# Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum

**Article** *in* Bioinformatics · June 2005

DOI: 10.1093/bioinformatics/bti254 · Source: PubMed

Jeffrey S Morris
University of Texas MD Anderson Cancer Center
**329** PUBLICATIONS **17,931** CITATIONS

SEE PROFILE

Kevin Robert Coombes
The Ohio State University
**540** PUBLICATIONS **28,206** CITATIONS

SEE PROFILE

John Koomen
Moffitt Cancer Center
**332** PUBLICATIONS **9,879** CITATIONS

SEE PROFILE

BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 9 2005, pages 1764–1775
doi:10.1093/bioinformatics/bti254

Genome analysis

# Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum

Jeffrey S. Morris[1,*,†], Kevin R. Coombes[1,†], John Koomen[2], Keith A. Baggerly[1] and Ryuji Kobayashi[2]

[1]Department of Biostatistics and Applied Mathematics and [2]Department of Molecular Pathology,
The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Received on September 21, 2004; revised on December 6, 2004; accepted on December 27, 2004
Advance Access publication February 22, 2005

**ABSTRACT**

**Motivation:** Mass spectrometry yields complex functional data for which the features of scientific interest are peaks. A common two-step approach to analyzing these data involves first extracting and quantifying the peaks, then analyzing the resulting matrix of peak quantifications. Feature extraction and quantification involves a number of interrelated steps. It is important to perform these steps well, since subsequent analyses condition on these determinations. Also, it is difficult to compare the performance of competing methods for analyzing mass spectrometry data since the true expression levels of the proteins in the population are generally not known.

**Results:** In this paper, we introduce a new method for feature extraction in mass spectrometry data that uses translation-invariant wavelet transforms and performs peak detection using the mean spectrum. We examine the method's performance through examples and simulation, and demonstrate the advantages of using the mean spectrum to detect peaks. We also describe a new physics-based computer model of mass spectrometry and demonstrate how one may design simulation studies based on this tool to systematically compare competing methods.

**Availability:** MATLAB scripts to implement the methods described in this paper and R code for the virtual mass spectrometer are available at http://bioinformatics.mdanderson.org/ software.html

**Contact:** jefmorris@mdanderson.org

**Supplementary information:** http://bioinformatics.mdanderson.org/supplements.html

## 1 INTRODUCTION

Mass spectrometry is being used increasingly to detect disease-related proteomic patterns in complex mixtures of proteins derived from tissue samples, or from more easily obtained biological fluids such as serum, urine or nipple aspirate fluid (Paweletz *et al.*, 2000, 2001; Wellmann *et al.*, 2002; Adam *et al.*, 2002, 2003; Zhukov *et al.*, 2003; Schaub *et al.*, 2004). These proteomic patterns can potentially be used for identifying biomarkers for early diagnosis, prognosis, monitoring disease progression or response to treatment, or identifying which patients are most likely to benefit from particular treatments.

The mass spectrometry instruments most commonly used in clinical and biological applications rely on a matrix-assisted laser desorption and ionization (MALDI) ion source and a time-of-flight (TOF) detection system. A typical dataset arising in MALDI–TOF contains tens or hundreds of spectra, with each spectrum containing tens of thousands of intensity measurements representing an unknown number of protein peaks. From a modeling viewpoint, these spectra can be considered complex functional data in which the key features of scientific interest are the peaks. While comprehensive functional data analytic approaches are possible (Morris and Carroll, 2004; Billheimer, unpublished report), a common two-step approach focuses on the peaks. The first step involves feature extraction and quantification, in which one identifies the peak locations and quantifies each peak in each spectrum. This requires one to deal with several modeling issues simultaneously, including calibration of the spectra, baseline correction, normalization and denoising. Assuming that one finds $p$ peaks from $n$ spectra, this yields a $p \times n$ matrix of 'protein expression levels'. The second step consists of using this matrix to search for proteins that may be differentially expressed between experimental conditions or correlated with clinical outcomes, perform unsupervised clustering or apply supervised learning methods to perform discrimination and classification.

