

Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform

Kevin R. Coombes¹, Spiridon Tsavachidis¹, Jeffrey S. Morris¹, Keith A. Baggerly¹, Mien-Chie Hung², and Henry M. Kuerer³

Departments of ¹Biostatistics and Applied Mathematics, ²Molecular and Cellular Oncology, and ³Surgical Oncology, The University of Texas M. D. Anderson Cancer Center

Running Title: Improved Quantification of SELDI Spectra Using Wavelets

Keywords: mass spectrometry, SELDI, peak detection, peak quantification, wavelets, denoising, undecimated discrete wavelet transform, reproducibility

Contact Address:

Kevin R. Coombes

Department of Biostatistics and Applied Mathematics, Box 447

The University of Texas M. D. Anderson Cancer Center

1500 Holcombe Blvd.

Houston, TX 77030

Email: krc@odin.mdacc.tmc.edu

Phone: 713-794-4154

FAX: 713-745-4940

Abstract

Background: Mass spectrometry, especially surface enhanced laser desorption and ionization (SELDI) is increasingly being used to find disease-related proteomic patterns in complex mixtures of proteins derived from tissue samples or from easily obtained biological fluids such as serum, urine, or nipple aspirate fluid. Questions have been raised about the reproducibility and reliability of peak quantifications using this technology. For example, Yasui and colleagues opted to replace continuous measures of the size of a peak by a simple binary indicator of its presence or absence in their analysis of a set of spectra from prostate cancer patients.

Methods: We collected nipple aspirate fluid from breast cancer patients and from healthy women. Samples were pooled to form a single quality control (QC) sample, separated into aliquots, and stored. We extracted protein from an aliquot of the QC sample and hybridized it to two spots on each of three SELDI ProteinChip arrays using weak cation exchange (WCX2; CIPHERGEN). The experiment was repeated on four successive days. We developed a novel algorithm for low level processing of SELDI spectra, including denoising with the undecimated discrete wavelet transform (UDWT), baseline correction, peak detection and quantification. We evaluated this algorithm for consistency and reproducibility across the 24 replicate spectra.

Results: The UDWT provides a computationally efficient method for decomposing mass spectra into a noise component and a signal component consisting of peaks and baseline. The noise levels are consistent and mostly uncorrelated across spectra. The subsequent baseline correction step yields spectra consisting of isolated peaks or peak clusters separated by flat regions. Our method detects more peaks than the method implemented in the CIPHERGEN software, and the peaks it detects are reproducibly found across replicate spectra. After normalization to the total ion current and log transformation, the mean coefficient of variation of the peak heights is 10.6%. Software to implement these methods is freely available.

Conclusions: The method for low-level processing of SELDI spectra described here provides substantial improvements over existing methods. In particular, denoising spectra using the

UDWT appears to be an important step toward obtaining more accurate results. It both improves the reproducibility of the peak quantifications and supplies tools that will make it possible to investigate the variations in the technology more carefully.

1 Introduction

Mass spectrometry, especially surface enhanced laser desorption and ionization (SELDI), is increasingly being used to find disease-related proteomic patterns in complex mixtures of proteins derived from tissue samples or from easily obtained biological fluids such as serum, urine, or nipple aspirate fluid [Paweletz et al., 2000; Paweletz et al., 2001; Wellmann et al., 2002; Petricoin et al., 2002; Sauter et al., 2002; Adam et al., 2002; Adam et al., 2003; Zhukov et al., 2003; Schaub et al., 2004]. These proteomic patterns can potentially be used for early diagnosis, to predict prognosis, to monitor disease progression or response to treatment, or even to identify which patients are most likely to benefit from particular treatments.

Ciphergen Biosystems (Fremont, CA) developed the SELDI technology, which enables selective protein retention using distinct, factory-prepared, chromatographic surfaces [Hutchens and Yip, 1993]. In brief, SELDI technology begins by applying a biological sample to a precoated stainless steel slide. The coating enhances the surface to preferentially bind a specific class of proteins based on their physiochemical properties. Different coatings yield different “chip types” that bind different classes of proteins. The company produces a variety of ProteinChip® arrays, including reverse phase, cation exchange, anion exchange, and immobilized metal affinity surfaces. By selling both ProteinChip arrays and an inexpensive mass spectrometry instrument, Ciphergen has made proteomics tools available to a wide variety of biological and clinical researchers.

When it is applied to the ProteinChip array, the biological sample is mixed with an energy absorbing matrix (EAM) such as sinapinic acid, which causes the mixture to crystallize as it dries. The array containing the sample is then placed into a vacuum chamber, and the crystal is hit with light pulses from a nitrogen laser. The matrix molecules absorb energy from the laser and transfer it to the proteins. This causes the proteins to desorb and ionize, resulting in a cloud of ionized protein molecules in the gas phase. Next, a brief electric field is applied, which accelerates the ionized proteins into a flight tube where they drift until they strike a detector that records the time of flight. Given the known length of the tube and the applied voltage, a quadratic transformation can be used to derive the mass-to-charge ratio (m/z) of the protein from

the time of flight. The spectral data that results from this experiment consists of the sequentially recorded numbers of ions arriving at the detector (the intensity) coupled with the corresponding m/z values. Peaks in the intensity plot ideally correspond to individual proteins.

Questions have been raised about the reproducibility and reliability of peak quantifications using SELDI technology [Sorace and Zhan, 2003; Baggerly et al., to appear]. For example, Yasui and colleagues [2003] opted to replace continuous measures of the size of a peak by a simple binary indicator of its presence or absence in their analysis of a set of spectra from prostate cancer patients. More ominously, Rogers and colleagues [2003] have reported that the sensitivity and specificity of a neural-network-based classifier using SELDI data fell from initial values of 81.8–83.3% on an independent validation set contemporaneous with the training set to 41.0–76.6% on a second independent validation set studied ten months later. These disturbing results continued to hold in qualitative form even when a binary classification was used for the peaks.

While we certainly believe that there is substantial variability in the technology, and that changes in data acquisition protocols or parameters can drastically affect mass spectrometry profiles, we also believe that part of this variability arises from the application of inadequate algorithms for the low-level processing of mass spectrometry data. This low-level processing involves a number of complicated steps. The goal is to identify the locations of peaks and to quantify their sizes accurately. To reach this goal, one must typically remove baseline artifacts (attributable to the matrix molecules that form an integral part of the technology), normalize across spectra, and separate the true signal from the underlying noise (both electrical and chemical). These processing steps interact in complicated ways. For example, one common normalization method is to divide by the “total ion current”, which is obtained mathematically by summing the observed intensities over all (or a large portion) of the spectrum. This method is motivated by the idea that the total ion current is a surrogate for the total amount of protein in the sample being measured. If the baseline correction algorithm starts with the raw spectrum and ensures that the corrected signal never becomes negative, however, then electronic noise contributes a substantial portion of the total ion current, and one normalizes, in part, to the noise.

Statistically, the low-level processing of mass spectra reduces to decomposing the observed signal into three components: true signal, baseline, and noise. One might try to decompose a spectrum using a model represented schematically by the equation

$$f(t) = B(t) + N * S(t) + \epsilon(t)$$

where $f(t)$ is the observed signal, $B(t)$ is the baseline, $S(t)$ is the true signal, N is a normalization factor, and $\epsilon(t)$ is the noise. At present, this model is of limited utility, since we do not have an effective characterization of the individual components. The true signal can, in principle, be modeled as a sum of independent, possibly overlapping, peaks, each corresponding to one protein. Approximate shapes of the peaks can be estimated empirically by simulating the physical process by which a time-of-flight (TOF) mass spectrometer collects data. White noise is a plausible model for the final term in the model, based on the notion that it arises primarily from electronic noise in the detector. We do not, however, have a good theoretical model for the baseline, aside from the vague intuition that it consists of a very low frequency component of the observed signal. This intuition is difficult to use without making it more precise, because the shape of the true peaks changes within a spectrum, becoming significantly lower and broader at later times and higher masses.

We have previously described practical methods for the low-level processing of mass spectra [Coombes et al 2003; Baggerly et al, 2003]. Those methods tried to deal with the interactions between processing methods (and the intertwining of the three components of the signal) by iteratively attempting to isolate one component at a time. Although those methods looked promising, they suffered from two weaknesses. First, they were computationally intensive, making them an impediment to rapid, interactive processing of large sets of spectral data. Second, they were forced upon us because we had not identified effective tools for isolating at least one of the three components of the observed signal in a single pass.

