested, but the effects on the decomposition were not strictly dependent on the particular wavelet chosen. The Coifmann wavelet of degree 1 was finally selected because its shape closely approximates the profile of a small chromatographic peak.

MODELING APPROACH

Noise Characterization. Noise in LC−MS data sets is mostly either random or chemical. Its origins are still under investigation, but seem mostly related to the detection of the LC mobile phase and buffers.[19] The random noise is represented by small spikes uniformly distributed in both the chromatographic and the mass domains. The chemical noise, in contrast, is known to present a different behavior in the two domains. In the mass domain, it has been described as a periodic background, which is difficult to remove because it has a pattern very similar to and often overlapping with that of the signal.[20] In the chromatographic domain, it appears rather as a slowly varying baseline[21] whose trend can differ significantly over contiguous SICs.

To characterize the chemical noise in both the chromatographic and the mass domains, we had to abstract away from the signal model presented in eq 2 and to reset every assumption that could bias our reasoning.

For this reason, a three-dimensional map of a whole experiment was thoroughly examined (Figure 3). This inspection confirmed both the periodicity of the elution traces of the chemical noise and the superposition of noise and signal. However, it also unveiled the morphological feature that was exploited afterward to discriminate between peptides and chemical noise.

As is clear from the figure, a SIC can be described by means of two simple features: presence or absence of a peptide peak and presence or absence of the chemical noise (being the random noise always present). These two considerations were the basis to formulate our new mathematical model to fully characterize a SIC and alternative to (1):

$$F(t) = S(t) + \varepsilon(t) + [B(t) + k_B \varepsilon(t)] \qquad (3)$$

In this model, $F(t)$ is the observed signal, $S(t)$ is the peptide signal, $B(t)$ is the baseline, completely ascribed to the chemical

[11] Stolt, R.; Torgrip, R. J. O.; Lindberg, J.; Csenki, L.; Kolmert, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2006**, *78*, 975–983.

[12] Carrillo, B.; Kearney, R. E.; Yanofsky, C.; Lekpor, K.; Bell, A.; Boismenu, D. *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2004**, *1*, 220–223.

[13] Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A. *J. Mass Spectrom.* **2001**, *36*, 849–865.

[14] Shao, X. G.; Leung, A. K.; Chau, F. T. *Acc. Chem. Res.* **2003**, *36*, 276–283.

[15] Alsberg, B. K.; Woodward, A. M.; Kell, D. B. *Chemom. Intell. Lab. Syst.* **1997**, *37*, 215–239.

[16] Walczak, B.; Massart, D. L. *Trends Anal. Chem.* **1997**, *16*, 451–463.

[17] Leung, A.; Chau, F.; Gao, J. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 165–184.

[18] Lang, M.; Guo, H.; Odegard, J. E.; Burrus, C. S.; Wells, R. O. *IEEE Signal Processing Lett.* **1996**, *3*, 10–12.

[19] Cech, N. B.; Enke, C. G. *Mass Spectrom. Rev.* **2001**, *20*, 362–387.

[20] Kast, J.; Gentzel, M.; Wilm, M.; Richardson, K. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 766–776.

[21] Danielsson, R.; Bylund, D.; Markides, K. E. *Anal. Chim. Acta* **2002**, *454*, 167–184.

