

Feature Extraction Methods for Analyzing Mass Spectrometry Data in Biomedical Applications Using the Mean Spectrum

Jeffrey S. Morris^{a 1 2}, Kevin R. Coombes^{a2}, John Koomen^b, Keith A. Baggerly^a, and Ryuji Kobayashi^b

*Departments of ^aBiostatistics and Applied Mathematics and ^bMolecular Pathology
The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA.*

Summary: Mass spectrometry methods are actively being used to discover disease-related proteomic patterns in biological samples. These methods yield highly structured functional data characterized by many peaks that correspond to individual proteins present in the sample. One approach in analyzing these data is to first perform peak detection and quantification, then use the quantified peak intensities in further analyses. We refer to this approach as feature extraction. If feature extraction is used, it is crucial to optimize the peak detection procedure, since subsequent analyses condition on these determinations. In this paper, we introduce a new peak detection algorithm that uses wavelet transforms and operates on the mean spectrum. This method is relatively easy to implement, and it avoids the peak matching problem inherent when performing peak detection on individual spectra. Also, peaks tend to be enhanced in the average spectrum, since by the Central Limit Theorem the noise level is reduced while features that are part of the signal tend to be reinforced. We describe our method, and demonstrate the benefits of performing peak detection on the average spectrum through real examples and simulation studies. Our results suggest that working with the average spectrum increases sensitivity and specificity of peak detection. Our simulation studies use a virtual mass spectrometer we have developed which models the actual physical process undergirding the instrument, and appears to yield reasonably realistic-looking mass spectrometry data. The simulation studies represent an important contribution in themselves, since they describe how to conduct such studies to compare methods for analyzing mass spectrometry data.

Keywords: Central Limit Theorem, Denoising, Functional data analysis, Mass spectrometry, MALDI-TOF, Peak detection, Proteomics, Wavelets.

Short title: Peak Detection for Proteomics Data

¹*Address for correspondence:* Jeffrey S. Morris, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Unit 447, Houston, TX 77030-4009, USA. Email: jefmorris@mdanderson.org

²Contributed equally to this paper

1 Introduction

Mass spectrometry is actively being used to discover 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; 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.

The mass spectrometry instruments most commonly applied to clinical and biological problems use a matrix-assisted laser desorption and ionization (MALDI) ion source and a time-of-flight (TOF) detection system. Briefly, to run an experiment on a MALDI-TOF instrument, the biological sample is first mixed with an energy absorbing matrix (EAM) such as α -cyano-4-hydroxycinnamic acid, which causes the mixture to crystallize as it dries. The metal plate containing the crystallized sample is placed into a vacuum chamber. The crystal is then struck with light pulses from a nitrogen laser. The matrix molecules absorb energy from the laser and transfer it to the proteins, causing them to desorb and ionize, producing a cloud of ionized protein molecules in the gas phase. Next, an electric field accelerates the ionized proteins into a flight tube where they drift until they strike a detector that records the time of flight. Knowing the length of the tube and the applied voltage, researchers can use a quadratic transformation to derive the approximate mass-to-charge ratio (m/z) of the protein from the observed time of flight. The spectral data that result 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 represent proteins that are present in the sample.

A typical data set arising in a clinical application of MALDI-TOF contains tens or hundreds of spectra; each spectrum contains tens of thousands of intensity measurements representing an unknown number of protein peaks. We see two general approaches one could take in analyzing these data. The first is a functional data analysis approach (FDA)-(Ramsay and Silverman, 1997), whereby the entire set of spectra are modeled as functions. Morris and Carroll (2004) and Billheimer (2004) take this approach. The second approach is feature extraction, which consists of two steps. The first step is to preprocess the spectra and extract the meaningful features, then the second is to perform analyses on these extracted features. Each approach has advantages and disadvantages, but for this paper we focus on feature extraction. Feature extraction leads

to a great degree of data reduction, summarizing tens of thousands of intensity measurements with only hundreds or thousands of features. As long as the scientifically meaningful features of the spectra are properly extracted, there should be very little loss of useful information by using this approach.

Qu, et al. (2003) used feature extraction to analyze mass spectrometry data. The features they extracted were wavelet coefficients. This approach is effective for dimension reduction, since wavelets can yield a parsimonious representation for many classes of functions, including functions comprised of convolutions of peaks. In their analysis, they were able to represent 48,538 intensity measurements with just 1,271 wavelet coefficients. However, this approach does not take advantage of the knowledge that the scientifically meaningful features of the spectra are the peaks, which correspond to proteins that are present in the sample. It is not straightforward to map wavelet coefficients back to peaks, and indeed there is not a one-to-one correspondence between them. Thus, we focus directly on the peaks in our feature extraction. We have found that this tends to yield even greater dimension reduction than the wavelets, and the results are more readily interpretable since if a peak is found to be interesting, the biologists can try to find the identity of the corresponding protein from the molecular mass of the peak.

Feature extraction requires extensive low-level processing in order to identify the locations of peaks and to quantify their sizes accurately. After properly preprocessing a set of n spectra and quantifying p peaks per spectrum, one should be left with a $p \times n$ matrix of “peak expression levels”. This matrix is analogous to the matrix of gene expression levels produced by a microarray experiment. At this point, a MALDI-TOF data set can be analyzed using tools that have already been developed for microarrays. Since subsequent analyses condition on the preprocessed data, it is very important that the best possible preprocessing methods are used. Inadequate or incorrect preprocessing methods 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., 2003; Baggerly et al., 2004).

The low-level processing of mass spectra involves a number of complicated steps that interact in complex ways. Some elements of the preprocessing can be elucidated by the following model, which represents a decomposition of the observed signal into three components: baseline, true signal, and noise. Suppose we observe n spectra, each taken on the same equally-spaced grid of length T of times of flight $t_j, j = 1, \dots, T$. A model for $y_i(t_j)$, the observed spectral intensity

for spectrum i at time of flight t_j , is

$$y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + \epsilon_{ij}. \quad (1)$$

The true signal, $S_i(t)$, consists of a sum of possibly overlapping peaks, each corresponding to a particular biological molecule, e.g., a protein or peptide. The approximate shapes of peaks can be estimated empirically by simulating the physical process by which time-of-flight (TOF) mass spectrometers collect data (see Coombes, et al. 2004a), although here we do not attempt to parametrically characterize the shapes of the peaks. The normalization factor, N_i , is a constant multiplicative factor to adjust for possibly differing amounts of protein on each slide. The baseline function, B_i , represents a systematic artifact commonly seen in mass spectrometry data. This artifact is largely attributable to a cloud of matrix molecules hitting the detector in the early part of the experiment, as well as detector overload. Our only characterization of this function is that it should be smooth. In this paper, we assume that the errors are white noise, i.e. $\epsilon_{ij} \sim N(0, \sigma^2)$. We believe that the white noise assumption is plausible here since the additive noise is primarily electronic noise from the detector, although it is also reasonable to allow the noise variance to be some smooth function of the time of flight, $\sigma^2(t_j)$.

Following is a list of the steps comprising the low-level processing of MALDI-MS data:

1. **Calibration** maps the observed times of flight $t_j, j = 1, \dots, T$ to a set of inferred mass-to-charge ratios $x_j, j = 1, \dots, T$. This step aligns multiple spectra and yields molecular masses that can be used to ascertain the protein identity of a peak of interest.
2. **Filtering** removes the random noise, ϵ_{ij} , typically electrical or chemical in origin.
3. **Baseline subtraction** removes the baseline artifact $B_i(t)$.
4. **Normalization** corrects for systematic differences in the total amount of protein desorbed from the sample plate, represented by N_i in model (1).
5. **Peak detection and quantification** is the primary goal of low-level processing. It involves identifying the locations of peaks in the true signal, $S_i(t)$, and then quantifying the intensity of each peak for each spectrum, which is a rough surrogate for the amount of the corresponding protein desorbed from the sample.
6. **Peak matching** across samples is required because neither calibration nor peak detection is perfect. Thus, the analyst must decide which peaks in different samples correspond to the same biological molecule.

Calibration is best performed experimentally, using a sample containing a small number of proteins of known mass. Throughout this paper, we will assume that all spectra have been experimentally calibrated and, if necessary, interpolated so that they can reasonably be compared on both a common mass axis and a common time axis. In future research, we hope to address the problem of validating or improving the calibration based on evidence internal to the spectra.

We recently described a method for filtering and denoising spectra based on the undecimated discrete wavelet transform (UDWT) (Coombes et al., 2004b). We showed that the UDWT can be effectively used to isolate the noise component of a spectrum. As a result, it is easier to perform baseline subtraction, normalization, and reproducible peak detection and quantification on the resulting denoised spectrum. In that work, we also showed that our peak detection method based on this approach appears to outperform some other peak detection methods commonly used for mass spectrometry data (Fung and Enderwick, 2002, Yasui, et al., 2003a, 2003b).

Low-level processing using the UDWT, like other standard approaches, performs the first five processing steps separately on each individual spectrum. These approaches relegate the problem of peak matching to the final preprocessing step. As a result, peak matching becomes the first point in the analysis that acknowledges the existence of more than one spectrum. While we recognize that such an approach has advantages in a high-throughput setting, we also believe that it has some serious drawbacks. Peak matching is fraught with difficulties, and it frequently requires the analyst to make numerous *ad hoc* decisions. How many clock ticks must separate peaks before you call them different? What relative mass difference is large enough to call two peaks different? What signal-to-noise (S/N) ratio is large enough to believe that a computationally detected peak arises from a real biological molecule? If a peak appears in ten spectra with $S/N = 3$, is it as reliable as a peak that appears in only one spectrum with $S/N = 30$?

We propose to finesse the peak matching problem by performing peak detection on the mean spectrum; we return to individual spectra only to quantify peaks that have already been located using the mean spectrum. We observed previously that the mean spectrum can provide important insights into structure in proteomics data. For example, we used the mean spectrum to uncover the presence of systematic noise in a MALDI-TOF lung cancer data set (Baggerly et al., 2003). This systematic noise, apparently attributable to a computer buffer on the detector, was present in all spectra in the study, but was obscured by additional random noise in individual samples.

The idea of using the mean spectrum to get around the peak matching problem has been suggested by others (Carpenter et al., 2003). In their paper, however, Carpenter and colleagues only mentioned the idea tangentially and focused more attention on “binning” methods for data reduction. In the present paper, we take an in-depth look at the benefits to be gained by working with the mean spectrum. The Central Limit Theorem suggests that at any given point, the noise level should decrease in the mean spectrum by the square root of the number of samples. Thus, we hypothesize that using the mean spectrum for peak detection should yield increased sensitivity and specificity for peak detection. We investigate this possibility both by applying our algorithm to real data and by conducting a detailed simulation based on a physical model of a MALDI-TOF instrument. Specifically, we compare the performance of our previously described peak detection method based on the UDWT (Coombes, et al., 2004b) applied to individual spectra with a revised version of the method that operates on the mean spectrum.