In this paper, we present an improved method for the low-level processing of SELDI spectra. Our method relies on the undecimated discrete wavelet transform (UDWT) as a first step to isolate the noise component of a spectrum [Lang et al., 1995]. Wavelets have been used previously to denoise signals in a number of contexts, including capillary electrophoresis, magnetic resonance imaging, ultrasound blood flow, microneurography, and computed

tomography [Liu et al., 2003; Bao and Zhang, 2003; Placidi et al., 2003; Olhede and Walden, 2003; Diedrich et al., 2003; Harpen, 1999]. Qu and colleagues [2003] used an orthogonal (decimated) discrete wavelet transform (DWT) to study SELDI data derived from the serum of prostate cancer patients (see also Chau and Leung, 2000). The orthogonal DWT is extremely efficient computationally, but it is not shift-invariant. Thus, its denoising performance can change drastically if the starting position of the signal is shifted. The UDWT, by contrast, is shift-invariant. It has been reported to yield better visual and qualitative denoising, with a small added cost in computational complexity [Lang et al., 1996, Kamath et al., 2002]. In this paper, we show that the UDWT can be used to denoise mass spectrometry SELDI data. After denoising in this manner, separating background from true signal is considerably easier, and peaks can be rapidly identified and precisely quantified.

After preprocessing a set of n spectra and identifying and quantifying p peaks per spectrum, we are left with an $n \times p$ matrix of “peak expression levels”. This matrix is conceptually similar to the $n \times p$ matrix of gene expression levels produced by a typical microarray experiment. After proper preprocessing, then, SELDI data can be analyzed using many of the tools that have already been developed for microarrays, although we suspect that it will be important to take into account the elaborate covariance structures in the peak expressions resulting from common biological and chemical modifications of proteins. Inadequate or incorrect preprocessing methods, on the other hand, can result in data sets that exhibit substantial biases and make it difficult to reach meaningful biological conclusions [Sorace and Zhan, 2003; Baggerly et al., to appear].

2 Methods

2.1 Biological samples and generation of spectra

Samples of nipple aspirate fluid (NAF) from breast cancer patients and from healthy women were collected and prepared as described previously [Coombes et al., 2003]. Samples of cancer- and noncancer-associated NAF were pooled to create a quality control (QC) sample, divided into aliquots, and stored at -80°C . Three weak cation exchange ProteinChip arrays (WCX2), each containing eight spots, were used for experiments on 4 successive days. On each day, an aliquot of the QC sample was incubated on two previously unused spots on each array and proteomic spectra were generated. All samples were prepared by the same technician and hand-spotted on the arrays. Each spot was used to produce spectra in a low mass range (up to 30,000 Daltons), with settings optimized for the range from 3,000 to 20,000 Da. Sixty-five shots were averaged using automatic data collection protocols in the Peaks (CIPHERGEN) software program. Spectra were calibrated on a mass calibration curve constructed from bovine insulin (5,733.6 Da), bovine cytochrome *c* (12,230.9 Da), and equine myoglobin (16,951.5 Da).

2.2 Default spectral processing

Spectra were processed using CIPHERGEN ProteinChip Software, version 3.2, which is widely used since it is bundled with the instrument. The algorithms used by the software have been described previously [Fung and Enderwick, 2002]. We used the factory default settings for filtering, baseline correction, and normalization. We performed peak detection both with the default settings and with the sensitivity slider reset to maximum sensitivity. Raw spectra were exported from the CIPHERGEN software in XML format for analysis using other tools. The XML files containing the raw spectra are available from our web site (<http://bioinformatics.mdanderson.org/pubdata.html>).

2.3 Improved spectral processing

Wavelet denoising was accomplished using the UDWT as implemented in version 2.4 of the Rice Wavelet Toolbox (RWT), which is available from their web site (<http://www-dsp.rice.edu/software/rwt.shtml>). The RWT is a supplement to the mathematical software

package MATLAB (The MathWorks Inc., Natick MA). Our guidelines for selecting the parameters for wavelet denoising are described in the Results section. Because the implementation requires the inputs to have length equal to a multiple of a power of 2, we padded each spectrum to the nearest multiple of 1024 by reflecting the end of the spectrum. We used hard thresholding in the wavelet domain, based on a multiple of the median absolute deviation.

We performed baseline correction (by fitting a monotone local minimum curve) and normalization (dividing by the total ion current in a given mass range) using MATLAB scripts developed in house. We detected and quantified proteomic peaks using a MATLAB implementation of a streamlined version of the Simple Peak Finding algorithm described previously [Coombes et al., 2003]. This algorithm locates all local maxima in each denoised, baseline-corrected, normalized spectrum. The height of the peak is used to quantify peaks. For this purpose, a local maximum is defined as a point where the intensities change from increasing to decreasing (allowing for flat plateaus when the tops of peaks are more than one clock tick in width). The signal-to-noise ratio of a peak is estimated as the height above baseline divided by a median-smoothed version of the wavelet-defined noise. To match peaks across spectra, we pooled the list of detected peaks and combined peaks that differed in location by no more than 7 clock ticks or in relative mass by 0.3%. All MATLAB scripts used for the analysis are available from our web site (<http://bioinformatics.mdanderson.org/software.html>).

3 Results

Consistency of mass calibrations across samples was assessed visually by plotting false-color heat maps of samples by clock ticks on both the original scale and a logarithmic scale (data not shown). Prominent peaks showed up as vertical bands in these plots; the bands appeared to be well-aligned, indicating that there was no need for further calibration in this data set.

3.1 *Selecting parameters for UDWT-based denoising*

There are two basic parameters that must be set in order to use the UDWT to denoise the spectra: which wavelets to use and what threshold level to set. We tried several different wavelets, and found that Daubechies wavelets of degree 6 – 20 gave comparable results (data not shown). We selected a Daubechies wavelet of degree 8 for all further analysis.

In general, we expect information about the true peaks in the time domain to be represented by a small number of (relatively large) coefficients in the wavelet domain, and we expect noise to be distributed (at low levels) over most wavelet coefficients. These expectations drive our denoising strategy. We start by transforming from the time domain to the wavelet domain. Next, we compute the median absolute deviation (MAD) of the wavelet coefficients, which yields a robust estimate of their variability. We then set to zero all coefficients below some threshold (expressed as a multiple of 0.67 MAD in the RWT implementation), and transform back to the time domain. Two variants of thresholding are discussed in the wavelet literature. With hard thresholding, all coefficients below the threshold are set to zero, and all coefficients above the threshold are retained unchanged. With soft thresholding, the coefficients above the threshold are shrunk toward zero by subtracting the threshold value. When using an orthogonal DWT, hard thresholding is known to give better l_2 performance while soft thresholding has better smoothness properties. When using the UDWT, it has previously been observed that hard thresholding simultaneously retains good smoothness properties and good l_2 performance (Lang et al., 1995). By visual inspection of several samples, we confirmed that hard thresholding retained more of the true peak shape than soft thresholding in our spectra (data not shown).

We also examined the effects of the threshold parameter, which determined the amount of smoothing. We made plots of the residuals showing the difference between the original raw spectrum and the processed smooth spectrum for different threshold values (Figure 1). The scale of the noise increased with increasing threshold, which presumably occurred because the higher thresholds removed more of the true peak signal and added it to the noise component. This trend was particularly evident at the highest threshold considered. When we used a threshold of 40, several clearly visible peaks remained in the noise spectrum, especially in the region from 12,000 to 14,000 clock ticks. A threshold of 20 removes a noticeable fraction of those same peaks.