In recent years, there has been a great deal of methodological research on the second step of this approach, whereby statistical data mining techniques are applied to the matrix of expression levels. Much of this development has taken place in the context of microarrays, but in general the same methods also may be applied to mass spectrometry proteomics. Hastie *et al.* (2001) provide an excellent review of such methods. However, the second step presumes the validity of the first.

It is important to perform the first step well, since subsequent analyses condition on these determinations. It has been shown that the use of inadequate or ineffective methods in the first step may make it difficult to extract meaningful biological information from these data (Sorace and Zhan, 2003; Baggerly *et al.*, 2003, 2004). There have been several recent papers dealing with these issues (Yasui *et al.*, 2003a; Baggerly *et al.*, 2003; Coombes *et al.*, 2003; Malyarenko *et al.*, 2005).

In this paper, we focus on the first step. We describe a comprehensive approach for performing feature extraction and quantification

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

for mass spectrometry data. Our method is easy to implement and algorithm-based and, as we will show, it appears to perform very well in practice. We perform peak detection on the average spectrum, which leads to greater sensitivity and specificity while eliminating the difficult and intrinsically error-laden step of matching peaks detected on individual spectra. While conceptually straightforward, to our knowledge, no existing software or papers take this approach. Carpenter *et al*. (2003) mention the idea in passing, but they focus their attention on 'binning' methods for data reduction. We illustrate the advantages gained by using the average spectrum to detect peaks, through analysis of sample datasets and more systematically, as described below.

We also introduce a simulation-based approach for systematically comparing different methods for analyzing mass spectrometry data. Comparison of algorithms is complicated by the fact that, in general, the true expression levels of the proteins in the population are not known. In order to obtain a gold standard against which one may compare methods, it is necessary to simulate the data reliably. We have developed a computer model, a 'virtual mass spectrometer', that is based on the physical principles underlying the instrument and can be used to generate realistic virtual spectra (Coombes *et al*., 2005a). In this paper, we demonstrate how to design simulation studies using this tool. The simulation study we perform compares the peak detection performance of the feature extraction method described in this paper with a similar method that performs peak detection on the individual spectra rather than using the average spectrum.

The remainder of this paper is organized as follows. In Section 2, we provide details on mass spectrometry, present a conceptual model for mass spectrometry data and discuss the specific issues arising in feature extraction and quantification. In Section 3, we describe two methods for feature extraction and quantification, one which operates on individual spectra and the other in which the peak detection is performed on the average spectrum. In Section 4, we describe the virtual mass spectrometer and outline how to use it to conduct simulation studies. Section 5 examines the method's performance using examples and a simulation study, demonstrating that using the average spectrum for peak detection provides improved sensitivity and specificity compared to a similar approach based on individual spectra. Discussion and conclusions are given in Section 6.

## 2 ELEMENTS OF FEATURE EXTRACTION AND QUANTIFICATION

To run an experiment on a MALDI–TOF instrument, the biological sample is first mixed with an energy absorbing matrix, which causes the mixture to crystallize as it dries. The metal plate containing the crystallized sample is then placed into a vacuum chamber and the crystal is 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. An electric field accelerates the ionized proteins into a flight tube, where they drift until they strike a detector that records the TOF. 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 TOF. The spectral data that result from this experiment consist of the sequentially recorded numbers of ions (the intensities) arriving at the detector coupled with the corresponding $m/z$ values. Peaks in the intensity plot represent proteins that are present in the sample.

Feature extraction and quantification for these data involve a number of steps that interact in complex ways. Some elements of this process are elucidated by the following conceptual model. Suppose we observe $n$ spectra, each taken on the same equally-spaced grid of length $T$ of TOFs $t_j, j = 1, \ldots, T$. We model the log-transformed intensities, since we find that this transformation makes the data more symmetric and decouples the relationship between the mean and variance. A model for $y_i(t_j)$, the observed log spectral intensity for spectrum $i$ at TOF $t_j$, is

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

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 a peptide. The approximate shapes of peaks can be estimated empirically by simulating the physical process by which TOF mass spectrometers collect data (see Coombes *et al*., 2005a), 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 spectrum-specific variability, e.g. to adjust for differing amounts of protein ionized and desorbed from each slide. The baseline function, $B_i$, represents a systematic artifact commonly seen in mass spectrometry data. This artifact is believed to be attributable to a cloud of matrix molecules hitting the detector in the early part of the experiment, or to detector overload (Malyarenko *et al*., 2005). Our only characterization of this function is that it should be smooth. In this paper, we assume that the errors are mean-zero Gaussians with the variance a smooth function of $t$, i.e. $\epsilon_{ij} \sim N\{0, \sigma^2(t_j)\}$.