2 Peak Detection Methods

We now describe the two peak detection methods we consider in this paper. All MATLAB scripts for both methods are available from our web site (<http://bioinformatics.mdanderson.org/software.html>).

2.1 Peak Detection Using UDWT on Individual Spectra (SUDWT)

Following are the steps used to preprocess the spectra by applying the UDWT-based method introduced in Coombes, et al. (2004b) to the individual spectra. We will refer to this method as the SUDWT, or single spectrum, undecimated wavelet transform-based peak detection method.

1. Ensure that the individual spectra are well calibrated.
2. Denoise the individual spectra using the undecimated discrete wavelet transform, as implemented in version 2.4 of the Rice Wavelet Toolbox (RWT), which is available from the web site of the Digital Signal Processing Group at Rice University, Houston, Texas (<http://www-dsp.rice.edu/software/rwt.shtml>).
 - The denoising works by computing the wavelet coefficients for the observed signal, then performing hard thresholding. In hard thresholding, all coefficients less than a

threshold value are set to zero, while all coefficients greater than the threshold remain unchanged. The threshold is the product of a thresholding parameter η and a robust estimate of the noise, the median absolute deviation (MAD) divided by 0.67.

- Unlike the more commonly used decimated discrete wavelet transform, the undecimated discrete wavelet transform is translation-invariant, which makes it more effective for denoising.
 - We have found that the choice of wavelet basis does not strongly impact the denoising, although the choice of the thresholding parameter η does.
 - This method decomposes the observed signal into two components: the denoised signal and the noise.
3. Estimate the noise level across the spectrum using a median filter, i.e., by applying the MAD/0.67 estimate to the noise in a sliding window.
 4. Estimate and remove the baseline artifact by computing a monotone local minimum curve on the denoised signal.
 5. Normalize the spectrum by dividing by the total ion current, defined to be the mean intensity of the denoised and baseline corrected spectrum.
 6. Identify peaks on the denoised, baseline corrected, and normalized spectrum.
 - (a) Find all local maxima and the associated peak endpoints.
 - (b) Compute the signal-to-noise ratio (S/N) at each local maximum by taking the ratio of the intensity at the maximum to the local noise estimate.
 - (c) Let ϕ be a S/N threshold. All local maxima with S/N above ϕ are considered peaks.
 7. Match peaks across spectra. First pool the list of detected peaks across spectra, then combine peaks that differ in location by no more than δ_t clock ticks or δ_m in relative mass.
 8. Quantify peaks using either the intensity of the local maximum or computing the area under the curve (AUC) for the region defined to be the peak.

2.2 Peak Detection Using the UDWT on Mean Spectrum (MUDWT)

We now describe an adaptation of this algorithm that uses the average spectrum for peak detection. We refer to this method as MUDWT, or the mean-spectrum, undecimated discrete

wavelet transform-based peak detection method.

1. Ensure that the individual spectra are well calibrated.
2. If necessary, use interpolation to put all spectra on the same time scale.
3. Compute the mean of all (interpolated) raw spectra.
4. Apply the UDWT-based method described above to denoise, baseline correct, and find peaks in the mean spectrum. This method finds all local maxima in the average spectrum, and identifies an interval containing each peak. Keep only those peaks above the S/N threshold ϕ . The noise reduction inherent in the averaging should allow the use of a smaller S/N threshold.
5. Quantify the identified peaks in the individual spectra.
 - (a) Apply the UDWT to denoise and baseline correct each individual spectrum. We generally recommend choosing a smaller η for quantification than when doing peak detection to reduce bias in quantifying the peak.
 - (b) Normalize an individual spectrum by dividing by the total ion current.
 - (c) Peaks can be quantified either by height (maximum) or area (sum) within the specified interval.

Note that the algorithm is applied to the mean spectrum based on the original raw spectra with no processing other than calibration. This may seem surprising at first, but there is a good reason why this works. A peak is something that stands out above the noise and above the baseline, ideally in multiple spectra. These properties should be preserved (and, with respect to the noise, enhanced) in the mean spectrum. This can be seen in the model for the sample mean spectrum, $\bar{y}(t_j)$, which is

$$\bar{y}(t_j) = \bar{B}(t_j) + \overline{NS}(t_j) + \bar{\epsilon}_j, \quad (2)$$

where $\bar{y}(t_j) = n^{-1} \sum y_i(t_j)$, $\bar{B}(t_j) = n^{-1} \sum B_i(t_j)$, $\overline{NS}(t_j) = n^{-1} \sum N_i S_i(t_j)$, and $\bar{\epsilon}_j = n^{-1} \sum \epsilon_{ij}$, with all sums taken over the n samples and $\bar{\epsilon}_j \sim N(0, \sigma^2/n)$. Since the baseline is smooth, the peaks will stand out above the baseline component in the mean spectrum, and will in many cases be more prominent because of the noise reduction caused by the averaging. We also realize that there are advantages in not having to rely on other processing algorithms before computing averages.

[Insert Figure 1 here]

The success of the proposed method depends on having the spectra reasonably well calibrated at the beginning. This property can be assessed visually by preparing a “heat map” of the raw spectra (Figure 1). In this figure, the vertical axis is an arbitrary ordering of the samples, the horizontal axis represents time, and the values displayed are the base-2 logarithms of the intensities. The largest peaks are easy to see in these plots, and it is easy to check that they are properly aligned across spectra. Minor inaccuracies in calibration should not cause a problem: they result in peaks in the mean spectrum that are somewhat broader than the peaks found in individual spectra. This naturally spreads out the interval where we look for peaks, allowing the data to guide us to an efficient solution to the peak matching problem.

It may be necessary to perform some normalization before averaging the spectra. Gross disparities between spectra in the overall signal level will allow some spectra to dominate others, which is evident by the fact that the mean signal $\overline{NS}(t_j)$ in (2) is a weighted average of the individual signals, with the weights being the normalization factors N_i . This will also reduce the amount of noise reduction achieved by using the mean spectrum. However, in our experience we have found that spectra that are part of the same project are typically measured on similar scales, so we have not found it necessary to perform normalization prior to computing the mean spectrum for the data we have encountered to date.

3 Examples Demonstrating the Benefit of Using the Mean Spectrum

To illustrate some of the advantages gained by using the mean spectrum, we applied our algorithm to several real data sets.

3.1 The Noise in the Mean Spectrum Decreases by \sqrt{n}

The first data set was described in our previous paper (Coombes et al., 2004b). It consists of 24 spectra acquired from the same pooled sample of nipple aspirate fluid using a variant of MALDI-TOF known as surface enhanced laser desorption and ionization (SELDI). In practice, SELDI differs from MALDI-TOF in two ways. First, it replaces a bare metal sample plate with a ProteinChip array (CIPHERGEN, Inc., Fremont, CA) containing eight chemically prepared spots. Different surface preparations preferentially bind different classes of proteins. Second,

SELDI experiments are usually performed on a relatively low resolution MALDI-type instrument manufactured by Ciphergen.

[Insert Figure 2 here]

Figure 2 shows a portion of one individual spectrum and the corresponding portion of the mean over the 24 spectra. As expected, the scale of the noise is decreased by about a factor of 5. This claim is supported on a global scale by a plot of the noise removed by applying the UDWT to both an individual spectrum and to the mean spectrum (Figure 3).

[Insert Figure 3 here]

3.2 Peak Finding on the Mean Spectrum Appears to be More Sensitive

We compared the peaks found on the mean spectrum with the peaks that were first found in individual spectra and then matched across spectra. In the analysis of this data set reported in our previous paper using the SUDWT method, we found 174 sets of matched peaks (Coombes et al., 2004b). When we used the mean spectrum (MUDWT), we found 227 peaks. The differences between the two collections of peaks are the following:

1. Five of the matched-individual peaks have no corresponding peak in the mean spectrum.
2. Nineteen pairs of matched-individual peaks are collapsed into a single peak in the mean spectrum.
3. Four matched-individual peaks are resolved as double peaks in the mean spectrum.
4. The mean spectrum contains 73 peaks that were not found as individual matched peaks.

Figure 4 contains typical examples of each class of differences between the two methods.

[Insert Figure 4 here]

For the SUDWT method, we combined peaks together if they were within $\delta_t = 7$ clock ticks or $\delta_m = 0.003$ on the relative mass scale, and we used a S/N threshold of $\phi = 10$ to define the peaks. We investigated how this S/N threshold affected the numbers of peaks found (Table 1). Interestingly, the number of peaks does not increase monotonically as the signal-to-noise threshold is decreased. This phenomenon is a direct result of the peak matching problem. As

we lower the threshold, we find more peaks in each individual spectrum. As the number of low S/N peaks increases, there is a greater chance that a spurious peak will occur in regions where, because of the matching rules, two separate peaks coalesce into one.

[Insert Table 1 here]

To test this idea, we repeated the comparison between the peaks found in the mean spectrum with matched peaks from individual spectra at the threshold ($S/N > 4$) with maximum sensitivity. In this case, the four categories of peaks described above contain 17 peaks found individually but not in the mean, 25 pairs of collapsed peaks, 6 doubled peaks, and 36 peaks in the mean spectrum that are never found in an individual spectrum.

The examples shown in Figure 4 are representative of the differences we have seen between the two methods. In most cases, visual confirmation leads us to believe that using the mean spectrum provides a list of peaks that is closer to the truth.

3.3 Small but Consistent Peaks are Easily Seen in the Mean Spectrum

If we see a small bump at the same location in many spectra, our intuition suggests that it corresponds to a real protein peak. If a small bump occurs extremely rarely, however, then we think it is likely a spurious feature. By contrast, a large bump that occurs in even one spectrum is also believable. In our previous attempts to identify peaks in individual spectra and match them across spectra, we adopted *ad hoc* filtering rules along these lines, combining the number of times a peak was found with its S/N ratio. Working with the average spectrum automatically takes this idea into consideration, allowing us to effectively borrow strength across spectra.