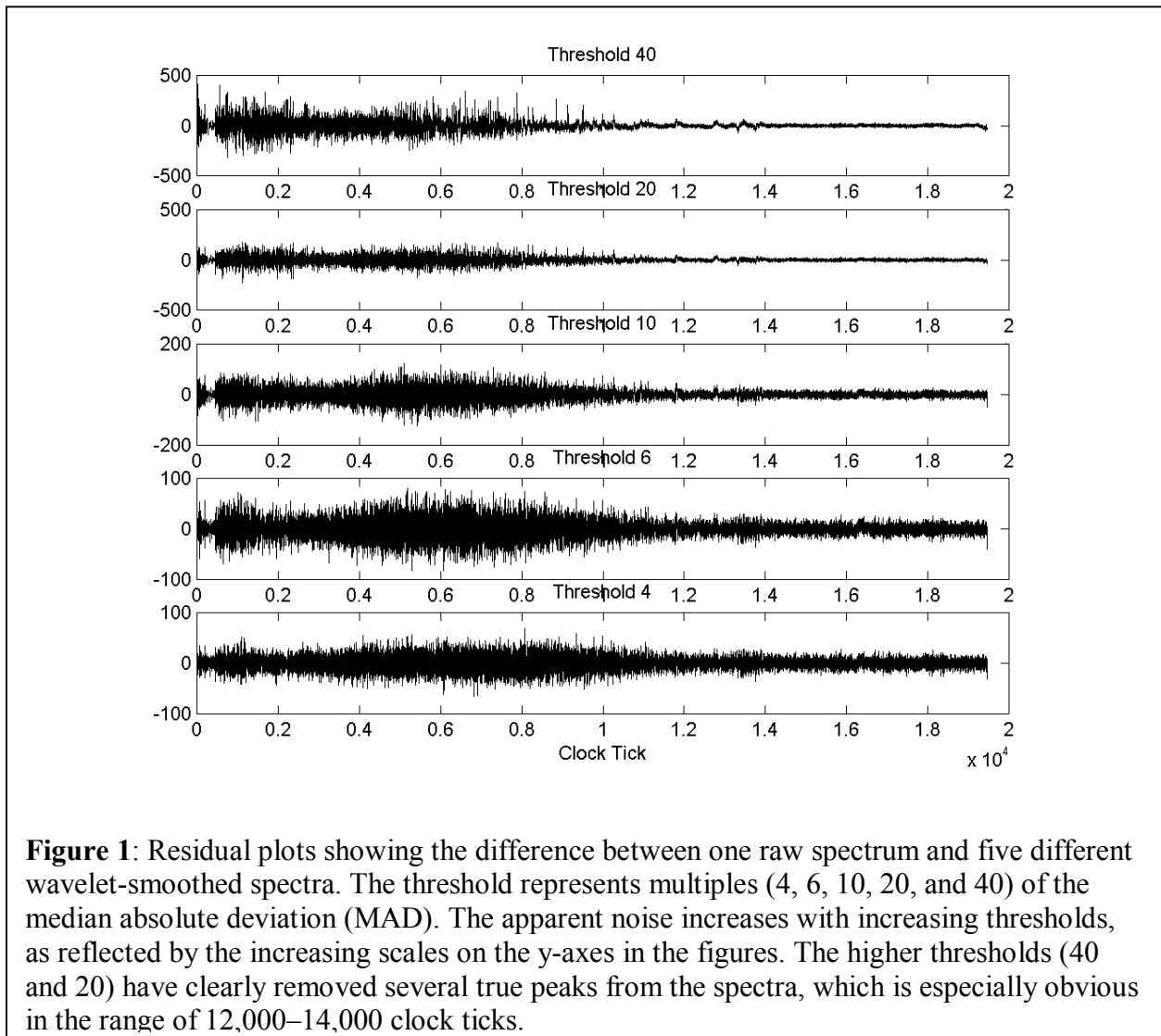


Figure 1: Residual plots showing the difference between one raw spectrum and five different wavelet-smoothed spectra. The threshold represents multiples (4, 6, 10, 20, and 40) of the median absolute deviation (MAD). The apparent noise increases with increasing thresholds, as reflected by the increasing scales on the y-axes in the figures. The higher thresholds (40 and 20) have clearly removed several true peaks from the spectra, which is especially obvious in the range of 12,000–14,000 clock ticks.

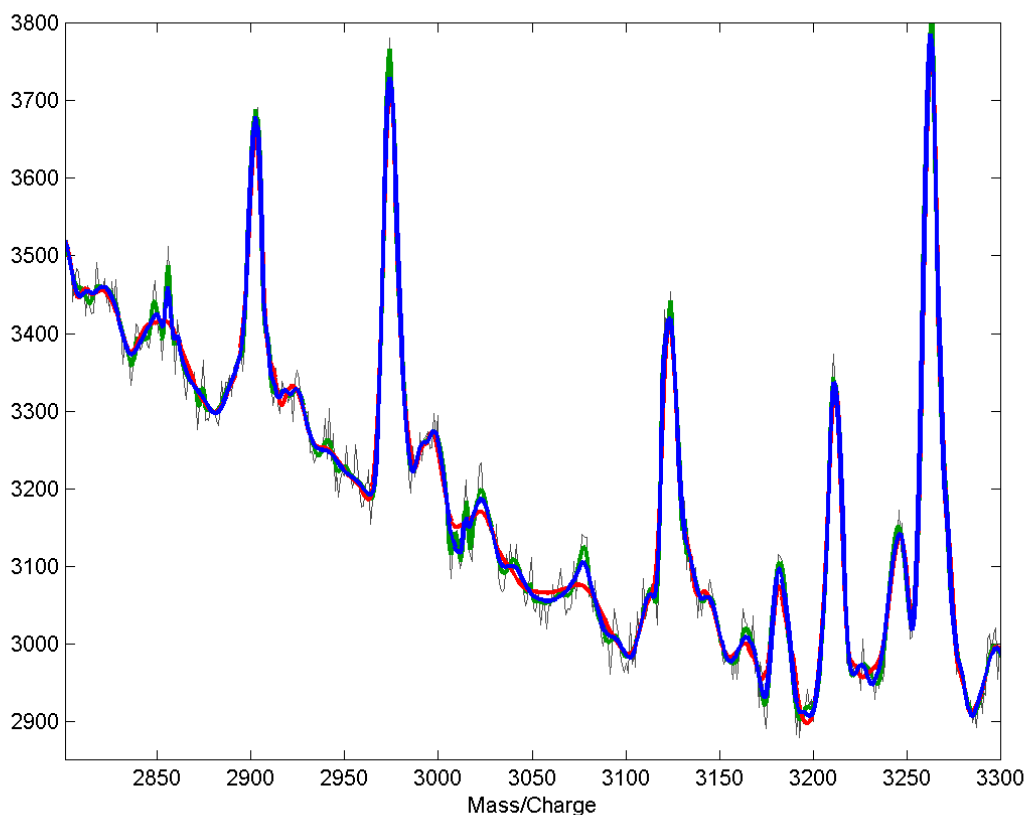


Figure 2: Plot of a portions of a raw spectrum (gray) with overlays of wavelet smooths using three different thresholds (green = 4, blue = 6, red = 10). With the threshold set at 4, the method undersmooths the data. Thresholds of about 6–10 appear to give the right amount of smoothing.

To examine more closely the effects of these parameters, we plotted a single raw spectrum with the five smoothers produced by three levels (4, 6, 10) of the threshold parameter (Figure 2). We concluded that setting the threshold at 10 is likely to give accurate smoothing without overfitting spurious features of the data. In applying this method to other SELDI data sets, we have found that threshold levels in the range 6–10 seem to consistently work well (data not shown). It seems likely that different threshold values will be required for spectra optimized over different mass ranges or for spectra generated using other mass spectrometry instruments, but we have not yet analyzed this problem completely.

3.2 The noise level is consistent across spectra

We processed all 24 spectra with the UDWT using a threshold of 10 and plotted a heat map of the noise (Figure 3). The general noise level appeared to be nearly the same across spectra, suggesting that this represented additive noise introduced by the instrument. In most regions, the noise appeared to be uncorrelated across spectra. However, there are regions of faint vertical banding at later clock ticks (or higher mass levels), suggesting that a few small peaks were consistently smoothed away at this threshold level.