The following steps are necessary to perform feature extraction and quantification with MALDI data. (1) Calibration maps the observed TOFs $t_j, j = 1, \ldots, T$ to a set of inferred $(m/z)$ ratios $x_j, j = 1, \ldots, T$. 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, $\epsilon_{ij}$, 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 Equation (1). (5) Peak detection and quantification 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. If the peak detection is done on individual spectra, (6) then peak matching across the samples is necessary to decide which peaks in different samples correspond to the same biological molecule.

## 3 METHODS: FEATURE EXTRACTION AND QUANTIFICATION

We now describe two methods for performing feature extraction and quantification. Both methods are based upon the undecimated discrete wavelet transform (UDWT), and are motivated by the conceptual model (1) presented in Section 2. The first method operates on the individual spectra, while the second uses the average spectrum for peak detection.

### 3.1 Peak detection using UDWT on individual spectra (SUDWT)

The following steps are used to preprocess the spectra by applying the UDWT-based method introduced in Coombes *et al*. (2005b) to extract features from individual spectra. We will refer to this method as the SUDWT (single spectrum undecimated discrete wavelet transform)-based peak detection method.

(1) Ensure that the individual spectra are well calibrated. 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 a common time axis. If further calibration is necessary, methods such as those described in Eilers (2003, 2004) can be used to align the spectra.

(2) Denoise the individual spectra via wavelet regression using the UDWT. We use the implementation in version 2.4 of the Rice Wavelet Toolbox (RWT), which is available at 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 $\eta$ and a robust estimate of the noise, the median absolute deviation (MAD) divided by 0.67. Because most signals can be represented by a small number of wavelet coefficients, yet white noise is distributed equally among all wavelet coefficients; this approach denoises with minimal attenuation of the features of the signal. This approach is similar to other wavelet regression procedures developed in recent years. However, unlike the decimated discrete wavelet transform typically used in many of these methods, the overcomplete UDWT is translation-invariant, which leads to more effective denoising. We have found that the choice of wavelet basis does not strongly impact the denoising, although the choice of the thresholding parameter $\eta$ does. This denoising step partitions the raw spectrum into estimates of the denoised signal and a 'noise spectrum' containing the noise residuals, $\epsilon_{ij}$.

(3) Estimate the noise level across the spectrum using a median filter, i.e. by applying the MAD/0.67 estimate to the estimate of the noise in a sliding window.

(4) Estimate and remove the baseline artifact, $B_i(t)$ 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. After these first five steps, we are left with an estimate of the true signal, $S_i(t)$.

(6) Identify peaks on the denoised, baseline corrected and normalized spectrum. First, find all local maxima and the associated peak endpoints. Second, 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. Let $\phi$ be a S/N threshold. All local maxima with S/N $> \phi$ are considered peaks.

(7) Match peaks across the spectra. First, pool the list of detected peaks across the spectra, then combine the peaks that differ in location by no more than $\delta_t$ clock ticks or $\delta_m$ in relative mass. This results in a number of 'peak bins' defined across spectra. One may also specify a maximum total bin width in order to reduce the chance of nearby peaks being incorrectly coalesced into the same peak group. We label each unique peak group by the $m/z$ value at the midpoint of its peak bin.

(8) Quantify the peaks for each individual spectrum using the maximum log intensity within each peak group.

In Coombes *et al*. (2005b), this method was shown to perform very well on sample datasets when compared to other commonly-used peak detection methods (Fung and Enderwick, 2002; Yasui *et al*., 2003a,b).