[Insert Figure 5 here]

To illustrate this idea, we considered a publicly available data set (<http://www.ncifdaproteomics.com>). This data set contains spectra from serum samples from 216 women, 95 of whom were healthy and 121 with ovarian cancer. The data were collected using a Qstar mass spectrometer, which combines a quadrupole ion source with a time-of-flight ion detector. The basic structure of the data is similar to that produced by a MALDI-TOF instrument. The authors of the initial study of this data found a peak near 8602 daltons that appeared to be more abundant in ovarian cancer patients than in healthy women. They also pointed out that this peak never achieved a signal-to-noise ratio of greater than about 1.5, which meant that it would be filtered out by most

processing based on peak finding. The peak clearly stands out in the mean spectrum (Figure 5). Given our earlier observation about the noise levels, we would expect S/N to approximately equal $1.5 * \sqrt{100} \approx 15$ in the mean of either group of samples, making it easy to find.

3.4 The Mean Spectrum Can Find Peaks Only Present in a Few Samples

Some may argue that a possible concern with using the mean spectrum for peak finding is that proteins present in a small subset of spectra may not be detected. Biologically, such proteins may be important, especially if they are only present in a small number of cancer samples (Coombes, Wang, and Baggerly, 2004). We believe that peaks that are present at a moderately high intensity will still be detected.

To support this claim, we computed a number of statistical summaries for a set of pancreatic cancer spectra from an experiment conducted at M D Anderson Cancer Center. The data set contained MALDI-TOF spectra from the blood sera of 124 individuals, 83 with pancreatic cancer and 41 without. Figure 6 contains plots of the pointwise mean, maximum, minimum, and 90th and 10th percentiles of these spectra. There are large peaks in the maximum spectrum at 11,500 and 11,600 daltons. These peaks are barely discernible in the 90th percentile spectrum, but would be clearly detected on the mean spectrum.

[Insert Figure 6 here]

4 Simulation Studies

The previous section contained anecdotal evidence of the advantages of the algorithm based on the mean spectrum. For a more systematic and comprehensive evaluation of the method's performance, we conducted simulation studies where the number and location of true peaks were known. These simulation studies make use of a virtual MALDI-TOF mass spectrometer (Coombes, et al., 2004a), which simulates spectra using a physical model of a linear MALDI-TOF instrument with ion focus delay.

4.1 A Virtual MALDI-TOF Mass Spectrometer

Mass spectrometry data are highly structured functional data. The spectra contain many peaks that correspond to proteins present in the biological sample. There is a systematic relationship between the masses and shapes of the peaks – proteins at low masses give tall, sharp peaks while

proteins at higher masses result in shorter, broader peaks. The actual shapes of the peaks are affected by numerous factors, including the isotope distributions of the elements of the proteins, the distributions of the initial velocities of the proteins' ions as they are desorbed from the sample plate, and the time resolution of the instrument's ion detector. In order to simulate mass spectrometry data that are as realistic as possible, we used a virtual MALDI-TOF mass spectrometer (Coombes, et al., 2004a) that is based upon the instrument's physical principles to generate the virtual spectra for our simulation study.

In principle, the virtual MALDI-TOF instrument works just like a real instrument; a *virtual sample* is fed into the instrument, and out comes a *virtual spectrum*. The virtual sample consists of a list of the molecular masses and abundances of a set of proteins assumed to be present in the biological sample. The abundance of each protein is measured in numbers of ions (positively charged protein molecules). To be precise, the abundance does not measure the number of molecules in the sample, but the number of molecules successfully ionized and desorbed from the sample. The virtual instrument simulates the actual physical process the ions undergo as they are desorbed from the plate, focused from the sample plate to a first grid, accelerated through an electric field produced by two charged grids, and then allowed to drift through a field-free tube from the second grid to the detector. The actual times-of-flight for each ion are computed using basic physics principles, then aggregated to form the virtual spectrum. Virtual calibration samples consisting of proteins of known masses are then run to obtain a mapping of time-of-flight to molecular mass. These calibration samples can also be used to map abundances to expected peak intensities. Empirical investigations have revealed that for a fixed protein abundance, the inverse of the log peak intensity is linearly related to the mass (Coombes, et al. 2004a).

The virtual instrument has various dials and settings, many of which correspond to the actual physical characteristics of a MALDI-TOF instrument, including the distance from the sample plate to the first grid D_1 , the distance D_2 and voltage drop V_1 between the second grid, the length of the flight tube L , the delay time until the electric field is produced δ , and the time resolution of the detector τ . These parameters can be set to mimic a particular instrument of interest. When struck with the laser, there is a stochastic distribution of initial velocities at which the ions are desorbed from the sample plate (Gluckmann and Karas, 1999; Karas, et al., 2003), which we model with a Gaussian distribution with mean μ and variance τ^2 , following Beavis and Chait (1991). The isotopic distributions of the organic elements comprising the proteins are

modeled through Bernoulli distributions. Isotopic prevalences are well known, so the parameters of the Bernoulli distributions are well-informed from the existing literature. Finally, we add an exponential baseline curve and white noise to represent the detector noise factor.

[Insert Figure 7 here]

While the results from any virtual instrument are based on simplifying assumptions, we believe that this tool generates virtual spectra that look much like the mass spectra emanating from a real MALDI-TOF instrument. Figure 7 contains a spectrum from a real MALDI-TOF instrument at M D Anderson Cancer Center, along with a virtual spectrum obtained from our tool with matching instrument settings. While not perfect, we see that the tool seems to do a reasonable job simulating realistic spectra.

4.2 Virtual Experiments

A MALDI-TOF experiment consists of taking n samples from a biological medium of interest (e.g., blood serum), spotting them on a plate, then running them through a mass spectrometer. Ideally, these n samples represent a random sample from a biological population of interest on whose proteome we wish to make inference.

In order to run a *virtual MALDI-TOF experiment*, we need to first characterize the *virtual population* from which our virtual samples will be drawn. This virtual population consists of the list of all detectable proteins present in the medium of interest for at least one sample in the reference population, along with the abundance distributions for each protein across samples. Let p represent the total number of detectable proteins present in the population. For a given protein j of mass x_j , we summarize its distribution across samples by three quantities: π_j , the protein's prevalence, or the proportion of samples in the population containing this protein; and m_j and s_j , the mean and standard deviation log peak intensity of the protein peak across samples in the population which contain the protein.

A virtual MALDI-TOF experiment is then conducted by randomly generating n virtual samples from the virtual population, for example assuming Bernoulli distributions for the prevalences and Gaussians for the log intensities, then running these samples through the virtual mass spectrometer to obtain n virtual spectra.

4.3 Setting up the Simulation

In order to maximize the realism in our simulation study, we based our virtual populations on a real MALDI-TOF data set. We used the pancreatic data set introduced in Section 3.4, which contained 124 spectra from blood sera of subjects with and without pancreatic cancer. We applied the SUDWT method to these data using wavelet threshold $\eta = 10$, signal-to-noise threshold $\phi = 20$, and time and mass peak matching tolerances of $\delta_t = 7$ and $\delta_x = 0.002$, respectively, and detected $p = 76$ peaks over all the samples. For each peak, we estimated the prevalence π and mean and standard deviation log intensity (m and s) across the 124 samples.

Because the efficacy of a peak detection algorithm depends on many interrelated factors, including the locations and spacings of the protein peaks, we did not want the entire simulation study to depend on this single set of true protein peak masses. Thus, we generated 100 virtual populations, each with characteristics similar, but not identical, to the pancreatic cancer data set. In this way, we were able to average over variability in these factors in our assessments of the methods' performances.

We noticed that the prevalences of the peaks in the pancreatic data set were well modeled by a Beta(0.5, 0.5) distribution, and the joint distribution of $\{\log(x_j), m_j, s_j\}^T$ across peaks was well modeled by a multivariate normal distribution with mean vector $(8.78, 9.34, 0.99)^T$ and covariance matrix

$$\Sigma = \begin{bmatrix} 0.536 & -0.108 & 0.104 \\ -0.108 & 0.503 & 0.057 \\ 0.104 & 0.057 & 0.156 \end{bmatrix}.$$

Recall that x_j , m_j , and s_j are the m/z value, mean log intensity, and standard deviation log intensity across samples containing the protein. These distributions characterized the variability of locations and intensities for the peaks in this data set in a way that also took into account the interrelationships among the different factors. For example, they accounted for the fact that peaks at lower masses tended to have larger mean log intensities ($\rho_{12} = -0.21$), and peaks with larger mean log intensities tended to also be more variable over samples ($\rho_{23} = 0.36$).

Each virtual population contained 150 true protein peaks. We obtained each peak's true mass x_j , prevalence π_j , and mean and standard deviation m_j and s_j by sampling from the distributions described above. For each virtual population, we ran one virtual experiment by taking n virtual samples from the population and obtaining the corresponding virtual spectra. The virtual MALDI-TOF instrument's settings were made to match the settings on a MALDI-

TOF instrument at M D Anderson Cancer Center, with assumed machine noise level σ .

We applied the SUDWT and MUDWT peak detection methods to the spectra from each virtual experiment and obtained a list of found peak locations $\{x_{U,j}^*\}, j = 1, \dots, p_U^*$ and $\{x_{M,j}^*\}, j = 1, \dots, p_M^*$, where p_U^* and p_M^* were the number of peaks found by the SUDWT and MUDWT methods, respectively. In preliminary studies, we determined that a wavelet threshold level of $\eta = 20$ worked well for both methods, and the tolerance settings $\delta_t = 7$ and $\delta_x = 0.002$ seemed optimal for the SUDWT method, so we kept these parameters fixed for all simulations. We found that the results were sensitive to the S/N threshold, so we ran each method using multiple thresholds. For the SUDWT, we used thresholds of $\phi=5, 10, 15, 20$, and 40 , and for the MUDWT we divided these quantities by \sqrt{n} to obtain roughly comparable thresholds. We assessed the relative performance of the two peak-finding methods, as described below.

We ran five simulation scenarios. For the first, the machine noise level was chosen to be comparable to the pancreatic data set ($\sigma = 66$) and each experiment consisted of $n = 100$ samples. The second two simulations also used $n = 100$ but the spectra had more or less noise ($\sigma = 200$ or $\sigma = 22$) than the pancreatic data. The final two simulations had the same noise level ($\sigma = 66$) as the pancreatic cancer data, but had larger ($n = 200$) or smaller ($n = 33$) sample sizes per experiment. The randomly generated spectra for each simulation study occupied on the order of 2-20GB of hard drive space, and required 6-18 hours to generate. The application of the two methods to the data from each simulation study took 16-48 hours to run in Matlab on a 3.0 GHz Pentium IV Windows 2000 machine with 2GB RAM.