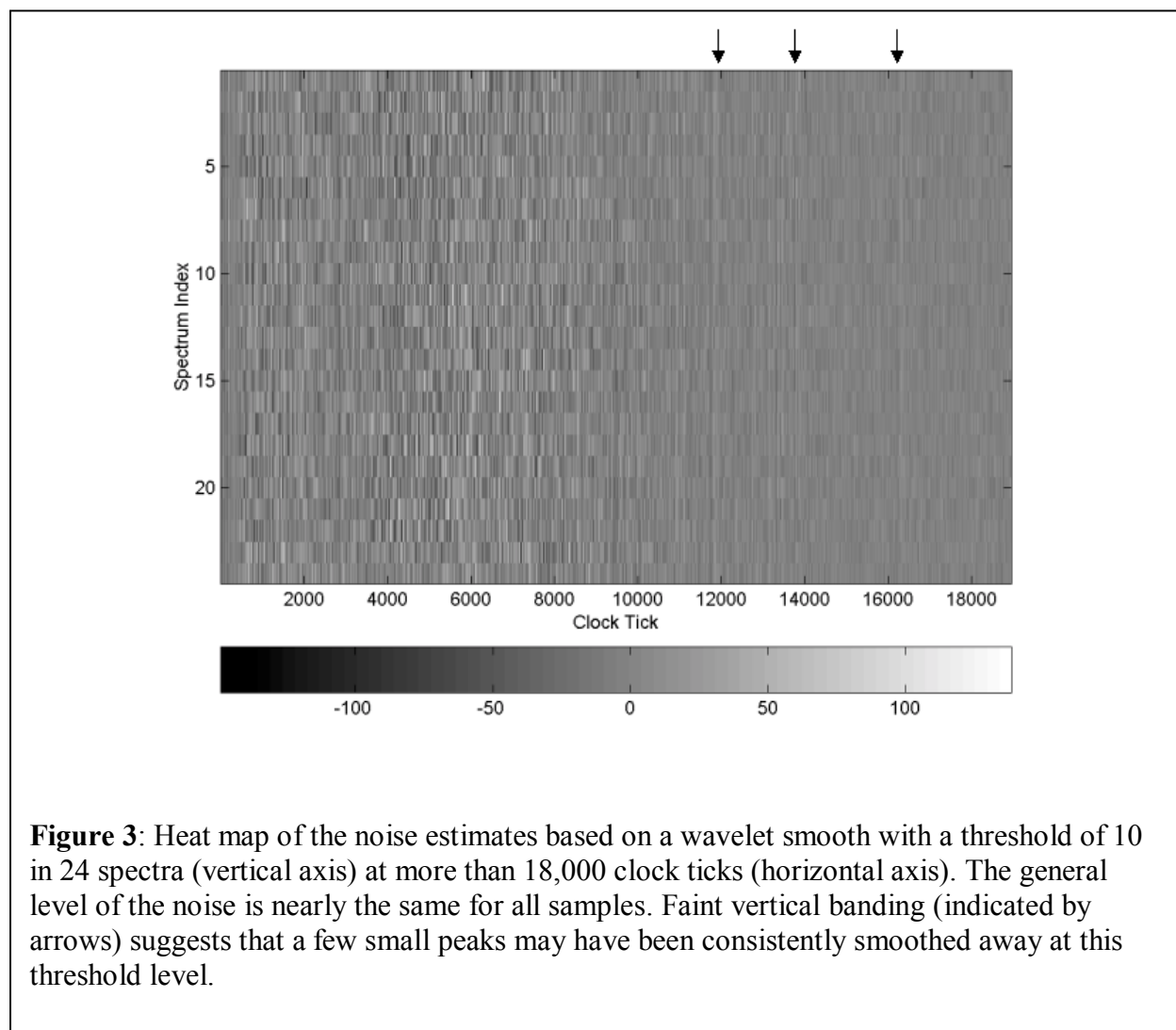
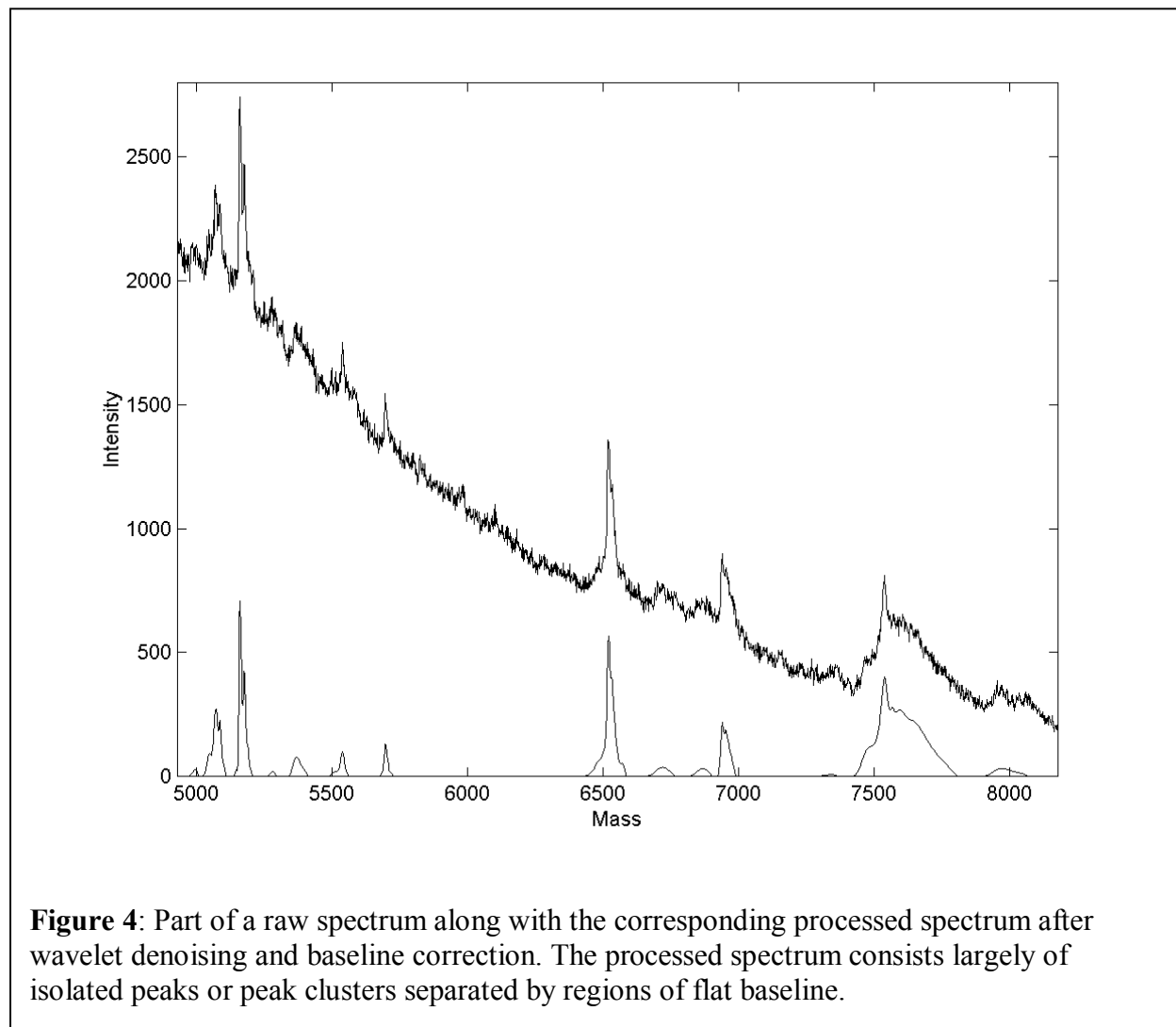


Figure 3: Heat map of the noise estimates based on a wavelet smooth with a threshold of 10 in 24 spectra (vertical axis) at more than 18,000 clock ticks (horizontal axis). The general level of the noise is nearly the same for all samples. Faint vertical banding (indicated by arrows) suggests that a few small peaks may have been consistently smoothed away at this threshold level.

3.3 Baseline correction retains peak shape and flattens non-peaks

We began by eliminating the region of the spectrum below 950 Daltons/charge. This region is typically dominated by noise from the matrix molecules, leading to extreme variability and

unreliability. Depending on the parameters used during data acquisition, this region may also contain extensive areas of saturation, where the number of ions hitting the detector exceeds its ability to count them. Saturation was frequently seen in the region below 950 Daltons in the spectra used for the present analysis, thus motivating our decision.