### 3.2 Peak detection using the UDWT on the mean spectrum (MUDWT)

We now introduce an adaptation of this algorithm that uses the average spectrum for peak detection. We refer to this method as MUDWT (the mean-spectrum undecimated discrete wavelet transform)-based peak detection method.

(1) Ensure that the individual spectra are well calibrated and, if necessary, use interpolation to put all spectra on the same time scale.

(2) Compute the mean spectrum, averaging over all raw spectra.

(3) Apply steps 2–6 described in the SUDWT method 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. The left and right endpoints of the interval are determined by the $m/z$ values at the nearest local minima to the left and right, respectively, of the local maximum. If the local minima are not well defined because there are multiple adjacent measurements with identical minimum values, then the $m/z$ value of the identical value closest to the local maximum is used. Keep only those peaks above the S/N threshold $\phi$. The noise reduction inherent in the averaging should allow the use of a smaller S/N threshold than the SUDWT. We label the individual peaks by the $m/z$ value of the local maximum in the mean spectrum.

(4) Quantify the identified peaks in the individual spectra. First, denoise, baseline correct and normalize each individual spectrum using steps 1–5 described in the SUDWT method to get estimates of the true signals $S_i(t)$. We generally recommend choosing a smaller wavelet threshold parameter $\eta$ when quantifying than when detecting peaks to reduce bias in the quantifications. Next, quantify each peak using the maximum log intensity on the individual spectra within the interval defining the peak on the average spectrum. This approach allows the peak quantification to be robust to slight misalignments across spectra.

Note that the peak detection 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. The presence of baseline does not affect our ability to detect the 'bumps' in the mean spectrum corresponding to peaks.

The success of the proposed method depends to an extent 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 (Fig. 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 simply result in peaks in the mean spectrum that are somewhat broader than the peaks found in individual spectra.

## 4 PERFORMING SIMULATION STUDIES IN MASS SPECTROMETRY

Here we introduce a simulation-based approach for systematically comparing different methods for analyzing mass spectrometry data. Data are simulated from a 'virtual mass spectrometer', a computer-based model of a MALDI–TOF instrument with ion focus delay we have developed that is based on the physical principles

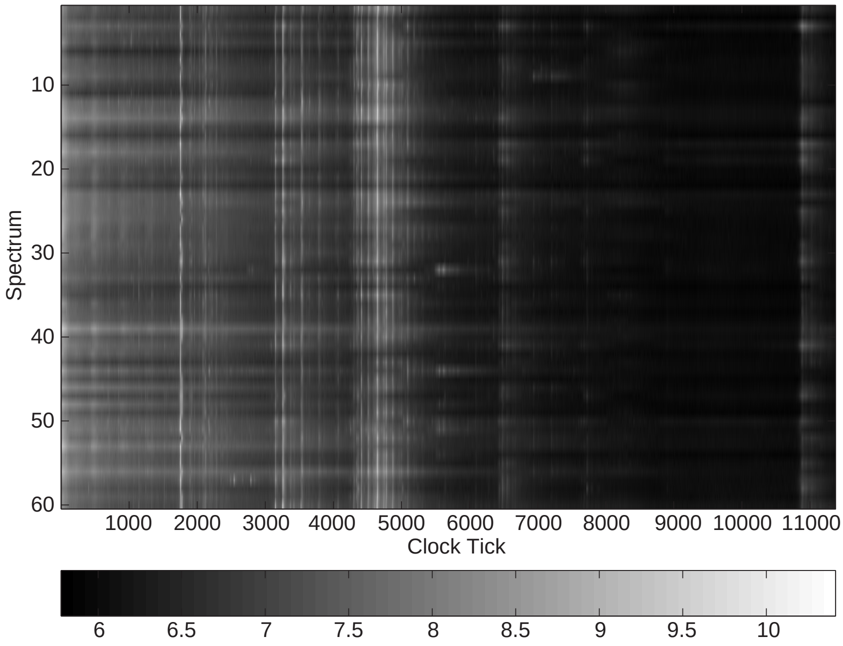

**Fig. 1.** Checking Calibration. Heat map of the logarithmic intensities of 60 spectra related to pancreatic cancer. Bright vertical lines are the largest peaks, which are well-aligned across spectra.

underlying the instrument (Coombes *et al.*, 2005a). We describe in general how to use this tool to perform simulation studies, and we set up a specific simulation study to compare the SUDWT and MUDWT methods described in Section 3 with respect to peak detection.