4.4 Summarizing Simulation Results

We assessed how well the two methods performed peak detection by comparing the lists of peaks found by the SUDWT and MUDWT methods, $x_{U,j}^*; j = 1, \dots, p_U^*$ and $x_{M,j}^*; j = 1, \dots, p_M^*$, with the true locations of the protein peaks in the virtual population, $x_j; j = 1, \dots, p$. For a single data set, it was easy to make this assessment visually by simply plotting the mean spectrum and marking the locations of found and true peaks. However, it was not feasible to do this for all simulated data sets, so we needed to find a strategy for automatically summarizing the results.

Peak detection can be viewed as a special type of classification problem, since each m/z value on the spectrum has a true state (peak or not), and each peak detection method will classify each m/z value into one of the two states (peak or not). See Table 2.

[Insert Table 2 here]

This classification paradigm is difficult to apply to this context, for various reasons. First, the continuous nature of the m/z values x makes it difficult to automatically determine matches in the lists of true and found peaks. Because of the stochastic components in the virtual mass spectrometry instrument, the precise location of a given protein peak is not the same across all samples, and as a result the locations of found peaks do not perfectly match the location of the corresponding true peak. To address this problem, we defined a tolerance interval around each true peak inside of which any found peak was considered a match. Specifically, we considered a true peak at x_i and a found peak at x_j to be a match if $|x_i - x_j| < \gamma x_i$, where $\gamma = 0.003$ is the tolerance parameter. Using this criterion, we were able to decide whether each found peak matched with a true peak or not.

A second problem we encountered is that the specificity $D/(B + D)$ is always near 1. This is because D is the number of m/z values correctly called “not peaks”, which is nearly always $\gg B$ since the number of m/z values is in the 10,000s and the number of peaks is usually in the 100s. Thus, we decided that the most useful summary measures in this context were the sensitivity $= A/(A + C)$, measuring the proportion of true peaks detected, and the false discovery rate $FDR = B/(A + B)$, measuring the proportion of found peaks that did not match any true peak.

However, sensitivity and FDR are not sufficient summaries of performance in this setting, because it is possible for a single found peak to match multiple true peaks, or conversely for multiple found peaks to match a single true peak. The first case is unavoidable when multiple true peaks happen to fall within the tolerance limit of one another. This biases the sensitivity upwards, but should affect both methods equally. We found that the second case seems to occur when numerous secondary bumps on the primary peaks are detected as independent peaks. This is often a result of undersmoothing in the MUDWT method (see Figure 8) or using too small of a tolerance limit in the peak matching for the SUDWT method. This problem is associated with a high FDR, since there also tend to be many secondary peaks that fall outside the tolerance limits for matching true peaks. Besides the sensitivity and FDR, we also reported the proportion of these two types of multiple matching, with MM_1 summarizing the proportion of found peaks matching multiple true peaks and MM_2 the proportion of true peaks matching multiple found peaks.

[Insert Figure 8 here]

Thus, for each simulation study, we summarized the performance of each method with four

summary measures: sensitivity, FDR, MM_1 , and MM_2 . We reported the mean and range for each of these quantities, computed across the 100 virtual experiments. We also reported a *comparison proportion* for each measure, which is the proportion of the time the MUDWT outperformed the SUDWT for a given data set plus one half of the proportion of the times they tied. We also reported the sensitivities split out by prevalence and abundance groups to identify scenarios in which each peak detection method seemed to outperform the other.

4.5 Simulation Results

As described above, we ran each simulation using various choices for the S/N threshold ϕ . Table 3 contains the mean sensitivity and FDR across the 100 virtual experiments with $n = 200$ and $\sigma = 66$ for $\phi \in \{5, 10, 15, 20, 40\}$ for the SUDWT and $\phi \in \{5, 10, 15, 20, 40\}/\sqrt{200}$ for MUDWT. In general, we found that the sensitivity and FDR for both methods were sensitive to the choice of ϕ . For the other simulation scenarios, we only report the results for a single choice of ϕ , chosen as follows. We first chose ϕ for the SUDWT by finding the value giving an FDR closest to 0.10, allowing it to be slightly higher if there was an accompanying large increase in sensitivity or making it slightly lower if that caused little decrease in sensitivity. We then chose ϕ for the MUDWT that gave the largest FDR less than or equal to the FDR for the SUDWT. For $n = 200$ and $\sigma = 66$, by this criterion we chose $\phi = 20$ for the SUDWT and $\phi = 40/\sqrt{200} = 2.82$ for the MUDWT. Across simulations, the ϕ used for the SUDWT ranged from 15-40, while the ϕ used for the MUDWT ranged from 2.82 to 4.

[Insert Table 3 here]

Table 4 contains the overall results for each simulation. The MUDWT achieved better mean sensitivity than the SUDWT for all simulation scenarios, and had higher sensitivity for at least 97 out of the 100 virtual experiments in each scenario. The FDR was also slightly lower for the MUDWT method in most cases. The multiple match proportions (MM_1 and MM_2) were comparable between the two methods, tending to be slightly higher for the MUDWT method. As expected, peak detection was generally more difficult for smaller sample sizes, and was also more difficult when the noise level was $\sigma = 200$ compared to $\sigma = 66$. For some unknown reason, the $n = 100, \sigma = 22$ scenario resulted in lower sensitivity and higher FDR for both methods when compared with the $n = 100, \sigma = 66$ scenario.

[Insert Table 4 here]

Table 5 contains the sensitivities for the peaks sorted into different prevalence groups. Recall that prevalence is the proportion of samples in the population expressing that protein. We classified each protein peak as either extremely rare ($\pi_j < 0.05$, 14% of peaks), rare ($0.05 < \pi_j < 0.20$, 16%), common ($0.20 < \pi_j < 0.80$, 40%), or prevalent ($\pi_j > 0.80$, 30%). Not surprisingly, the sensitivities increased as a function of the prevalence; more prevalent peaks were easier to detect with both methods. There was no evidence of improved sensitivity for the MUDWT method for extremely rare or rare peaks, i.e., those present in less than 20% of the samples. This was not surprising, since in these cases the benefit of averaging over the n samples was partially counteracted by the fact that the peak was absent in a vast majority of the samples. For some simulation scenarios, the SUDWT appeared to be more sensitive than the MUDWT for these very rare peaks, although the differences were relatively small in magnitude when compared with the advantages for the MUDWT found elsewhere. For protein peaks that were reasonably prevalent, the MUDWT method clearly dominated. The MUDWT had much higher sensitivity for peaks present in at least 20% of the population.

[Insert Table 5 here]

Table 6 contains the sensitivities for the peaks sorted by abundance groups, defined based on the mean \log_2 intensities across samples containing the protein (< 9 , $9 - 9.5$, $9.5 - 10$, and > 10). Sensitivity increased with abundance, as expected. Use of the average spectrum had the most benefit for the less abundant proteins. In the lowest abundance group (which accounted for 31% of the peaks), we found that the MUDWT had a higher sensitivity than the SUDWT at least 95% of the time, with mean sensitivity differences around 10-15%. We also saw large gains from using the average spectrum in the second and third abundance groups. There was much less improvement for the most abundant proteins, with the mean sensitivities nearly equivalent for some of the simulation scenarios we considered.

[Insert Table 6 here]

To investigate the possibility of a prevalence-by-abundance interaction, for the $n = 100$, $\sigma = 66$ simulations we computed the mean sensitivities for both methods sorted by groups defined by both prevalence and abundance (see Table 7). There was strong evidence of an interaction. The greatest relative benefit for the MUDWT over the SUDWT occurred for peaks with low abundance but high prevalence. For the lowest abundance/highest prevalence group, which accounted for, on average, 10% of the peaks, the SUDWT achieved a mean sensitivity of 0.76,

while the MUDWT achieved a mean sensitivity of 0.94. The MUDWT achieved higher sensitivity than the SUDWT for this group for 86 of the 100 virtual populations; the SUDWT achieved higher sensitivity in 5 of the 100; and the two methods tied in 9 out of the 100. These results agreed with our earlier claim that the benefit of using the average spectrum is maximized for low intensity peaks that are present across many spectra. Figure 9 shows one such example. The peak at 2835 is not discernable from the noise in the individual spectrum, but its reinforcement across spectra makes it evident in the mean spectrum. This peak was detected by the MUDWT method, but not the SUDWT.

[Insert Table 7 here]

[Insert Figure 9 here]

Conversely, the MUDWT tended to have lower sensitivity than the SUDWT for the extremely rare ($\pi < 0.05$) peaks that had high abundance. For the most abundant/least prevalent group, accounting for, on average, 2% of the total number of peaks, the mean sensitivity was 0.52 for the SUDWT and 0.40 for the MUDWT. For this group, the SUDWT achieved higher sensitivity than the MUDWT for 40 of the 100 virtual populations; the MUDWT was higher for 7 out of the 100; and the methods tied for 53 of the 100. While it was clear that the SUDWT performed better for this subset of peaks, the MUDWT still achieved at least as high a level of sensitivity for 60 of the 100 virtual populations.

5 Discussion

Averaging is a fundamental principle underlying many statistical methods, and the Central Limit Theorem is arguably the fundamental theorem of statistics. In this paper, we put these simple ideas to work in order to improve peak detection for mass spectrometry data. We have demonstrated in real data examples and through our simulation studies that use of the mean spectrum leads to increased sensitivity for peak detection. This effect is especially strong for the low intensity peaks, which are frequently the peaks in which biomedical investigators are most interested. There may be a slight tradeoff for some of the rarest peaks, although our simulation studies suggest that this difference is small compared to the improvements seen elsewhere. Peak detection based on the average spectrum is also more straightforward and computationally efficient than peak detection based on the individual spectra. On our Pentium-IV 3.0GHz Windows 2000 machine with 2GB RAM, our MATLAB code for peak detection

and quantification takes 45 seconds for one experiment containing 100 spectra when using the MUDWT, while the SUDWT on the same data takes roughly 8 minutes.

In order to implement our MUDWT method, two parameters must be set: the wavelet threshold η and the S/N threshold ϕ . We found that $\eta = 20$ and $\phi = 4$ tended to work best for our simulated data, but it is difficult to know how well these settings will transfer to other data sets. For η , there is a careful balance to strike since making it too small results in undersmoothing, which causes the procedure to find many spurious secondary peaks, while making it too large results in oversmoothing that may eliminate some of the low intensity peaks. If ϕ is made much smaller than 4, we have found in our simulations that the false discovery rate is greatly increased while few new true peaks are discovered. We recommend starting with these levels, then visually inspecting plots of the raw and wavelet smoothed average spectrum to check whether it seems to be detecting features that appear to be peaks. Automatic methods for selecting these parameters would be welcome and would make this method easier to use.