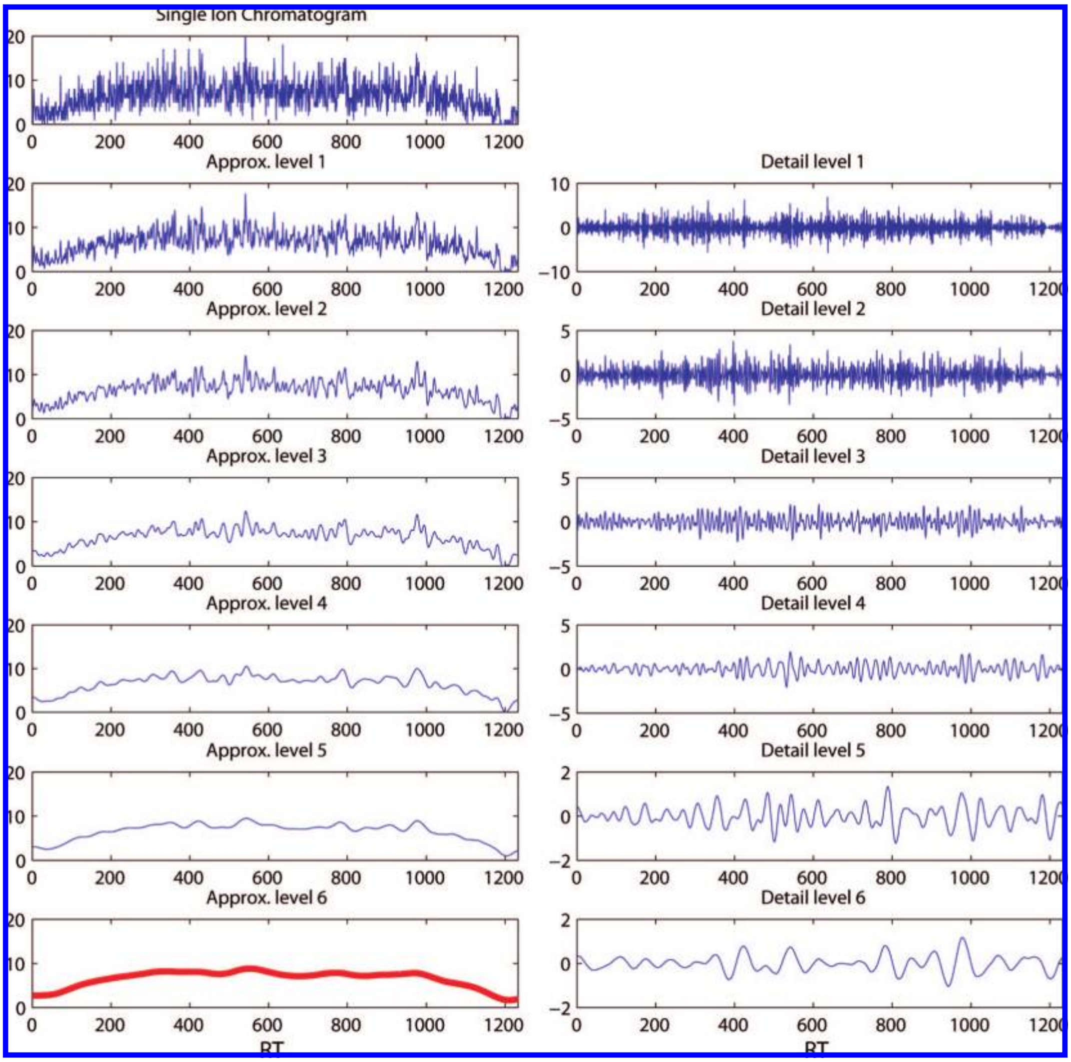


Figure 2. Wavelet decomposition of a single ion chromatogram from the first to the sixth level. The juxtaposition of approximations A_j and details D_j allows a coherent view of low- and high-frequency components at the different scales.

background, $\epsilon(t)$ is the stochastic noise, modeled as white noise, and k_B is a multiplicative constant, proportional to $B(t)$, which takes into account the heteroscedasticity of the stochastic noise. Given this model, four cases are possible:

$$F_1(t) = \varepsilon(t) \tag{F1}$$

$$F_2(t) = S(t) + \varepsilon(t) \tag{F2}$$

$$F_3(t) = \varepsilon(t) + [B(t) + k_B\varepsilon(t)] \tag{F3}$$

$$F_4(t) = S(t) + \varepsilon(t) + [B(t) + k_B\varepsilon(t)] \tag{F4}$$

Noise Rejection. Given the new model, it was clear that a good rejection of the noise could be achieved by subtracting the chemical baseline and filtering out the high-frequency stochastic noise from each chromatogram.

Many techniques have been described in the chemometric literature both for baseline subtraction[22,23] and for signal denoising.[24] Here wavelet analysis was applied to accomplish both tasks, respectively by means of wavelet smoothing and wavelet denoising.[25] These two filtering techniques are very similar and are both applied to the transformed data set prior to back-transformation to the signal domain. The difference between the two methods is that while smoothing removes the high-frequency components of the transformed signal regardless of amplitude, denoising removes the small-amplitude components of the signal regardless of frequency.

Wavelet Smoothing. Figure 2 shows that a smoothed version of a chromatogram could be directly accessed at the high scales of the decomposition. In particular, the sixth level of approximation worked very well for our signals, while higher levels oversmoothed the baseline, and lower levels still contained components of the peptide peaks.

It should be noticed that the choice of the level was empirical, since the optimal level of decomposition is strongly dependent on the length of the signal[26] and on the average elution time of

[22] Gan, F.; Ruan, G.; Mo, J. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 59–65.

[23] Sanchez-Ponce, R.; Guengerich, F. P. *Anal. Chem.* **2007**, *79*, 3355–3362.

[24] Mittermayr, C. R.; Nikolov, S. G.; Hutter, H.; Grasserbauer, M. *Chem. Intell. Lab. Syst.* **1996**, *34*, 187–202.

[25] Barclay, V. J.; Bonner, R. F.; Hamilton, I. P. *Anal. Chem.* **1997**, *69*, 78–90.

[26] Perrin, C.; Walczak, B.; Massart, D. L. *Anal. Chem.* **2001**, *73*, 4903–4917.