### 4.1 A virtual mass spectrometer

When given a virtual sample, the virtual MALDI–TOF instrument produces a virtual spectrum. The virtual sample consists of a list of the molecular masses and abundances for a set of proteins assumed to be present in the biological sample. The abundance of a protein is the assumed number of molecules of that protein that have been ionized and desorbed from the sample. Given the initial velocities at which the ions are desorbed from the plate, the virtual instrument simulates the actual physical process the ions undergo as they are 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, which records the number of ions striking it in a fixed time interval. The actual TOFs for each ion are computed using basic physics principles, then aggregated to form the virtual spectrum. Virtual calibration samples consisting of proteins on a grid of known masses are then run to obtain a mapping of TOF to molecular mass and to map abundances to expected peak intensities.

Many of the dials and settings on the virtual instrument are 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 $\delta$ and the time resolution of the detector $\tau$. 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 $\mu$ and variance $\tau^2$, 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 Gaussian noise to represent additive noise factors.

Mass spectrometry data are highly structured. There is a systematic relationship between the $m/z$ values and the peak characteristics—proteins at low $m/z$ values yield 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 isotopic distributions of the elements in the proteins, the distributions of the initial velocities of the ions as they are desorbed from the sample plate and the time resolution of the instrument's ion detector. Since our virtual mass spectrometer is based on the key physical principles underlying the instrument, our simulated spectra reflect these characteristics. Note, however, that our virtual mass spectrometer does not model the ionization and desorption processes and so the abundances for a peak correspond to the number of molecules of the corresponding protein successfully ionized and desorbed from the sample, and not to the number of molecules of that protein actually present in the sample.

While any virtual instrument is based on simplifying assumptions, we believe that this tool generates virtual spectra that have characteristics similar to the spectra emanating from a real MALDI–TOF instrument. Figure 2 contains a spectrum from a real MALDI–TOF instrument at MD Anderson Cancer Center, along with a virtual spectrum obtained from our tool with matching instrument settings.

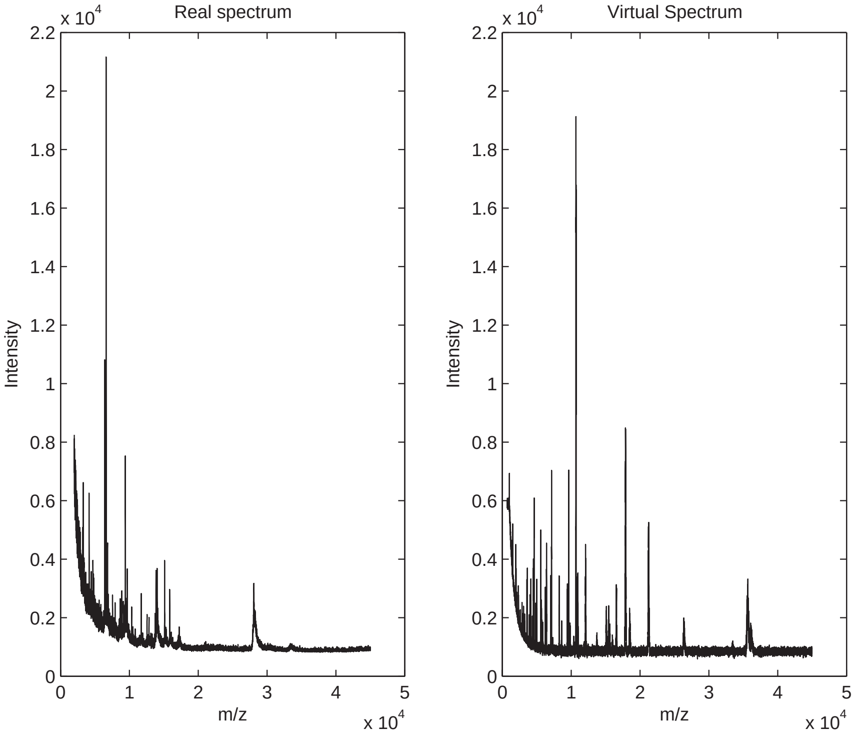

**Fig. 2.** Real and Virtual Spectra. Plot of a true MALDI–TOF spectrum and a virtual MALDI–TOF spectrum from our virtual mass spectrometer.