The MUDWT appears to have more problems with undersmoothing, especially when the noise level is low, as evidenced by the high MM_2 rates (see Figure 8). To combat this problem, we suggest combining together peaks that have high correlations (> 0.95 , for example) by summing their quantifications. This largely eliminates the problem caused by undersmoothing, since the secondary peaks should be very highly correlated with each other if they correspond to the same true protein peak. An added benefit of this practice is that it combines information across peaks representing doubly-charged versions of same protein, as well as matrix adducts and other alterations not expected to be biologically meaningful. This reduces the dimensionality of the data further without resulting in a loss of information, since if correlation is so high, there is very little additional information contained in the redundant peaks. Thus, we don't need to fear that this practice will cause us to miss out on important alterations of proteins that are informative for the underlying biological processes, e.g., phosphorylated proteins.

We have shown that using the average spectrum improves peak detection using the method based on the UDWT algorithm that was introduced by Coombes, et al., (2004b). However, the idea of using the mean spectrum for peak detection is general, and could be paired with other peak detection methods. We expect that similar or greater relative improvements could be realized when other peak detection methods to the average spectrum instead of the individual spectra. Also, while this paper focused exclusively on preprocessing mass spectrometry data, our mean function, UDWT-based peak detection procedure can be generalized to perform feature

extraction for other types of functional data where the features of interest are peaks. This includes image data, which can be viewed as functional data with a two-dimensional domain.

We have also demonstrated how the virtual mass spectrometer introduced in Coombes, et al. (2004a) can be used to conduct a simulation study. Since the virtual instrument is based upon the key physical principles underlying the real instrument, it yields virtual spectra that look much like real mass spectrometry data. Thus, we feel that this is a valuable tool for studying the technology and comparing statistical methods for analyzing these types of data. There is still more work to do to improve the virtual instrument, however. Common alterations of proteins, such as matrix and sodium adducts (additions of matrix or sodium molecules) or neutral losses of water, ammonia, or carbon, should also be incorporated into the modeling. Also, causes of the baseline artifact need to be better understood, so that a more realistic model for the baseline that is based on the technology can be used in lieu of the exponential curve used here.

We have mentioned that functional data analysis and feature extraction represent two possible approaches for analyzing mass spectrometry data. In this paper, we have focused on feature extraction of the peaks, which has the advantages of reducing the dimensionality of the data in a scientifically meaningful way, since the peaks represent the proteins, the fundamental units of interest in the data. Because subsequent analyses are performed only on the detected peaks, it is crucial to use an effective method for peak detection. The MUDWT method introduced in this paper appears to work quite well. No matter the level of confidence in one's methods for peak detection and the other preprocessing steps, however, we strongly advise against blindly trusting the output of these methods. The data analyst should always take some time to look back at the raw spectra to visually verify that the results found are supported by the data.

6 Acknowledgements

The authors would like to thank Henry Kuerer and Mien-Chie Hung for permission to use the nipple aspirate fluid SELDI data set; I.J. Fidler, Stanley Hamilton, James Abbruzzese, Donghui Li, and Nancy Shih for permission to use the pancreatic cancer MALDI-TOF data set; and Drs. Emmanuel Petricoin and Lance Liotta for making their QStar data publicly available on their web site (<http://www.ncifdaproteomics.com>). This research was supported in part by NIH/NCI grants R01 CA-107304 and P50 CA070907.

7 References

- Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, and Wright GL Jr. (2002). Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* **62**, 3609–3614.
- 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, and Terrett JA. (2003). Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. *Journal of Biological Chemistry* **278**, 6482–6489.
- Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC, and 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**, 1667–1672.
- Baggerly KA, Morris JS, and Coombes KR. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20**, 777–785.
- Beavis RC and Chait BT. (1991) Velocity distributions of intact high mass polypeptide molecule ions produced by matrix assisted laser desorption. *Chemical Physics Letters* **181**, 479–484.
- Billheimer D. (2004). A functional data approach to MALDI-TOF protein analysis. *Unpublished Report*.
- Carpenter M., Melath M., Zhang S., and Grizzle W.E. (2003). Statistical processing and analysis of proteomic and genomic data. *Proceedings of the Pharmaceutical SAS Users Group*, Miami, FL.
- Coombes KR, Kooman JM, Baggerly KA, Morris JS, and Kobayashi R. (2004a). Understanding the characteristics of mass spectrometry data through the use of simulation. *M.D. Anderson Biostatistics Technical Report* **UTMDABTR-002-04**.
- Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, and Kuerer HM. (2004b). 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. *M.D. Anderson Biostatistics Technical Report* **UTMDABTR-001-04**.
- Coombes KR, Wang J, and Baggerly KA (2004). A statistical method for finding biomarkers from microarray data, with application to prostate cancer. *M.D. Anderson Biostatistics Technical Report* **UTMDABTR-007-04**.
- Fung ET and Enderwick C. (2002). ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques* **Suppl.**, 34–41.
- Gluckmann M and Karas M (1999). The initial ion velocity and its dependence on matrix, analyte, and preparation method in ultraviolet matrix-assisted laser desorption/ionization.

Journal of Mass Spectrometry **34**(5), 467–477.

- Karas M, Bahr U, Fournier I, Gluckmann M and Pfenninger A (2003). The initial-ion velocity as a marker for different desorption-ionization mechanisms in MALDI. *International Journal of Mass Spectrometry* **226**(1), 239–248.
- Morris JS and Carroll RJ. (2004). Wavelet-based functional mixed models. *M.D. Anderson Biostatistics Technical Report* **UTMDABTR-006-04**.
- 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, and Petricoin EF 3rd. (2001). Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Disease Markers* **17**, 301–307.
- Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, and Liotta LA. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577.
- Qu Y, Adam BL, Thonquist M, Potter JD, Thompson ML, Yasui Y, Davis J, Schellhammer PF, Cazares L, Clements M, Wright GL Jr., and Feng Z. (2003). Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics* **59**, 143–151.
- Ramsay J and Silverman BW. (1997). *Functional Data Analysis* Springer, New York.
- Schaub S, Wilkins J, Weiler T, Sangster K, Rush D, and Nickerson P. (2004). Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney International* **65**, 323–332.
- Sorace JM and 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, and Wernert N. (2002). Analysis of microdissected prostate tissue with ProteinChip arrays—a way to new insights into carcinogenesis and to diagnostic tools. *International Journal of Molecular Medicine* **9**, 341–347.
- Yasui Y, McLerran D, Adam BL, Winget M, Thornquist M, and Feng Z. (2003a). An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Biomedical Biotechnology* **2003**, 242–248.
- Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL Jr, Qu Y, Potter JD, Winget M, Thornquist M, and Feng Z. (2003b). A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4** 449-463.
- Zhukov TA, Johanson RA, Cantor AB, Clark RA, and Tockman MS. (2003). Discovery of

distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer* **40**, 267–279.

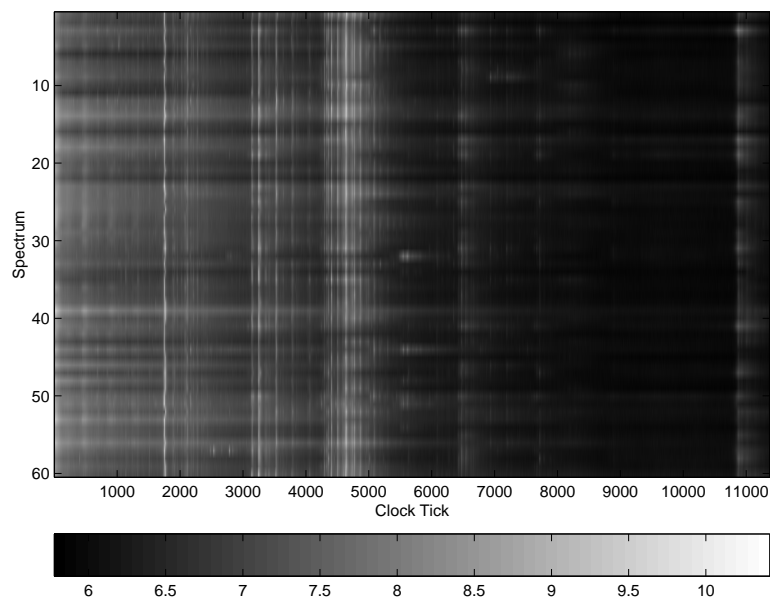


Figure 1: *Checking Calibration*. Heat map of the logarithmic intensities of 60 spectra related to pancreatic cancer. Bright vertical lines are the largest peaks, which are well-aligned across spectra.

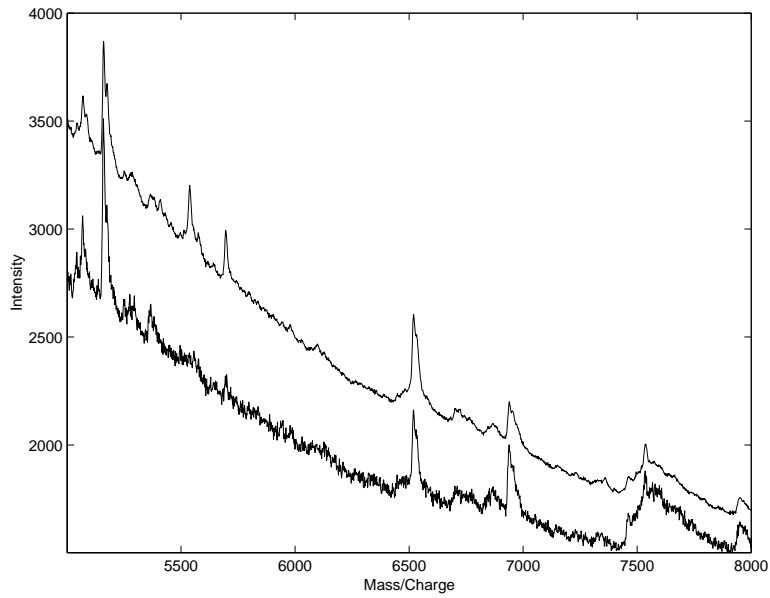


Figure 2: *Noise Reduction in the Mean Spectrum.* Portion of one raw spectrum (lower curve) and the mean of 24 replicate spectra (upper curve). The noise is reduced in the mean spectrum by a factor of about 5.