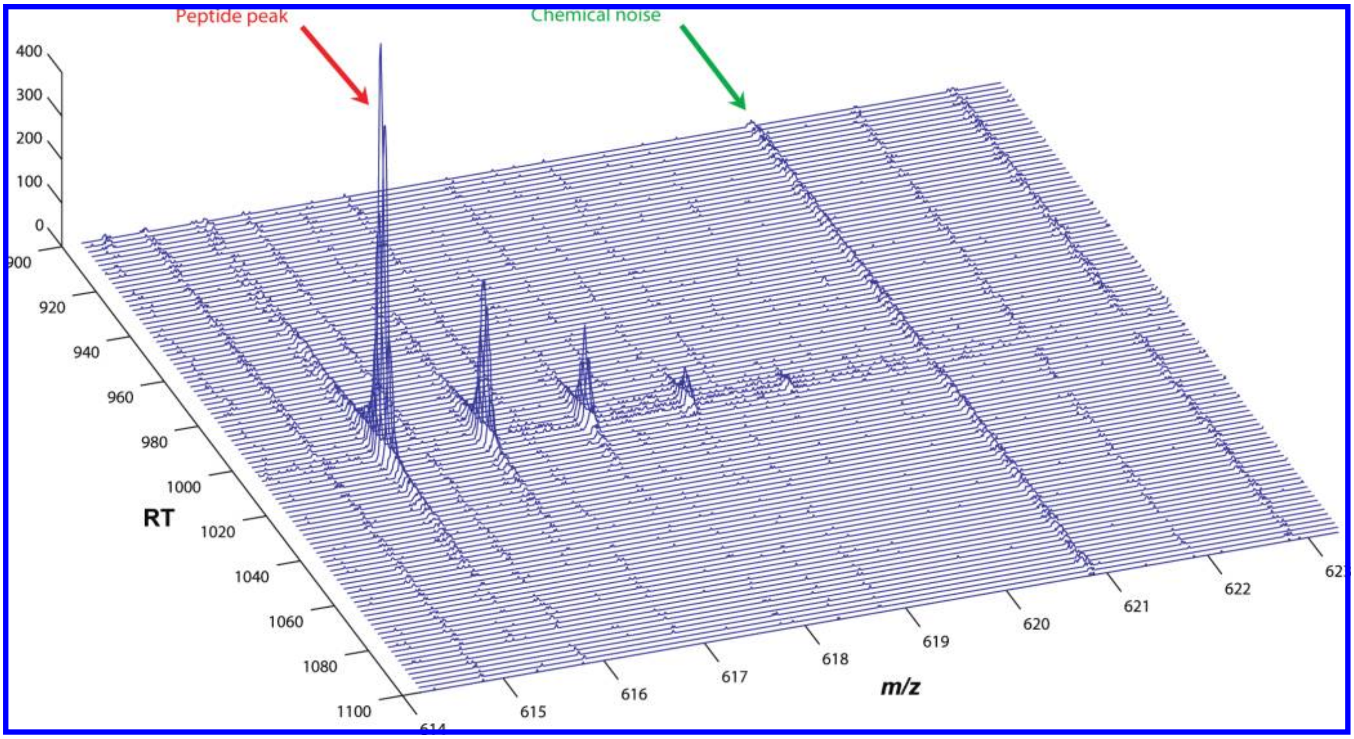


Figure 3. Elution trace of the chemical noise spread over the whole length of the LC−MS run. A peptide peak has a Gaussian profile limited over a few scans.

the peptides. The idea to identify the baseline with a particular approximation *Aj* of the decomposition and to subtract it from the signal had been already exploited.[27] The problem with this approach is that it introduces artifacts in the proximity of high-concentration peptide peaks, whose low-frequency component can be still detectable in the chosen approximation (Figure 4b). For this reason, we did not consider wavelet smoothing exhaustive.

Wavelet denoising is based on a technique called hard thresholding, which consists of fixing a threshold for each detail of the decomposition and setting to zero all the coefficients whose absolute value is below that threshold. The threshold value is usually calculated as a function of σ, the standard deviation of the stochastic noise. As we have already mentioned, the stochastic noise of our signals was essentially concentrated at the finest scale of the detail coefficients, cD_1. Thus, our problem could be reduced to calculating the standard deviation of cD_1. For the estimate of σ, we used the median absolute deviation (MAD) because it is very robust to the presence of outliers. It is worth noticing that, since the signal of interest was the baseline, the outliers were the transitory peptide peaks, whose possible occurrence in cD_1 was limited over a few scans. It should also be noticed that MAD allowed us to directly account for the heteroscedasticity of the noise, described in eq 3. In fact, SICs with MAD = 0 fell within cases F1 or F2 of our model (i.e., absence of chemical noise), and their stochastic noise was modeled as a zero-mean Gaussian white noise $N(0;1)$. In contrast, SICs with MAD $\neq$ 0 fell within cases F3 or F4 of our model (i.e., presence of chemical noise) and their stochastic noise was modeled as Gaussian white noise $N(0;\ \sigma)$, where $\sigma = \text{median}(\text{abs}(cD_1))/0.6745$ and 0.6745 is the 0.75 quantile of the standard normal distribution.

Figure 4 shows the comparison between the denoising and the smoothing steps and clearly evidences how the two methods yielded very similar results, except in the proximity of a peptide peak.

Strategy. Given the comparable results, we could formulate our strategy to approximate the baseline on the basis of the following steps: (1) smoothing of the original SIC and retention of the sixth level of approximation of the wavelet decomposition (Figure 5a); (2) denoising of the original SIC through hard thresholding (Figure 5a); (3) extraction of the common values (Figure 5b); (4) interpolation, by means of piecewise cubic hermite interpolating polynomial in the regions of the original SIC where the results of smoothing and denoising differ (Figure 5b).

The definitive filtered chromatogram could be simply obtained by subtracting the baseline from the wavelet denoised signal. The denoising step gave excellent results in finding the peptides' positions, but not in calculating their intensities, which were always underestimated. This issue could represent a limiting factor in terms of peptide quantification. Thus, we devised a further step, which extended and concluded our filtering strategy (Figure 5c). (5) Detection of the presumed peptide peaks (i.e., the points where denoised > baseline) and setting up of the filtered chromatograms is accomplished as follows: filtered = SIC − baseline, if SIC > baseline; filtered = 0, otherwise.