One side effect of eliminating the region below 950 Daltons is that the apparent baseline in these spectra then appears to be non-increasing across the entire range. Thus, we estimate the baseline by fitting a monotone local minimum curve to the denoised spectra. Denoising is critical to the success of this procedure; without it, the extremes of the noise (on the low end) will tend to drive the estimated baseline below the actual baseline, and the baseline-corrected spectra will tend to drift upward to the right [Baggerly et al., 2003]. We find that the combination of wavelet

denoising and a monotone minimum estimate of the baseline leads to processed spectra that retain the essential shapes of peaks and separates individual peaks by regions with flat baselines (Figure 4).

3.4 Our method finds more peaks than the standard CIPHERGEN algorithm

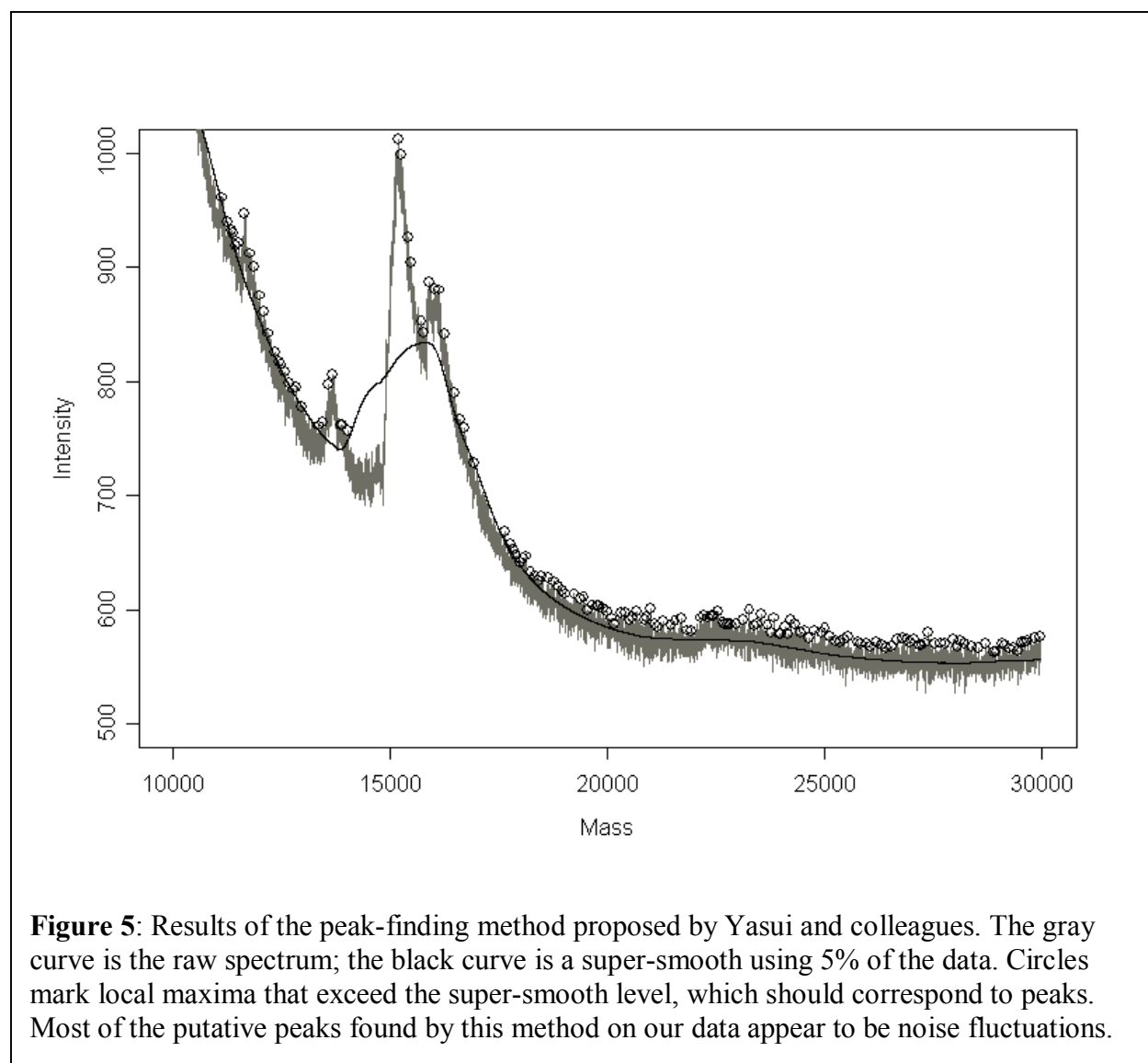
The smoothness of the processed spectra combined with the flatness of the baseline suggested that we could identify and quantify peaks by simply identifying the local maxima in a processed spectrum and recording their heights. When we used this method on our data, we found that the 24 processed spectra contained, on average, about 211 local maxima in the region above 950 Daltons/charge. We refined this list of potential peaks by considering the signal-to-noise ratio (S/N). We took the maximum height of a peak above its processed baseline to represent the signal. We also had local estimates of noise given by the residuals from the wavelet denoising process. We smoothed the local noise estimates by computing the mean of the absolute value of the noise in a window that was 500 clock ticks wide. With these definitions, each of the 24 spectra contained, on average, about 96 peaks with $S/N > 10$ and about 158 peaks with $S/N > 2$.

We compared our method for processing the spectra with the default method implemented in the CIPHERGEN ProteinChip Software [Fung and Enderwick, 2002]. The CIPHERGEN algorithm makes two passes to find peaks. It first identifies peaks in individual spectra with $S/N > 10$, and then refines those peaks across multiple spectra by adding peaks at the same mass location as long as $S/N > 2$. Using the same 24 spectra and the default settings, the CIPHERGEN algorithm found only 9 peaks per spectrum. When we increased the peak sensitivity to its maximum level (making no other changes), the CIPHERGEN algorithm found only 41 peaks per spectrum.

3.5 Comparison with the method of Yasui and colleagues

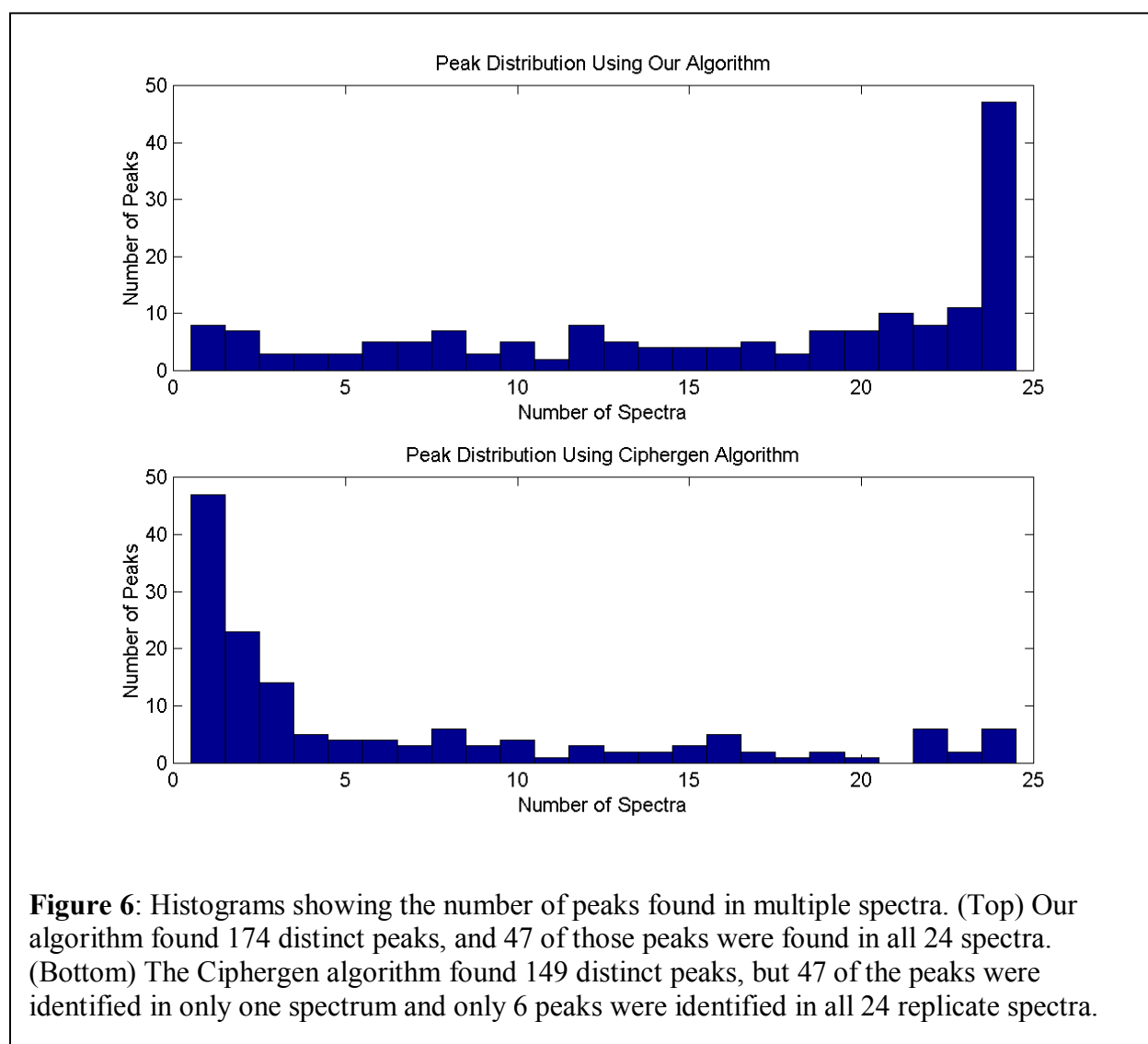
We also compared our method with the peak-finding method described by Yasui and colleagues [2003]. Their method does not attempt to quantify peaks; instead, they compute a binary indicator for the presence or absence of a peak. They define a point on the graph of the spectrum to be a peak if it satisfies two properties. First, it must be a local maximum in a fixed width