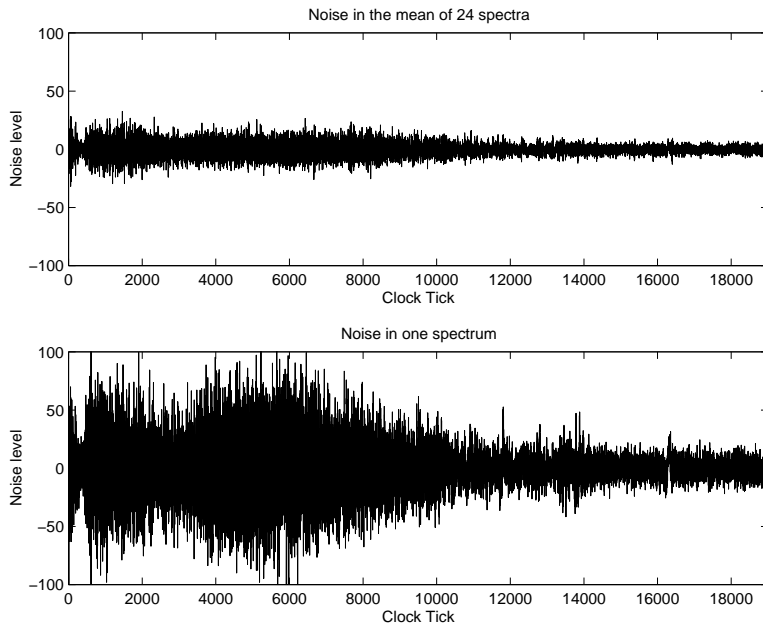


Figure 3: *Noise Levels.* Plots of the noise removed by the SUDWT in the mean of 24 spectra (top) and in one individual spectrum (bottom). The noise is reduced in the mean spectrum by a factor of about 5.

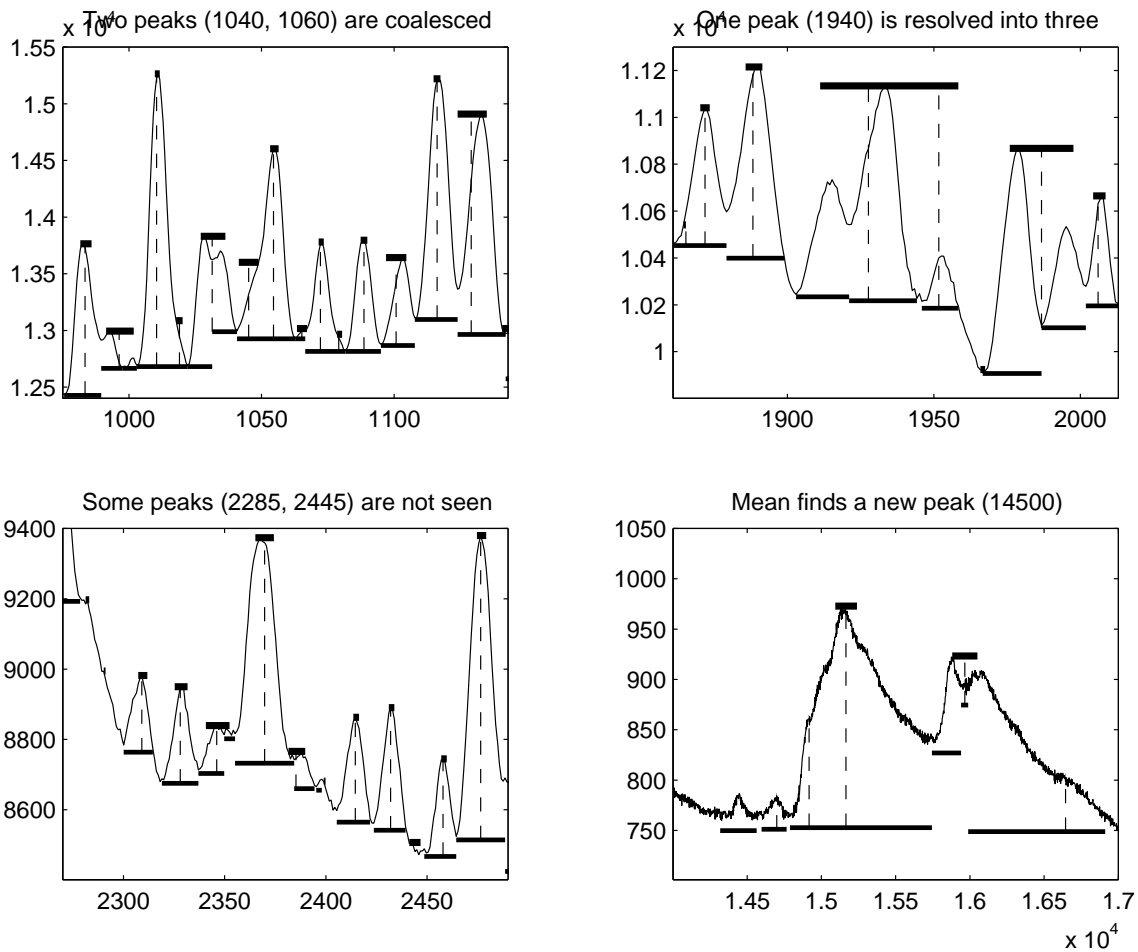


Figure 4: *Comparing Peak Detectors*. Plots of the mean spectrum illustrating differences in peak finding methods. Bars above peaks indicate regions where peaks in individual spectra were matched. Bars below peaks indicate the width of peaks found in the mean. Dotted vertical lines join peaks found by both methods.

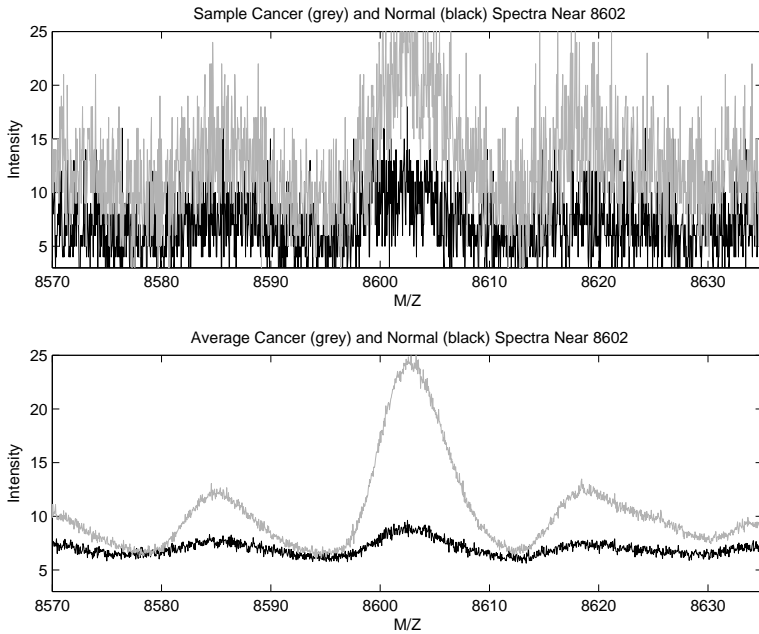


Figure 5: *Short Peak*. Plots showing an arbitrarily chosen normal and cancer spectrum (top) and the mean spectra across 95 normal spectra and 121 ovarian cancer spectra (bottom) in the neighborhood of a significant peak at 8602 daltons. This represents a protein that is more abundant in cancer patients. This peak would be difficult to detect on individual spectra, but is easily detected in the mean spectra for cancer and normal groups, and clearly would also be detected on the overall mean spectrum.

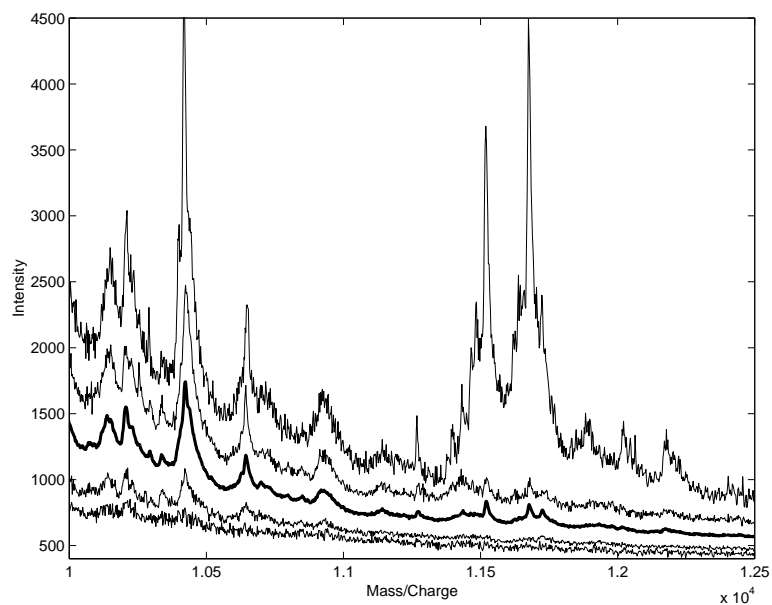


Figure 6: *Rare Peak*. Plots of spectra obtained by pointwise application of statistical functions to 60 spectra from a study of pancreatic cancer. From top to bottom, the spectra are the maximum, 90th percentile, mean, 10th percentile, and minimum. Note the occurrence of large peaks in the maximum at 11,500 daltons and 11,600 daltons that are barely present in the 90th percentile; these peaks are visible in the mean spectrum.