RESULTS AND DISCUSSION

Processing of a Single SIC. In Figure 5 is reported, as an example, the exit of the processing of a single SIC. The smoothed and the denoised signals, derived from the original SIC to obtain its filtered version, according to the scheme illustrated above, are plotted in Figure 5a, while the baseline is plotted in Figure 5b.

[27] Shao, X. G.; Cai, W. S.; Pan, Z. X. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 249–256.

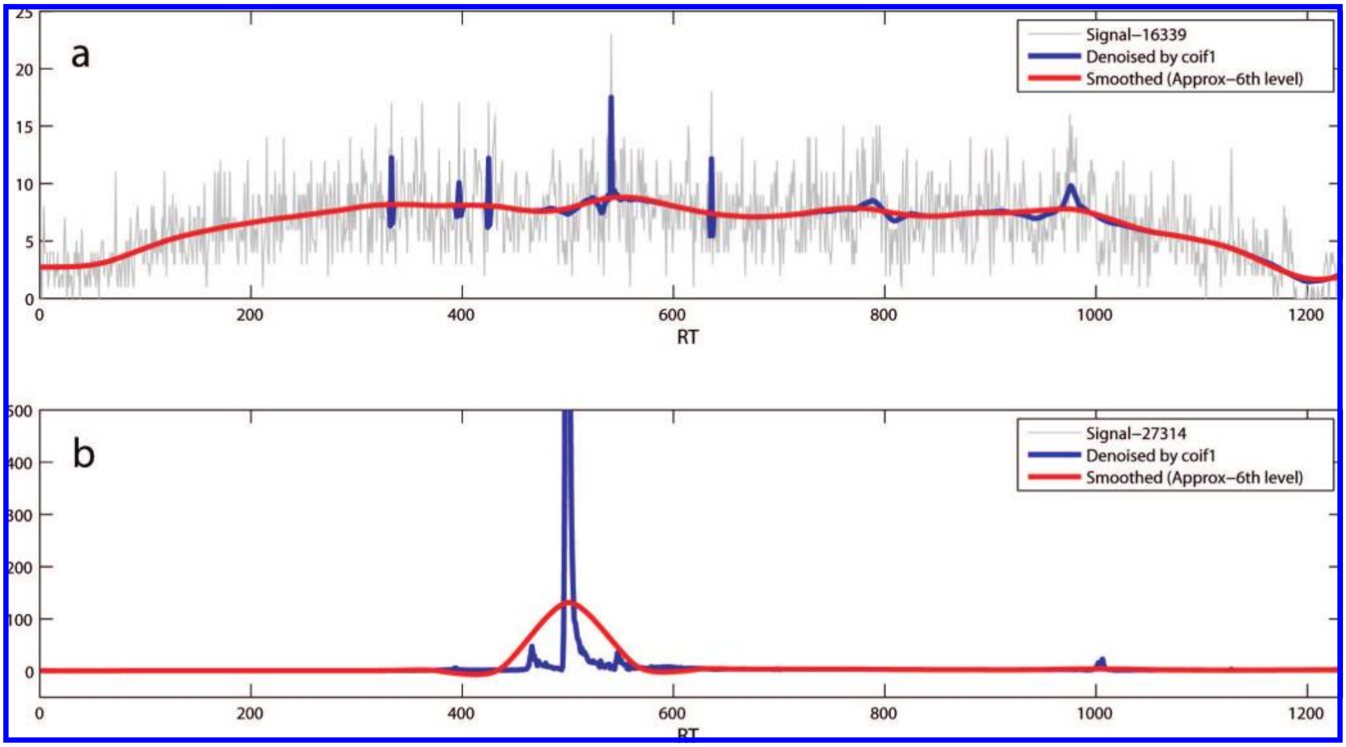


Figure 4. Comparison between wavelet smoothing and wavelet denoising. (a) The sixth level of approximation of the decomposition correctly identifies the baseline of a SIC. (b) The presence of a large peptide peak in the SIC is detected at the high scales of the decomposition. If we simply subtracted this approximation of the baseline from the original signal, we would completely lose the two peptide peaks close to the large one.