window. (They use a window that extends 20 clock ticks on either side.) Second, it must have an intensity value higher than the average intensity in a broad neighborhood, where this average is computed using the super-smoother method in a window containing 5% of the data points. When we applied this method to our data using their empirically determined parameters, it detected an average of 520 “peaks” per spectrum. Unfortunately, the vast majority of these peaks were contained in regions that looked flat to the naked eye (Figure 5). We were unable to find a combination of parameters (window width for the local maximum and window fraction for the super-smooth) that identified the obvious visible peaks without including large numbers of putative peaks that appeared to us to be spurious.



3.6 Our method finds reproducible peaks

In light of our experience with the method of Yasui and colleagues, it is clear that the number of peaks found per spectrum is not, by itself, an adequate measure of the quality of a peak finding algorithm. It is important to ascertain if the peaks being found by the algorithm correspond to real phenomena in the spectra. While we do not have knowledge of the “true” peaks in the spectra used for this investigation, we can look at the reproducibility of the method, since we have 24 spectra independently derived from the same starting material.



There is, of course, some drift in the locations of spectral peaks from one experiment to another. This drift is caused by a combination of uncertainty in the calibration (which provides the mapping from time to mass) and by the varying width of the peaks. Thus, we need to include a processing step that determines which peaks found in individual spectra should be identified as representing the same biochemical substance across spectra. To accomplish this task, we first considered only the set of peaks with $S/N > 10$. We coalesced two peaks if they differed in location by at most 7 clock ticks or if they differed in relative mass by at most 0.3%. These parameters were determined empirically by visually checking the spectra. We looked at isolated peaks, raising the threshold when a preponderance of isolated peaks could not be verified by eye in the raw spectra. We also looked across spectra in the neighborhood of peaks that occurred in many spectra to ensure that we were not collapsing visibly distinct peaks. With these parameters, there were only a few double peaks that were collapsed into single peaks, and these tended to be overlapping peaks. Next, we went back to the peaks with $2 < S/N < 10$, and added these to the list if they fell within the same distance limits (7 ticks or 0.3% of the mass) of a previously identified peak. (This second pass is equivalent to the method used in the CIPHERGEN software.) Using this method, we found a total of 174 distinct peaks across the 24 replicate spectra (Figure 6, top). Of these, 47 peaks were present in all 24 spectra. Moreover, 83 peaks were found in at least 20 spectra, 106 peaks were found in at least 15 spectra, and 130 peaks were found in at least 10 spectra. Visual inspection of the raw spectra in a neighborhood of peaks found 10 times suggest that failure to detect these peaks represents differences inherent in the SELDI technology rather than flaws in our processing methods; one example is illustrated in Figure 7.

By contrast, many of the peaks found by the CIPHERgen algorithm were located in regions of the spectra that appeared visually to be quite noisy. This observation suggested that, in addition to missing many peaks that were found by our method, their method might also detect numerous spurious peaks. If this were so, then we would expect many of the peaks they found to occur in only one of the replicate spectra. To test this idea, we combined the CIPHERgen peaks across spectra as described above, and plotted a histogram of the number of times a peak was found (Figure 6, bottom). The CIPHERgen algorithm only identified 6 peaks across all 24 spectra, and 47 of the 149 distinct peaks they identified were present in only one spectrum.

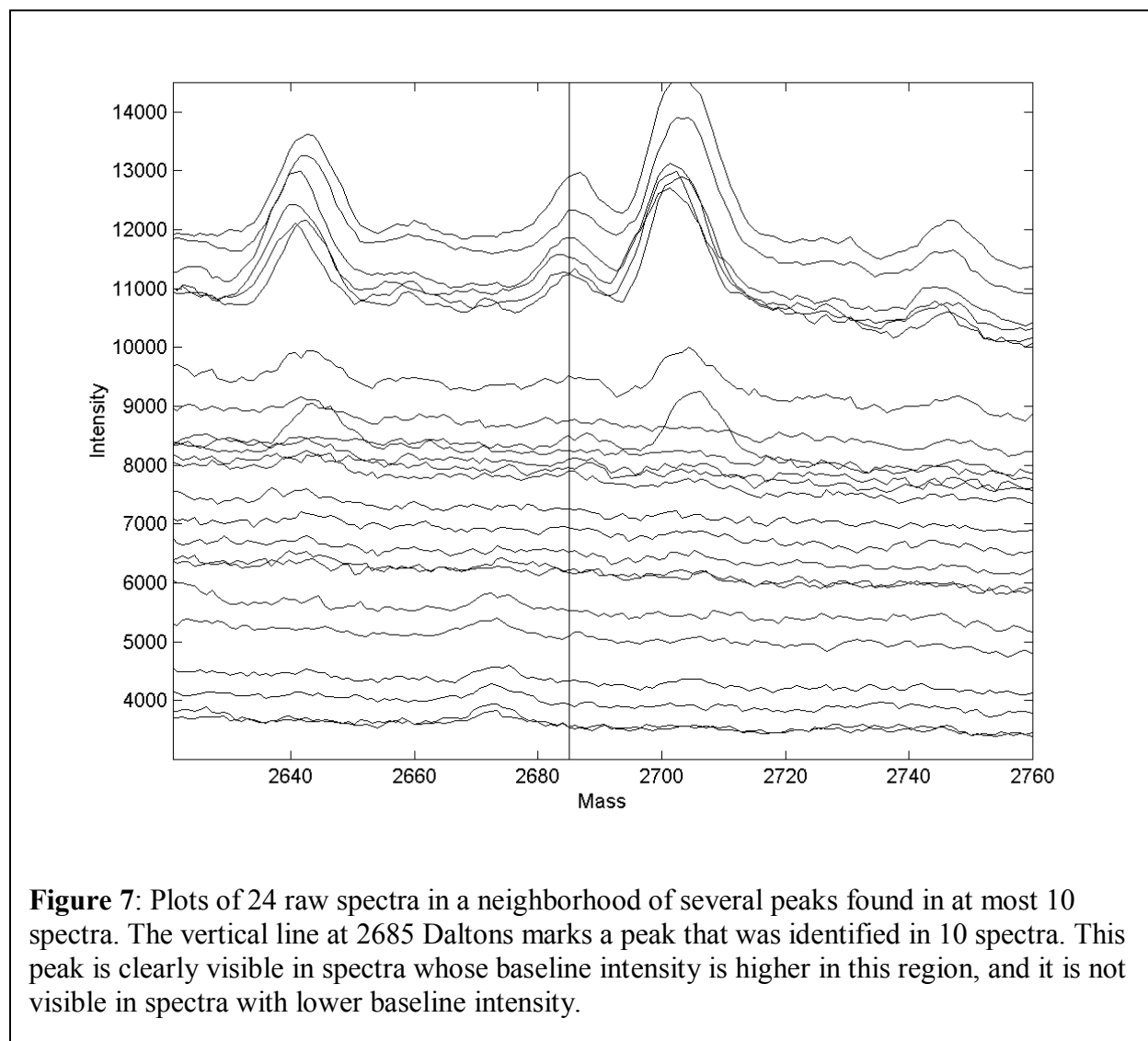


Figure 7: Plots of 24 raw spectra in a neighborhood of several peaks found in at most 10 spectra. The vertical line at 2685 Daltons marks a peak that was identified in 10 spectra. This peak is clearly visible in spectra whose baseline intensity is higher in this region, and it is not visible in spectra with lower baseline intensity.