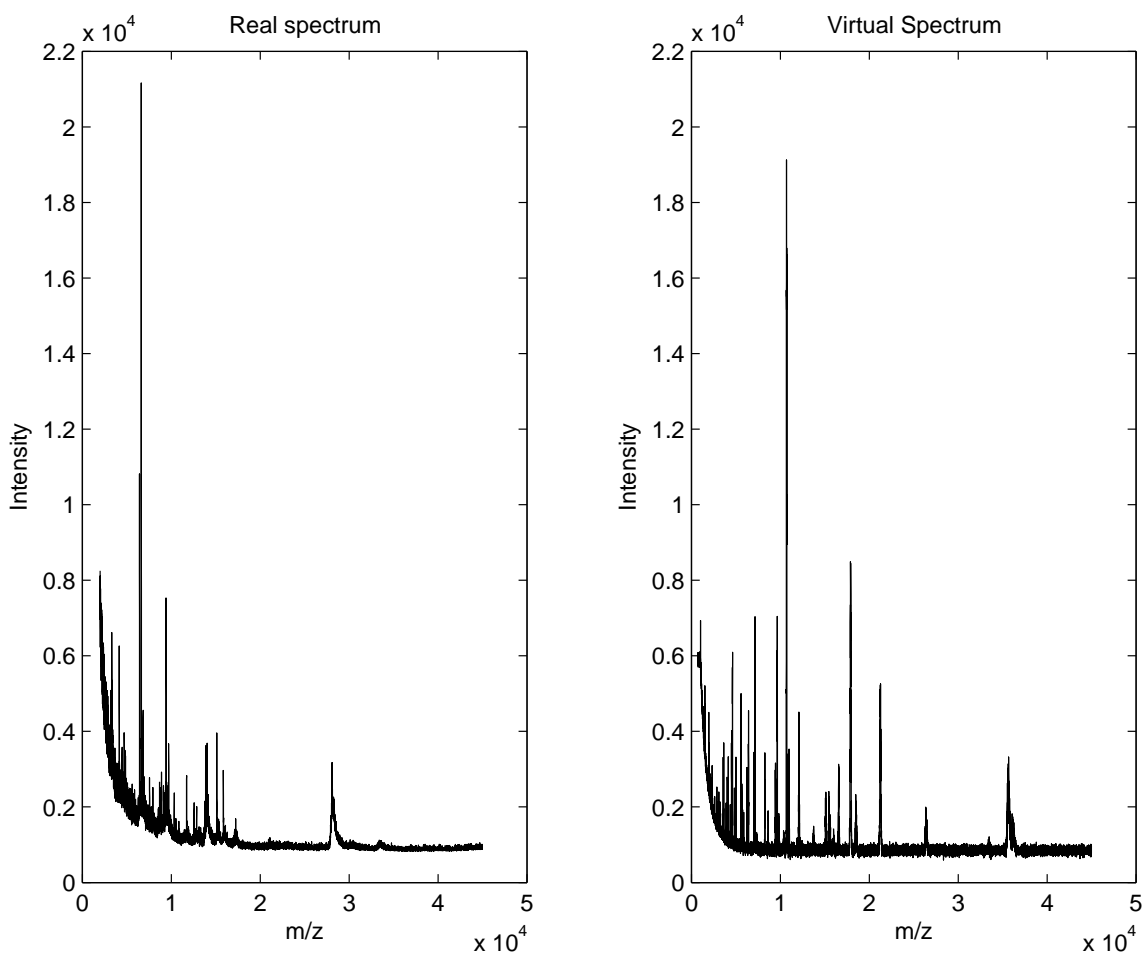


Figure 7: *Real and Virtual Spectra*. Plot of a true MALDI-TOF spectrum and a virtual MALDI-TOF spectrum from our virtual mass spectrometer.

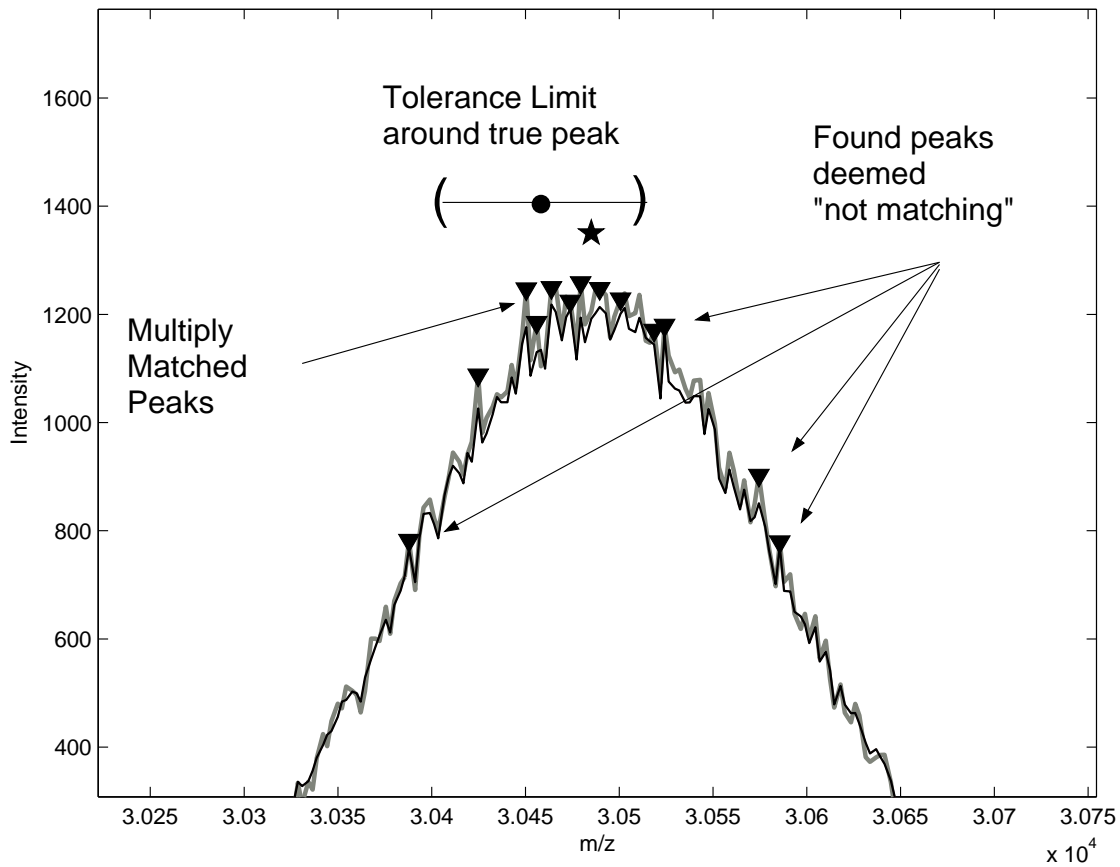


Figure 8: *Undersmoothed peak*. Two lines are plotted here in the local neighborhood of a high mass peak: the grey line is the mean spectrum while the black line is the wavelet smoothed mean spectrum, with threshold parameter 20, for one data set in the $n = 100$ and $\sigma = 22$ simulation. The dot marks a true peak at 30,458, and the interval marks the 0.2% tolerance limit to declare a found peak a match. The star marks the peak as found by the SUDWT method, while the triangles mark peaks found by the MUDWT method. Note that the threshold parameter 20 led to undersmoothing of the average spectrum, which caused many redundant secondary peaks near the true peak, which inflated MM_2 , and several secondary peaks outside the tolerance limit, which inflated the FDR.

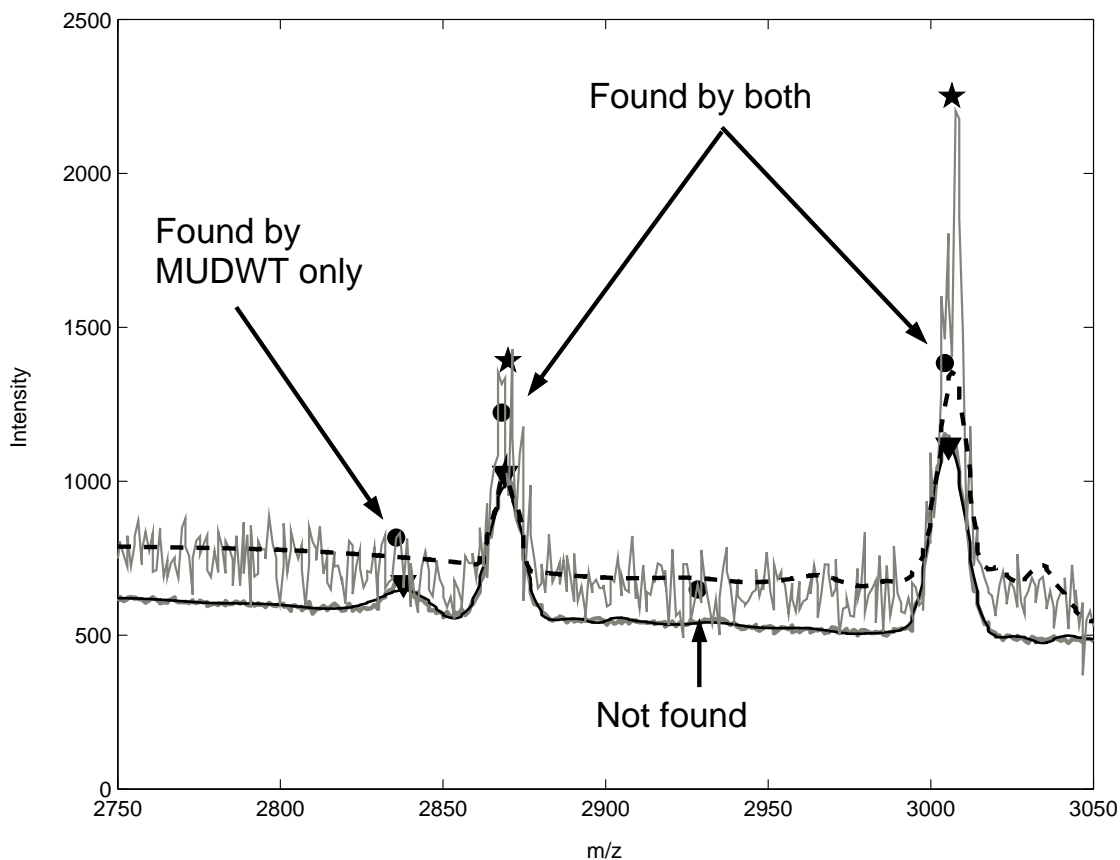


Figure 9: *Some Peaks*. Four lines from one of the data sets from the simulation with $n = 100$ and $\sigma = 66$ are plotted here: The two “noisy” grey lines are one selected raw spectrum and the mean spectrum. The dashed line is the wavelet-smoothed version of the individual spectrum, and the solid line is the wavelet-smoothed average spectrum. The dots mark true peaks, while the triangles mark peaks found by the MUDWT method and the stars mark peaks found by the SUDWT method. The peak at 2835 had low abundance ($m = 7.8, \pi = 0.40$) and was found by MUDWT, but not SUDWT. Peaks 2867 ($m = 8.8, \pi = 0.73$) and 3004 ($m = 9.0, \pi = 0.66$) were found by both methods. Peak 2928 was very rare ($m = 8.2, \pi = 0.04$) and went undetected by both methods.