## 4.2 Virtual experiments

A typical MALDI–TOF experiment consists of taking samples from a biological medium of interest (e.g. blood serum) from $n$ individuals, spotting them on a plate and 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 samples will be drawn. This population consists of the list of all detectable $m/z$ values corresponding to 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 peaks present in the population. For a given peak $j$ of mass $x_j$, we summarize its distribution across samples by three quantities: $\pi_j$, its prevalence or the proportion of samples in the population containing the protein corresponding to this peak; $m_j$ and $s_j$, the mean and standard deviation log peak intensity across samples in the population that contain the corresponding protein.

A virtual MALDI–TOF experiment is then conducted by randomly generating $n$ samples from the virtual population then running these samples through the virtual mass spectrometer to obtain $n$ spectra. Specifically, for each sample and peak, we first determine whether the corresponding protein is present by drawing a random Bernoulli($\pi_j$), then if it is present, we then generate the expected log peak intensity $y_j$ by drawing a random Normal($m_j, s_j$). These log peak intensities must then be mapped to numbers of molecules using a formula obtained from the virtual calibration samples described in Section 4.1 before being 'fed' into the virtual instrument. This mapping is based on two empirical observations regarding the virtual mass spectrometer. First, given a constant number of molecules, the expected inverse intensity of a peak is linearly related to the $m/z$ value (Coombes *et al.*, 2004). Second, the intensity of a peak in the virtual MALDI–TOF instrument is linearly related to the abundance, i.e. the assumed number of molecules of the corresponding protein ionized and desorbed from the sample. The details of this mapping are available as supplementary material from the first author.

Statistical simulation studies can be performed by running a number of virtual experiments, then comparing the performances of the methods across these experiments. The known proteins and abundance distributions for each virtual population can serve as a gold standard against which to evaluate the methods. The virtual populations can be determined using real data, and the simulation study can be made more robust to the population characteristics by averaging results over multiple virtual populations.

## 4.3 Details of simulation study

We now describe the details of our simulation study to compare the performance of the SUDWT and MUDWT methods with respect to peak detection. We based our virtual populations on data from a pancreatic cancer study conducted at MD Anderson Cancer Center, consisting of MALDI–TOF spectra from the blood sera of 124 individuals, 83 with pancreatic cancer and 41 without.

We applied the SUDWT method to these data, computed the prevalence of each detected peak, then for each computed the mean and standard deviation log intensity across those samples for which it was detected. We fit a beta distribution to the prevalences of the peaks and a multivariate normal distribution to the vector $\{\log(x_j), m_j, s_j\}^{\mathrm{T}}$ across peaks. We found these distributions fit the data well, and estimated the parameters of the beta to be $(0.5, 0.5)$. For the normal distribution, we got a mean vector $(8.78, 9.34, 0.99)^{\mathrm{T}}$ and covariance matrix $\Sigma$, with diagonal elements 0.536, 0.503 and 0.156, and off diagonal elements $-0.108, 0.104$ and 0.057, respectively.

We used these distributions to generate 100 virtual populations, each containing 150 true protein peaks, for each simulation scenario. Each peak's true mass $x_j$, prevalence $\pi_j$ and mean and standard deviation $m_j$ and $s_j$ were obtained by sampling from the distributions described above. For each virtual population, we ran one virtual experiment by taking $n$ samples from the population and obtaining the corresponding spectra. The virtual instrument's settings were made to match the settings on a MALDI–TOF instrument at MD Anderson Cancer Center. The additive noise was assumed to be stationary with variance $\sigma^2$. We obtained nearly identical results when allowing the variance to vary over $t$. Our use of multiple virtual populations ensured that our simulation study averaged over the different characteristics that can be present in a given population.