3.7 The coefficient of variation is small with our method

In order to compare the sizes of peaks across samples, we normalized each spectrum by dividing by the total ion current in the region above 950 Daltons/charge and then multiplying by the arbitrary constant 10,000. Normalization was performed after wavelet denoising and baseline correction. For each peak set that had been found in at least three spectra, we computed the mean and standard deviation of the normalized peak heights across the spectra where the peaks were detected. As we saw previously, the standard deviation increased roughly linearly with the mean intensity. Thus, we transformed the peak heights by computing the base-two logarithm, and we computed the mean log intensity and its standard deviation. As expected, the transformation removed the linear dependence of the standard deviation on the mean. The coefficient of

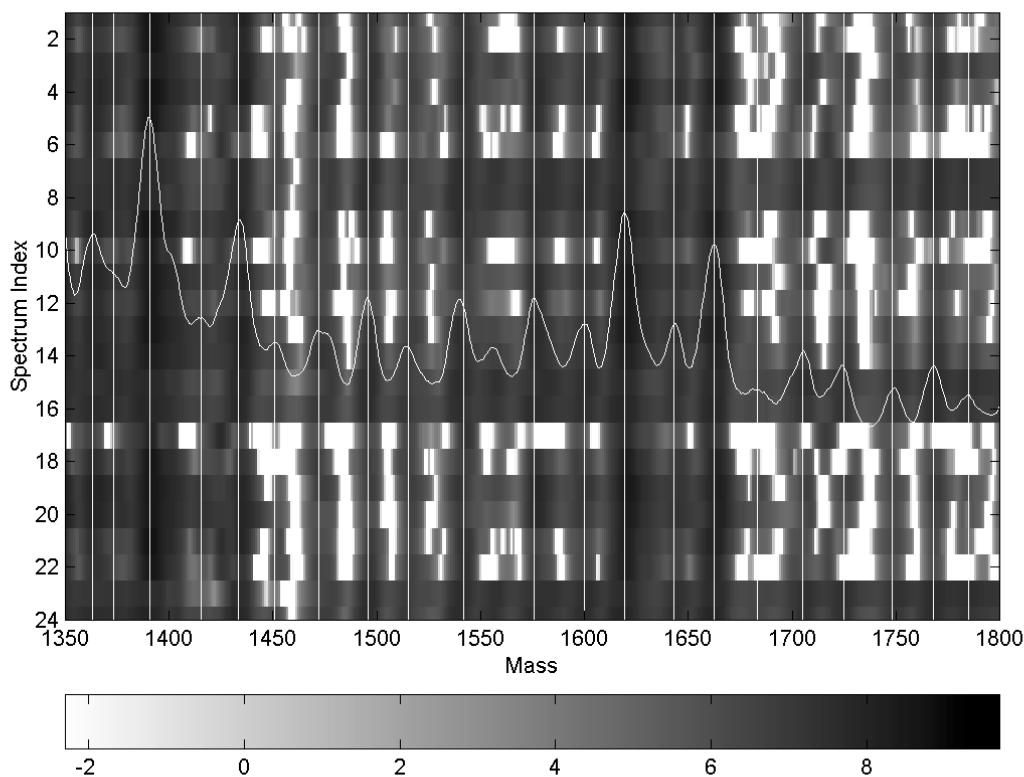


Figure 8: An artificial gel (heat map) of the log-transformed normalized intensities in 24 spectra in the mass range from 1350 to 1800 Daltons. Our algorithm detects 21 distinct peaks in this region (indicated by vertical lines), all of which appear to correspond to reproducible peaks that are visible both in the gels and in the overlaid plot of the mean of the 24 raw spectra.

variation of the log-transformed heights of the 159 peaks that were seen in at least 3 different spectra had a mean value of 10.6%, and ranged from 1.6% to 24.9%.

3.8 Our method appears to identify most of the true peaks in the spectra

We made one final visual inspection of the data to try to decide if the peaks that were detected by our method were indeed visible in most of the spectra. We plotted a heat map, or artificial gel, of the 24 spectra. On this image, we overlaid vertical lines at the points identified as peaks in at least 10 spectra, along with a plot of the mean spectrum (Figure 8). The peaks identified by our method were visible as well-aligned, darker bands in the heat map; they were also clearly visible as peaks in the mean spectrum. Moreover, our algorithm successfully identified almost all of the visible features.

4 Discussion

The difficulty in processing mass spectrometry data stems primarily from the fact that the observed signal consists of three components: the true peaks, baseline, and noise. Disentangling these three pieces is a complex task. Finding a method that successfully isolates one piece, however, can make the remaining task much simpler. The gains obtained by isolating and removing the noise component are particularly noteworthy. A denoised signal is typically much smoother than the original raw signal. As a consequence, it becomes easier to separate the low frequency baseline component from the true peaks, and it becomes extremely easy to identify peaks simply by locating all local maxima in the denoised spectrum.

Since their introduction, wavelets have been used to denoise signals in a wide variety of contexts. Recent biomedical applications have included capillary electrophoresis, magnetic resonance imaging, ultrasound blood flow, microneurography, and computed tomography [Liu et al., 2003; Bao and Zhang, 2003; Placidi et al., 2003; Olhede and Walden, 2003; Diedrich et al., 2003; Harpen, 1999]. Wavelet denoising works because most functional signals can be represented by a small number of wavelet coefficients, while white noise is distributed equally over all wavelet coefficients. Thus, if we set to zero all wavelet coefficients below a particular threshold, then we can remove the noise without biasing the signal much, since the larger wavelet coefficients that define the signal remain unaltered. The choice of threshold is governed by the usual bias-variance tradeoff that occurs in many statistical contexts. If the threshold is too large, then the signal may be altered. If the threshold is too small, then the level of denoising may be inadequate. For SELDI data, we have found that thresholds on the order of 6-10 times the MAD work well with the UDWT to isolate the noise. Other MALDI instruments may well have different noise characteristics and require different threshold settings. Appropriate thresholds can be found by plotting the noise estimates in one or two spectra at a range of threshold levels, as illustrated in Figure 1.

Qu and colleagues [2003] used an orthogonal (decimated) discrete wavelet transform (DWT) to study SELDI data derived from the serum of prostate cancer patients. They present the wavelet

transform as a tool for data reduction or feature selection, and they perform discriminant analysis on the wavelet coefficients that exceed a threshold. We have several objections to this approach. First, the wavelet features do not have a biological interpretation. By transforming back from the wavelet domain to the time domain, we can find actual peaks corresponding to physical proteins with specific mass-to-charge ratios. We can then perform additional experiments to isolate and identify those specific proteins, allowing us to develop alternative assays. Second, by working in the wavelet domain without considering the inverse transform, it is harder to justify the selection of a threshold for their method. We believe that the number of useful features that can be extracted from a spectrum is roughly equal to the number of peaks, and by using the inverse transform we can visually assess whether our thresholding is retaining the peaks while removing the noise. Finally, we believe that the UDWT performs better than the DWT.

Researchers who want to apply wavelets, whether for denoising or for data reduction, typically find themselves with an embarrassment of riches. First, there are a variety of fundamental wavelets. For our application, we tried Daubechies wavelets of different orders. We found, as expected, that the results were robust to this choice provided the order was sufficiently large. Second, each fundamental wavelet can be used to generate orthogonal bases (as in the DWT) or redundant bases (as in the UDWT). We found, when using the DWT for denoising, that it tended to create significant artifacts near the ends of the signal. These artifacts can change substantially by shifting the starting point of the signal being transformed. Similar phenomena have been described previously; with the DWT, there is usually a trade-off between the smoothness of the denoised signal (which can be improved with soft thresholding) and its squared-error performance (which can be improved with hard thresholding) [Lang et al., 1995; Lang et al., 1996; Coifman and Donoho, 1995]. The redundant bases of the UDWT provide a shift-invariant denoising method that can simultaneously produce more smoothness and better squared-error performance than the orthogonal DWT. Complex wavelets [Kingsbury, 1998] may provide similar benefits, but we have not yet explored them thoroughly in the context of mass spectrometry data.