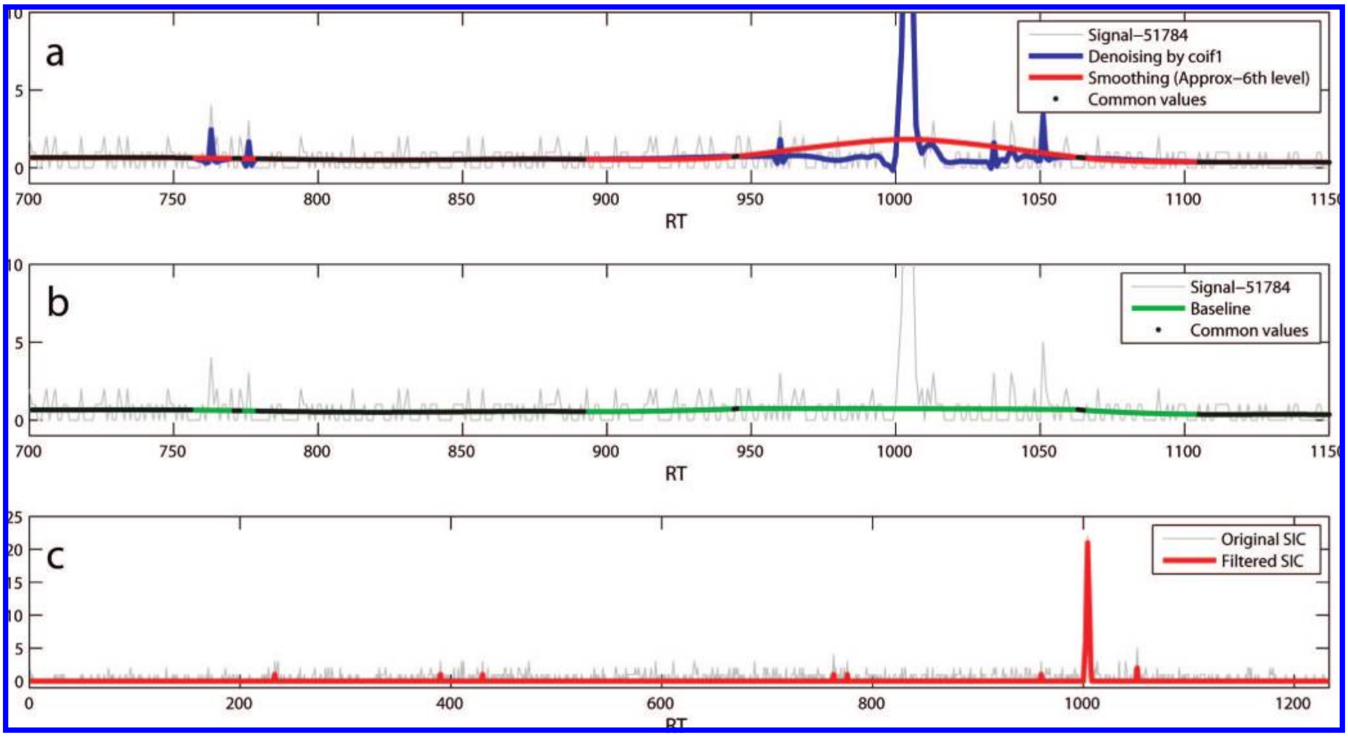


Figure 5. (a) Comparison between wavelet smoothing and wavelet denoising. (b) Extraction of the common values and approximation of the baseline. (c) The filtered chromatogram.

The definitive filtered chromatogram (Figure 5c) could be simply obtained by subtracting the baseline from the wavelet denoised signal, but the strategy adopted is preferable to prevent from underestimation of peak intensities.

Processing of a Whole LC−MS Run. Figure 6 shows how the iteration of the devised algorithm over all the SICs of a matrix permitted us to reject the chemical noise and the stochastic one from the entire LC−MS runs. It also shows that the algorithm