Table 1: Number of peak sets detected by matching individual peaks across samples.

| | | | | | | | | | | | |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Threshold | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| Number of Peaks | 174 | 176 | 179 | 181 | 189 | 206 | 208 | 205 | 204 | 191 | 179 |

Table 2: Peak Detection as Classification

| | | | |
|--------------------|----------|-------------------|----------|
| | | True State | |
| | | Peak | Not Peak |
| found state | Peak | A | B |
| | Not Peak | C | D |

Table 3: Mean sensitivity and FDR across 100 virtual populations for different S/N thresholds, $n=200$, $\sigma=66$ simulation

| SUDWT | | | MUDWT | | |
|-------|-------------|------|-------|-------------|------|
| S/N | Sensitivity | FDR | S/N | Sensitivity | FDR |
| 5 | 0.75 | 0.46 | 0.35 | 0.88 | 0.52 |
| 10 | 0.75 | 0.26 | 0.71 | 0.88 | 0.49 |
| 15 | 0.75 | 0.15 | 1.06 | 0.88 | 0.45 |
| 20 | 0.74 | 0.12 | 1.41 | 0.87 | 0.37 |
| 40 | 0.70 | 0.09 | 2.82 | 0.85 | 0.11 |

Table 4: *Overall results from the simulation study.* The top element in each box is the mean quantity over the 100 virtual experiments, and the bottom interval is the range. Sensitivity, FDR, MM_1 , MM_2 , and the comparison proportion are defined in Section 4.4.

| n | σ | Method | Sensitivity | FDR | MM_1 | MM_2 |
|-----|----------|------------|----------------------|----------------------|----------------------|----------------------|
| 100 | 66 | SUDWT | 0.75 (0.60, 0.85) | 0.09 (0.02, 0.26) | 0.10 (0.01, 0.22) | 0.16 (0.08, 0.27) |
| | | MUDWT | 0.83 (0.75, 0.92) | 0.06 (0.00, 0.41) | 0.12 (0.03, 0.25) | 0.19 (0.12, 0.33) |
| | | Comparison | 0.97 | 0.80 | 0.33 | 0.16 |
| 100 | 22 | SUDWT | 0.58 (0.43, 0.69) | 0.25 (0.11, 0.41) | 0.10 (0.03, 0.21) | 0.13 (0.06, 0.21) |
| | | MUDWT | 0.74 (0.61, 0.84) | 0.23 (0.10, 0.52) | 0.14 (0.07, 0.22) | 0.19 (0.13, 0.38) |
| | | Comparison | 1.00 | 0.63 | 0.16 | 0.03 |
| 100 | 200 | SUDWT | 0.70 (0.61, 0.80) | 0.08 (0.00, 0.17) | 0.11 (0.02, 0.22) | 0.16 (0.09, 0.25) |
| | | MUDWT | 0.78 (0.69, 0.87) | 0.05 (0.00, 0.45) | 0.08 (0.02, 0.17) | 0.17 (0.11, 0.26) |
| | | Comparison | 0.97 | 0.86 | 0.80 | 0.31 |
| 33 | 66 | SUDWT | 0.73 (0.63, 0.84) | 0.09 (0.01, 0.20) | 0.11 (0.02, 0.22) | 0.17 (0.11, 0.27) |
| | | MUDWT | 0.80 (0.74, 0.86) | 0.06 (0.00, 0.36) | 0.12 (0.03, 0.21) | 0.20 (0.11, 0.37) |
| | | Comparison | 0.99 | 0.85 | 0.37 | 0.27 |
| 200 | 66 | SUDWT | 0.75 (0.58, 0.87) | 0.12 (0.02, 0.46) | 0.11 (0.03, 0.20) | 0.17 (0.08, 0.25) |
| | | MUDWT | 0.85 (0.75, 0.91) | 0.11 (0.00, 0.31) | 0.12 (0.05, 0.22) | 0.20 (0.12, 0.36) |
| | | Comparison | 1.00 | 0.69 | 0.41 | 0.14 |

Table 5: *Sensitivity by prevalence group.* The sensitivities for peaks in different prevalence groups are given, along with the proportion of peaks in each prevalence group. The first number in each box is the mean sensitivity for the indicated method in that prevalence group across the 100 virtual experiments, while the interval on the second line indicates the range.

| n | σ | Method | Sensitivity by prevalence | | | |
|-----|----------|------------|---------------------------|----------------------|----------------------|----------------------|
| | | | <0.05 (14%) | 0.05-0.20 (16%) | 0.20-0.80 (40%) | >0.80 (30%) |
| 100 | 66 | SUDWT | 0.43 (0.20, 0.75) | 0.74 (0.50, 0.96) | 0.81 (0.60, 0.93) | 0.82 (0.59, 0.95) |
| | | MUDWT | 0.38 (0.16, 0.67) | 0.74 (0.54, 0.95) | 0.93 (0.78, 1.00) | 0.97 (0.89, 1.00) |
| | | Comparison | 0.25 | 0.49 | 1.00 | 0.99 |
| 100 | 22 | SUDWT | 0.39 (0.09, 0.67) | 0.62 (0.08, 0.85) | 0.62 (0.39, 0.83) | 0.60 (0.41, 0.88) |
| | | MUDWT | 0.39 (0.12, 0.63) | 0.66 (0.42, 0.88) | 0.81 (0.62, 0.94) | 0.84 (0.66, 0.97) |
| | | Comparison | 0.51 | 0.64 | 1.00 | 1.00 |
| 100 | 200 | SUDWT | 0.33 (0.00, 0.65) | 0.59 (0.35, 0.84) | 0.78 (0.63, 0.92) | 0.82 (0.67, 0.98) |
| | | MUDWT | 0.30 (0.07, 0.68) | 0.57 (0.29, 0.84) | 0.89 (0.73, 0.96) | 0.96 (0.87, 1.00) |
| | | Comparison | 0.40 | 0.40 | 0.95 | 1.00 |
| 33 | 66 | SUDWT | 0.32 (0.11, 0.61) | 0.60 (0.38, 0.80) | 0.82 (0.66, 0.94) | 0.86 (0.73, 0.98) |
| | | MUDWT | 0.32 (0.11, 0.57) | 0.62 (0.33, 0.83) | 0.91 (0.80, 1.00) | 0.98 (0.90, 1.00) |
| | | Comparison | 0.51 | 0.54 | 0.95 | 1.00 |
| 200 | 66 | SUDWT | 0.48 (0.16, 0.78) | 0.76 (0.50, 1.00) | 0.79 (0.62, 0.93) | 0.80 (0.55, 0.97) |
| | | MUDWT | 0.44 (0.19, 0.78) | 0.81 (0.54, 0.96) | 0.92 (0.85, 1.00) | 0.97 (0.89, 1.00) |
| | | Comparison | 0.38 | 0.70 | 0.97 | 0.98 |

Table 6: *Sensitivity by abundance group.* The sensitivities for peaks in different abundance groups are given, along with the proportion of peaks in each abundance group. The first number in each box is the mean sensitivity for the indicated method in that prevalence group across the 100 virtual experiments, while the interval on the second line indicates the range.

| n | σ | Method | Sensitivity by mean log intensity | | | |
|-----|----------|------------|-----------------------------------|----------------------|----------------------|----------------------|
| | | | <9.0 (31%) | 9.0-9.5 (27%) | 9.5-10 (23%) | >10 (19%) |
| 100 | 66 | SUDWT | 0.68 (0.48, 0.90) | 0.75 (0.53, 0.95) | 0.78 (0.51, 0.94) | 0.82 (0.56, 1.00) |
| | | MUDWT | 0.78 (0.60, 0.91) | 0.84 (0.68, 0.97) | 0.85 (0.68, 0.97) | 0.88 (0.70, 1.00) |
| | | Comparison | 0.97 | 0.89 | 0.84 | 0.78 |
| 100 | 22 | SUDWT | 0.56 (0.37, 0.76) | 0.58 (0.30, 0.89) | 0.61 (0.38, 0.81) | 0.61 (0.36, 0.94) |
| | | MUDWT | 0.70 (0.53, 0.85) | 0.73 (0.50, 0.86) | 0.75 (0.50, 0.86) | 0.78 (0.56, 0.91) |
| | | Comparison | 0.99 | 0.96 | 0.93 | 0.96 |
| 100 | 200 | SUDWT | 0.58 (0.40, 0.75) | 0.69 (0.45, 0.86) | 0.75 (0.52, 0.95) | 0.83 (0.61, 0.97) |
| | | MUDWT | 0.73 (0.55, 0.88) | 0.77 (0.61, 0.91) | 0.80 (0.63, 0.95) | 0.84 (0.56, 1.00) |
| | | Comparison | 0.98 | 0.86 | 0.78 | 0.54 |
| 33 | 66 | SUDWT | 0.63 (0.47, 0.77) | 0.73 (0.47, 0.86) | 0.78 (0.56, 0.93) | 0.84 (0.67, 1.00) |
| | | MUDWT | 0.75 (0.57, 0.88) | 0.80 (0.58, 0.93) | 0.83 (0.70, 0.95) | 0.85 (0.68, 1.00) |
| | | Comparison | 0.96 | 0.86 | 0.77 | 0.59 |
| 200 | 66 | SUDWT | 0.69 (0.49, 0.84) | 0.73 (0.52, 0.91) | 0.78 (0.54, 0.94) | 0.82 (0.48, 1.00) |
| | | MUDWT | 0.81 (0.71, 0.96) | 0.85 (0.69, 0.97) | 0.87 (0.68, 0.97) | 0.90 (0.74, 1.00) |
| | | Comparison | 0.98 | 0.89 | 0.86 | 0.79 |

Table 7: *Interaction of prevalence and abundance.* Relative performance of SUDWT and MUDWT for detecting peaks in $n = 100/\sigma = 66$ simulation with different combinations of prevalence and abundance. The first row in each cell contains the mean sensitivities across 100 virtual experiments for the SUDWT/MUDWT methods. The second row contains the comparison proportion p , measuring the proportion of the virtual experiments for which the MUDWT had higher sensitivity than the SUDWT plus one-half the proportion for which the methods tied.

| Sensitivity (SUDWT/MUDWT) | | Abundance(mean log₂ intensity) | | | |
|----------------------------------|-----------|---|----------------|---------------|---------------|
| | | <9.0 | 9.0-9.5 | 9.5-10 | >10 |
| Prevalence | <0.05 | 0.36/0.34 | 0.46/0.42 | 0.43/0.36 | 0.52/0.40 |
| | | 0.46 | 0.43 | 0.39 | 0.34 |
| | 0.05-0.20 | 0.62/0.65 | 0.72/0.76 | 0.80/0.78 | 0.86/0.83 |
| | | 0.55 | 0.53 | 0.44 | 0.48 |
| | 0.20-0.80 | 0.75/0.88 | 0.80/0.93 | 0.86/0.96 | 0.89/0.98 |
| | | 0.92 | 0.87 | 0.87 | 0.78 |
| | >0.80 | 0.76/0.94 | 0.83/0.96 | 0.86/0.99 | 0.87/0.99 |
| | | 0.91 | 0.87 | 0.87 | 0.80 |