We ran five simulation scenarios. The machine noise level for the first was chosen to be comparable to the pancreatic dataset ($\sigma = 66$) 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.

We applied the SUDWT and MUDWT methods to the spectra from each virtual experiment and obtained a list of found peak locations $\{x^*_{S,j}\}, j = 1, \dots, p^*_S$ and $\{x^*_{M,j}\}, j = 1, \dots, p^*_M$, with $p^*_S$ and $p^*_M$ being the number of peaks found by the two methods. 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. Since we found that the results were sensitive to the S/N threshold, 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 get a set of candidate thresholds.

### 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, $x^*_{S,j}; j = 1, \dots, p^*_S$ and $x^*_{M,j}; j = 1, \dots, p^*_M$, with the true locations of the protein peaks in the virtual population, $x_j; j = 1, \dots, p$. We treated peak detection as a special type of classification problem, since each $m/z$ value on the spectrum had a true state (peak or not), and the peak detection methods classified each $m/z$ value into one of the two states (peak or not). In order to devise an automatic method for summarizing the results that took into account the continuous nature of the $m/z$ values $x$, 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.

For each simulation, we summarized the performance of the peak detection by four measures, the sensitivity, the false discovery rate (FDR), $MM_1$ and $MM_2$. The sensitivity is the proportion of true peaks matching at least one found peak, while the FDR is the proportion of found peaks not matching any true peak. $MM_1$ summarizes the proportion of found peaks matching multiple true peaks and $MM_2$ summarizes the proportion of true peaks matching multiple found peaks.

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 dataset 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.

## 5 RESULTS

We have summarized the performance of our method, demonstrating the advantages of using the mean spectrum for peak detection, first through examples and then through simulation studies.

### 5.1 Examples

*The noise in the mean spectrum decreases by $\sqrt{n}$.* The first dataset was described in our previous paper (Coombes *et al*., 2005a). It consists of 24 spectra acquired from the same pooled sample of nipple aspirate fluid. Figure 3a 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 five. 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 (Fig. 3b).

*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 the spectra. In the analysis of this dataset reported in our previous paper using the SUDWT method, we found 174 sets of matched peaks (Coombes *et al*., 2005b). 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.

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.

*Small, 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 to be a spurious feature. By contrast, a large bump that occurs even in 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.