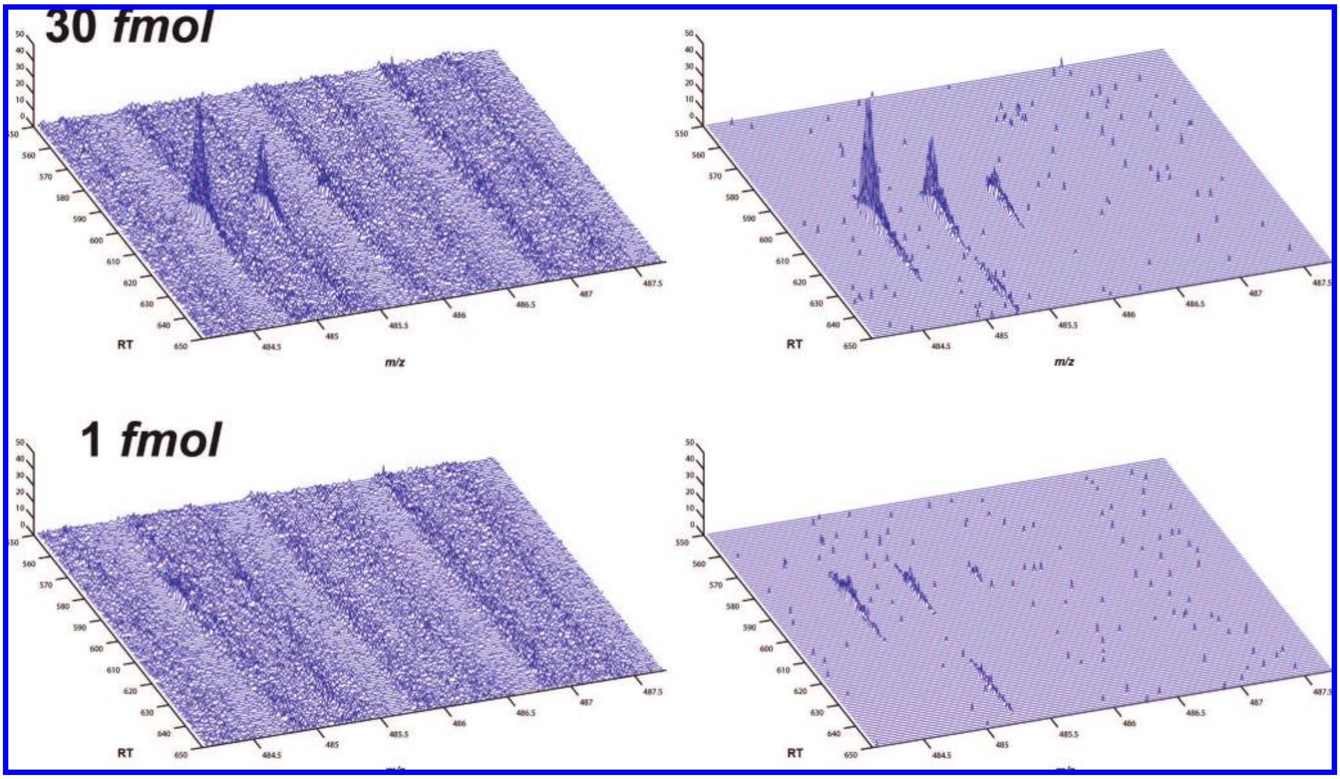


Figure 6. Noise rejection for the same peptide (the sixth from Table 1) at different concentrations. It is clear how the denoising step completely removed the stochastic noise, while the subtraction of the baseline rejected the chemical one. Obviously, the peaks shown at 30 fmol are high-concentration peaks and they do not present any detection problem. The strength of the filtering strategy becomes evident at the low concentrations, where the peptide peaks can be completely masked under the elution traces of the chemical noise.

Table 1. Peak Detection in the Processed (P) and the Unprocessed (U) Spectra[a]

		30 fmol			10 fmol			3 fmol			1 fmol		
	m/z	V	U	P	V	U	P	V	U	P	V	U	P
1	406.2224	√[b]	√	√	~130	X[c]	131−142, 146−148	X	X	X	X	X	X
2	407.7624	√	√	√	√	√	√	~500	X	499−505	~490	X	490−492
3	418.7267	√	√	√	√	√	√	√	√	√	~320	X	318−334
4	447.2879	√	√	√	√	√	√	~260	X	259−272	X	X	247
5	472.7679	~340	341−342	338−345	X	X	326, 330	X	X	X	X	X	X
6	484.7454	√	√	√	~585	X	582−590	~605	X	603−608	~585	585, 587	580−605
7	507.3065	√	√	√	√	√	√	~490	X	488−499	X	X	X
8	568.8013	√	√	√	√	√	√	X	X	X	X	X	X
9	626.3488	√	√	√	X	X	491	X	X	X	X	X	490−494
10	693.8858	√	√	√	√	√	√	~448	X	446−453	X	X	X
11	724.4133	~592	592	591−594	X	X	X	X	X	X	X	X	X
12	809.9336	X	X	X	X	X	X	X	X	X	X	X	X
13	540.2829	~520	521	519−526	~515	X	514−516	~530	X	527−530	~520	X	519−524

[a] The first column shows the peptides that are clearly detectable by visual inspection (V). The second and the third columns show the exact scan in which the peptides could be detected by means of Mascot Distiller, respectively before and after noise rejection. At 10 fmol, only 6 peptides were detected in the unprocessed spectra (U), while 11 became detectable in the processed (P) ones, with an improvement from 46 to 85%. At 3 fmol, the improvement was from 8 to 54%, and at 1 fmol it was from 8 to 46%. [b] √, clearly visible and detectable peptides. [c] X, invisible peptides.

worked as expected in selectively rejecting the chemical noise while preserving the low-concentration peptides. Clearly, at this point the cleaned matrices could also be used to retrieve the filtered mass spectra.

Both Figure 5c and Figure 6 show some isolated spikes, which are still present after denoising and which probably do not represent peptide peaks. These spikes could have been easily removed by increasing the denoising threshold, but we preferred to adopt a conservative approach. In fact, rather than risking loss of peptides, we preferred to retain some isolated spikes, which do not affect peaks and do not represent a problem in the detection phase, since they are not repeated over consecutive scans and they do not present an isotopic distribution.

Improvement of Peak Detection. To prove the efficacy of the denoising strategy at different concentrations, the algorithm was tested on the whole set of peptides recognized in the MS/MS run of the test data set. Figure 6 shows the effectiveness of the filtering algorithm even at the lowest concentration of our data set.

The performance of Mascot Distiller (Matrix Science, London, UK), a widely used peak detection software, was then evaluated