We have presented substantial evidence in this paper that our algorithm tends to find most of the true, reproducible peaks in SELDI data. It appears to be both more sensitive and more specific

than the algorithm implemented in version 3.2 of the Ciphergen ProteinChip software. We believe that these improvements are almost entirely due to the use of the UDWT to separate the noise from the remainder of the signal; the local maxima that remain in the denoised spectra are very easy to locate and almost always correspond to actual peaks.

Our method appears to be substantially more specific than the method proposed by Yasui and colleagues [2003], with the added advantage that we go beyond a binary indicator of the presence of a peak to a quantitative measure of its size. We suspect that the lack of specificity in the Yasui algorithm arises primarily from their use of fixed-size windows across the entire spectrum. (Our earlier methods for processing mass spectrometry data used windows of varying size [Baggerly et al, 2003, Coombes et al, 2003].) By requiring a peak to be a local maximum in a window that extends 20 clock ticks on either side, their method allows approximately one potential peak every 20 clock ticks. They then fit a super-smoothed curve to the data, which in regions where the peaks are sparse should lie only a small distance above the local median. Thus, each (non-overlapping) interval of 20 clock ticks that does not contain a true peak has nearly a 50% chance of contributing a spurious peak. Our spectra were 18,963 clock ticks long. If we assume that each spectrum contains 96 true peaks (the average number our method detected with $S/N > 10$), this computation would yield 426 spurious peaks, for a total of 522 peaks. If we assume instead that there are 174 true peaks (the number of distinct peaks we found when merging data across all 24 spectra), then the corresponding computation yields 387 spurious peaks, for a total of 561 peaks. Both of these computations are consistent with our finding that the Yasui method detected, on average, 520 peaks per spectrum.

Successful denoising of mass spectrometry data, although extremely important, is still only one aspect of low-level processing. It must be combined with methods for calibration, baseline correction, and normalization. We have not addressed calibration issues in the present paper, since calibration is best handled experimentally at the time the spectral data are acquired. The baseline correction method used in this paper (fitting a monotone minimum curve) is certainly not an adequate solution to the general problem. We have seen examples of SELDI and MALDI spectra where the simple assumption of a nonincreasing baseline does not apply; further research (and experimentation) is required to better understand the behavior of the baseline. Other classes

of mass spectrometry instruments, such as quadrapoles or electrospray devices, are also likely to generate data with very different characteristics, and may require alternative processing methods. Normalization also poses some open research problems. For instance, it is not at all clear that the assumption of equal concentrations of protein (which underlies the idea of normalizing to the total ion current) has a sound biological basis when comparing two or more classes of samples. All of these questions, however, can be addressed more easily if the noise is first removed using the undecimated discrete wavelet transform.

5 References

Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL Jr. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609-14.

Adam PJ, Boyd R, Tyson KL, Fletcher GC, Stamps A, Hudson L, Poyser HR, Redpath N, Griffiths M, Steers G, Harris AL, Patel S, Berry J, Loader JA, Townsend RR, Daviet L, Legrain P, Parekh R, Terrett JA. (2003) Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. *J. Biol. Chem.* **278**, 6482-9.

Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC, Coombes KR. (2003) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3**, 1667-72.

Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: Comparing data sets from different experiments. *Bioinformatics*, to appear.

Bao P, Zhang L. (2003) Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. *IEEE Trans. Med. Imaging* **22**, 1089-99.

Chao FT, Leung AKM. (2000) Application of wavelet transform in processing chromatographic data. In: Walczak B (ed.) *Wavelets in Chemistry*, Elsevier Science Publishers, pp. 205-223.

Coifman R, Donoho D. (1995) Translation invariant de-noising. In *Wavelets and Statistics*, New York: Springer-Verlag, pp. 125-150.

Coombes KR, Fritsche HA Jr, Clarke C, Chen JN, Baggerly KA, Morris JS, Xiao LC, Hung MC, Kuerer HM. (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.* **49**, 1615-1623.

Diedrich A, Charoensuk W, Brychta RJ, Ertl AC, Shiavi R. (2003) Analysis of raw microneurographic recordings based on wavelet de-noising technique and classification algorithm: wavelet analysis in microneurography. *IEEE Trans. Biomed. Eng.* **50**, 41-50.

Fung ET, Enderwick C. (2002) ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques Suppl.*, 34-41.

Harpen MD. (1999) A computer simulation of wavelet noise reduction in computed tomography. *Med. Phys.* **26**, 1600-6.

Hutchens TW, Yip TT. (1993) New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Commun. Mass Spectrom.* **7**, 576-580.

Kamath C, Fodor IK, Gyaourova A. (2002 November) Undecimated wavelet transforms for image denoising, Lawrence Livermore National Laboratory technical report UCRL-ID-150931.

Kingsbury N. (1998) The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters. In: *Proc. 8th IEEE DSP Workshop*, Bryce Canyon UT, USA, paper no. 86.

Lang M, Guo H, Odegard JE, Burrus CS, Wells RO Jr. (1995) Nonlinear processing of a shift invariant DWT for noise reduction. In: *Mathematical Imaging: Wavelet Applications for Dual Use*, SPIE Proceedings, vol. 2491, Orlando FL.

Lang M, Guo H, Odegard JE, Burrus CS, Wells RO Jr. (1996) Noise Reduction Using an Undecimated Discrete Wavelet Transform. *IEEE Signal Processing Letters* **3**, 10-12.

Liu BF, Sera Y, Matsubara N, Otsuka K, Terabe S. (2003) Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis. *Electrophoresis*; **24**, 3260-5.

Olhede SC, Walden AT. (2003) Noise reduction in directional signals using multiple Morse wavelets illustrated on quadrature Doppler ultrasound. *IEEE Trans. Biomed. Eng.* **50**, 51-7.

Paweletz CP, Gillespie JW, Ornstein DK, Simone NL, Brown MR, Cole KA, Wang QH, Huang J, Hu N, Yip TT, Rich WE, Kohn EC, Linehan WM, Weber T, Taylor P, Emmert-Buck MR, Liotta LA, and Petricoin EF. (2000) Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Development Research* **49**, 34-42.

Paweletz CP, Trock B, Pennanen M, Tsangaris T, Magnant C, Liotta LA, Petricoin EF. (2001) Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis. Markers* **17**, 301-7.

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572-7.

Placidi G, Alecci M, Sotgiu A. (2003) Post-processing noise removal algorithm for magnetic resonance imaging based on edge detection and wavelet analysis. *Phys. Med. Biol.* **48**, 1987-95.

Qu Y, Adam BL, Thornquist M, Potter JD, Thompson ML, Yasui Y, Davis J, Schellhammer PF, Cazares L, Clements M, Wright GL Jr, Feng Z. (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics* **59**, 143-51.

Rogers MA, Clarke P, Noble J, Munro NP, Paul A, Selby PJ, Banks RE. (2003) Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and

neural-network analysis: identification of key issues affecting potential clinical utility. *Cancer Res.* **63**, 6971-83.

Sauter ER, Zhu W, Fan XJ, Wassell RP, Chervoneva I, Du Bois GC. (2002) Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. *Br. J. Cancer* **86**, 1440-3.

Schaub S, Wilkins J, Weiler T, Sangster K, Rush D, Nickerson P. (2004) Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int.* **65**, 323-32.

Sorace JM, Zhan M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **4**, 24.

Wellmann A, Wollscheid V, Lu H, Ma ZL, Albers P, Schutze K, Rohde V, Behrens P, Dreschers S, Ko Y, Wernert N. (2002) Analysis of microdissected prostate tissue with ProteinChip arrays--a way to new insights into carcinogenesis and to diagnostic tools. *Int. J. Mol. Med.* **9**, 341-7.

Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL Jr, Qu Y, Potter JD, Winget M, Thornquist M, Feng Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4**, 449-63.

Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS. (2003) Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer* **40**, 267-79.