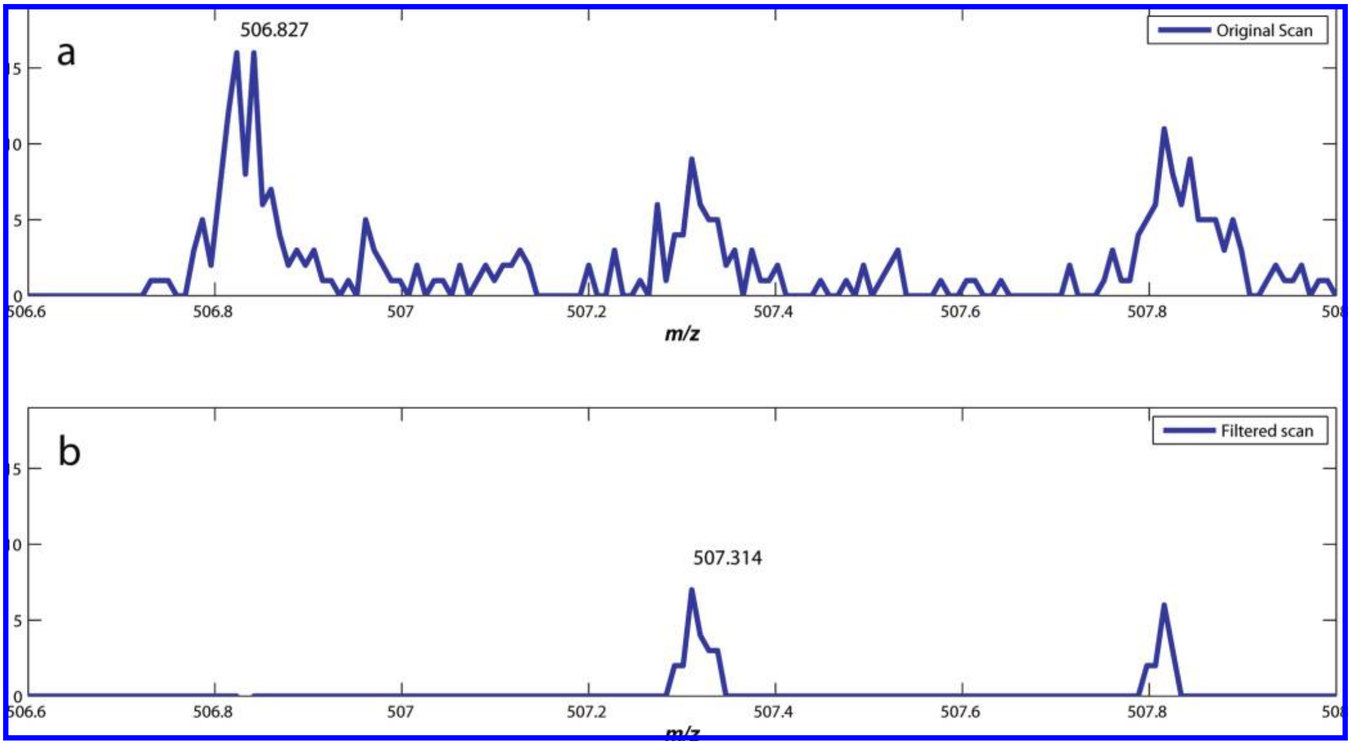


Figure 7. Misidentification of a peptide by Mascot Distiller.

on both the processed and the unprocessed spectra. The results are shown in Table 1. For each concentration, the first column of the table shows the peptides that were detected by visually inspecting the maps at the values of m/z and RT obtained from the Mascot search. At the highest concentrations, all of the peptides were clearly visible, and hence, these data have been omitted from the table. At decreasing concentrations, however, many of the peptides began to disappear "in the grass". At 30 fmol, for instance, three of them were hardly visible and one had completely disappeared. The situation became worse at the lower concentrations and, at 1 fmol, only four of the peptides were still visible. The second columns of the table show the exact scans in which the peptides could be detected by Distiller and demonstrate that the detection tool, when run on the unprocessed spectra, did not perform better than visual inspection. In contrast, the third column shows the detection results on the spectra filtered by our algorithm and evidence a clear improvement of the Distiller performances. In particular, it is evident that at the medium concentrations the improvement regarded the number of scans in which a peptide was detected, while at the lowest concentrations, we could detect even some peptides that had completely disappeared from the raw scans.

Detection of False Peaks. As we have stressed, a critical issue about noise in LC−MS data is that noise peaks can be morphologically identical to peptide peaks and this can cause false positive identifications. A typical example of a peak that can cause a misidentification is shown in Figure 7a. The unprocessed signal shows an isotopic distribution of three peaks at a distance of 0.5 Th (compatible with a 2+ peptide). After applying our filtering method, the first peak was recognized as chemical noise and eliminated from the signal (Figure 7b).

Before filtering, Distiller recognized the peak at 506.827 Th as the monoisotopic peak and the peak at 507.314 Th as the second peak of the isotope envelop. After noise removal, Distiller correctly identified the peak at 507.314 Th (one of the known peptides in our data set) as the monoisotopic peak. This means that the denoising algorithm could successfully improve specificity even in situations where a false peak is mimicking not only a correct distance from the other isotopes but also the correct intensity.

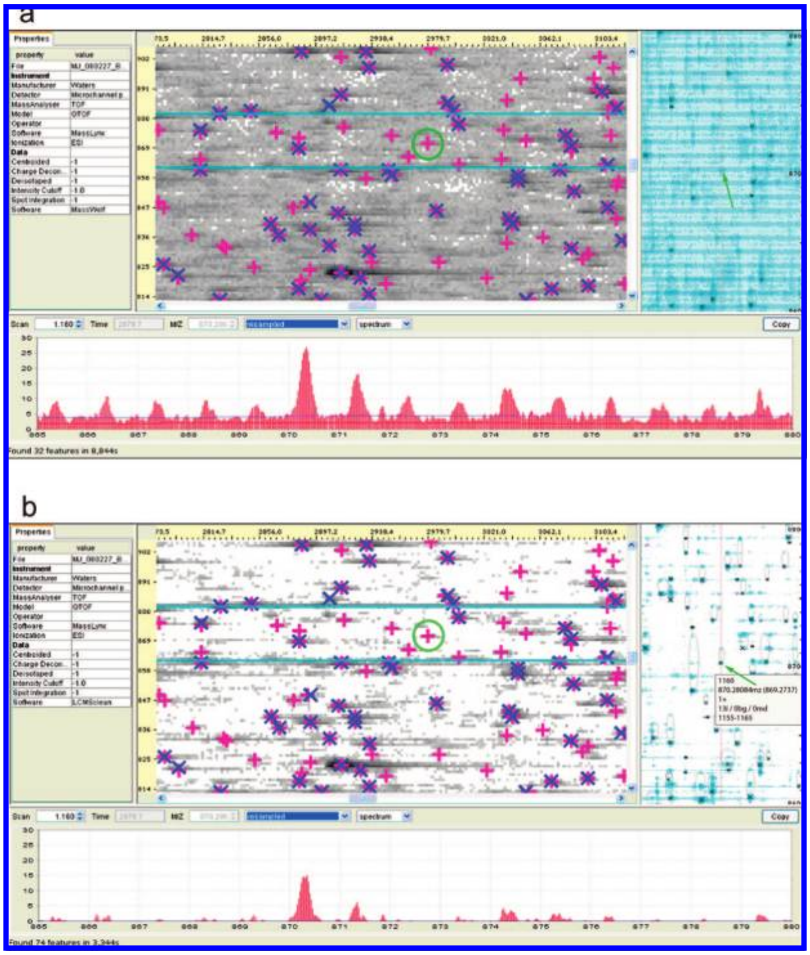
Extension to a Complex Sample. In this paper, we proposed a validation of our strategy on the base of a purposely devised test data set. This allowed us to follow the signal of a single purified protein at different concentrations. The proposed method, though, has performed efficiently also when applied to complex samples. The elution of a peptide peak in a SIC, in fact, is such a transient event in comparison to the total length of the chromatogram that the efficiency of the algorithm to estimate the baseline is practically unaffected even in the presence of many peptides.

The performance of msInspect,[28] a widely used peptide detection software, was evaluated in a highly crowded area of a complex sample, containing membrane proteins extracted from *B. subtilis* and digested by the unspecific enzyme proteinase K, which generates many closely related peptides that tend to coelute. Figure 8 demonstrates the improvement of a downstream peak detection analysis when performed after the application of the proposed method. The highlighted feature, for instance, is a typical example of a peptide that, in spite of its relatively high intensity, was completely masked by the chemical background and could be detected only after rejection of the chemical noise smears.

CONCLUDING REMARKS

We have developed a wavelet-based approach for noise rejection from HPLC−MS data sets. The method takes advantage of

[28] Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C. W.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *Bioinformatics* **2006**, *22*, 1902–1909.



a

property	value
File	MJ_080227_B
Instrument	
Manufacturer	Waters
Detector	Microchannel p
MassAnalyser	TOF
Model	QTOF
Operator	
Software	MassLynx
Ionization	ESI
Data	
Centroided	-1
Charge Decon..	-1
Deisotoped	-1
Intensity Cutoff	-1.0
Spot integration	-1
Software	MassWolf

Scan 1.160 Time MZ resampled spectrum Copy

Found 32 features in 8,844s

b

property	value
File	MJ_080227_B
Instrument	
Manufacturer	Waters
Detector	Microchannel p
MassAnalyser	TOF
Model	QTOF
Operator	
Software	MassLynx
Ionization	ESI
Data	
Centroided	-1
Charge Decon..	-1
Deisotoped	-1
Intensity Cutoff	-1.0
Spot integration	-1
Software	LCMSclean

Scan 1.160 Time MZ resampled spectrum Copy

Found 74 features in 3,344s

Figure 8. Noise rejection and peak detection in a complex sample, by means of msInspect. The software layout provides three different views of the data: a two-dimensional map of the experiment (upper left), a closeup of a selected area (upper right), and the profile of a single *m*/*z* spectrum (bottom). The blue crosses show the peptides found in the unprocessed data (a), and the pink crosses show the peptides found on the processed data (b).

the two-dimensional nature of the data to minimize the noise in the mass spectra by processing the signal in the chromatographic domain, where morphological features allow us to better distinguish signal from noise.

We have compared the performance of two widely used peak detection software (Mascot Distiller and msInspect) on both processed and unprocessed spectra. The analysis shows that the devised method allows a significant increase in sensitivity and specificity over uncleaned data, even at very low peptide concentrations. This result is obtained by unveiling the peptide peaks that are masked by the chemical noise and by removing the noise peaks that are mimicking the peptide signal.

Treatment of the data by the devised procedure is associated with a decrease of both chemical and random noise but does not affect the peptide peaks present in the data. For this reason, we believe that the method may represent an effective preprocessing strategy upstream to the peak detection for protein identification and in general to the application of data mining techniques on LC−MS maps.

ACKNOWLEDGMENT

This work was supported by grants from the Swedish Strategic Research Council (SSF) to CREATE Health (P.J.) and from the Foundation BLANCEFLOR Boncompagni-Ludovisi to S.C. We thank Mats Mågård and Liselotte Andersson for expert laboratory assistance and Marco Soria and Giorgio Arrigoni for helpful discussions.

Received for review January 23, 2008. Accepted April 4, 2008.

AC800166W