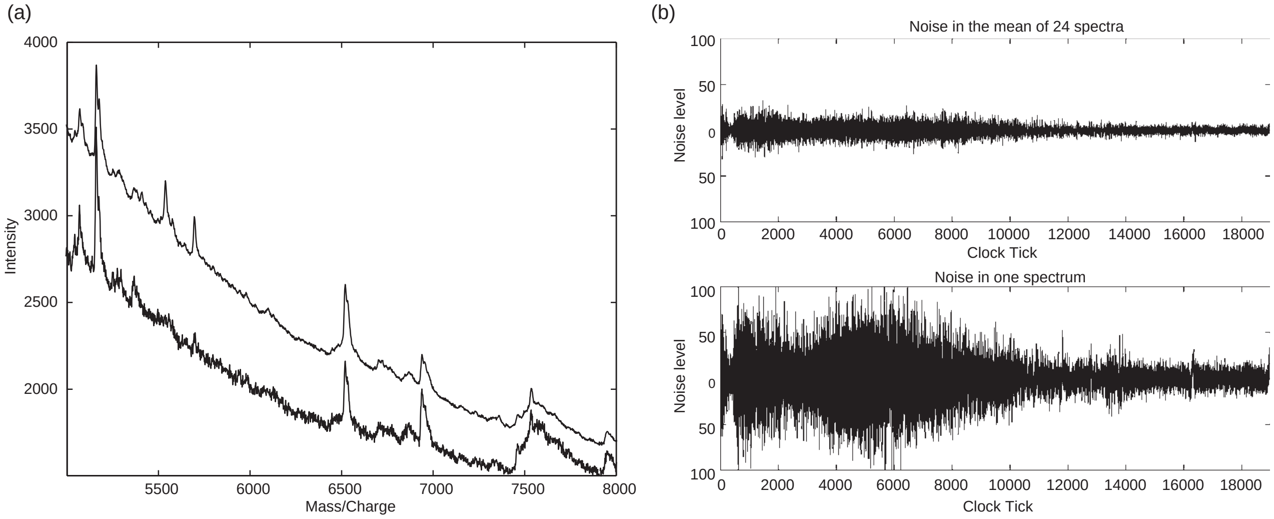

**Fig. 3.** Noise Reduction in the Mean Spectrum. Plots of (**a**) one individual raw spectrum (lower curve) and the mean of 24 replicate spectra (upper curve), and (**b**) the estimated noise removed by the UDWT in the mean of 24 spectra (top) and in one individual spectrum (bottom). The noise is reduced in the mean spectrum by a factor of about 5.

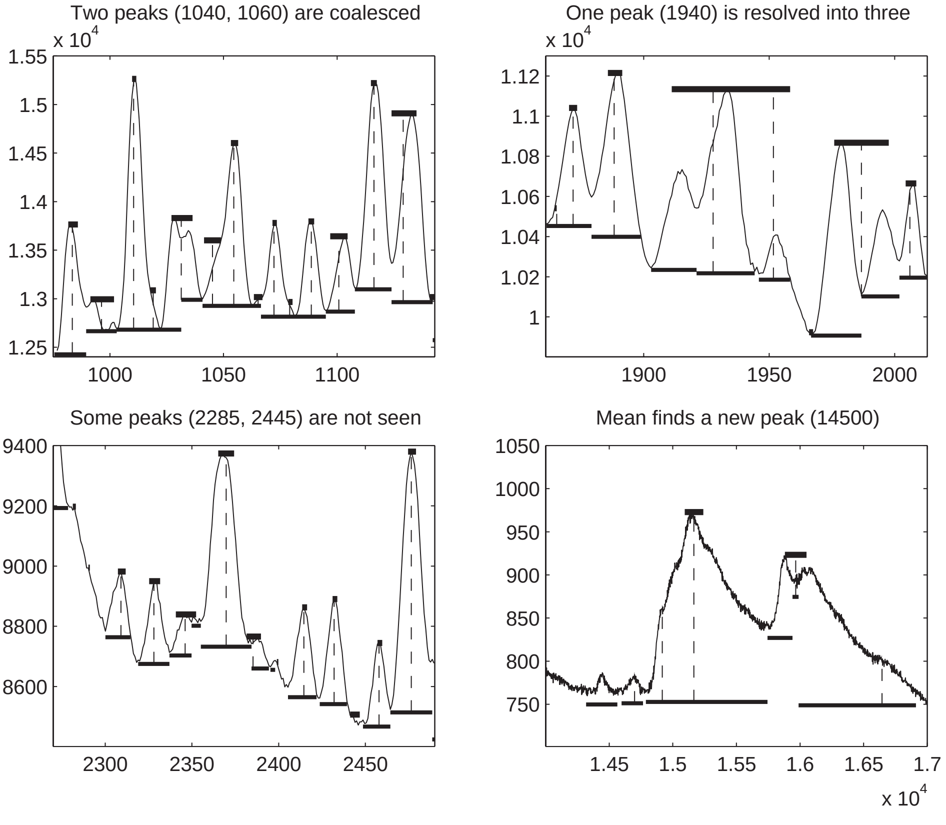

**Fig. 4.** Comparing Peak Detectors. Plots of the mean spectrum illustrating differences in peak finding methods. Bars above peaks indicate the peak bins, defining regions where peaks in individual spectra were matched. Bars below peaks indicate the width of peaks found in the mean spectrum. Dotted vertical lines join peaks found by both methods. In the upper left quadrant, two separate peak bins from the SUDWT at 1040 and 1060 Da show up as a single peak when using the average spectrum. In the upper right, three separate peaks found by the MUDWT at 1920, 1935 and 1950 Da are combined into the same peak bin by the SUDWT. In the lower left, two peaks (2285 and 2445 Da) are found by the SUDWT but not the MUDWT. In the lower right, the MUDWT finds a peak at 14 500 Da that is not found by the SUDWT.

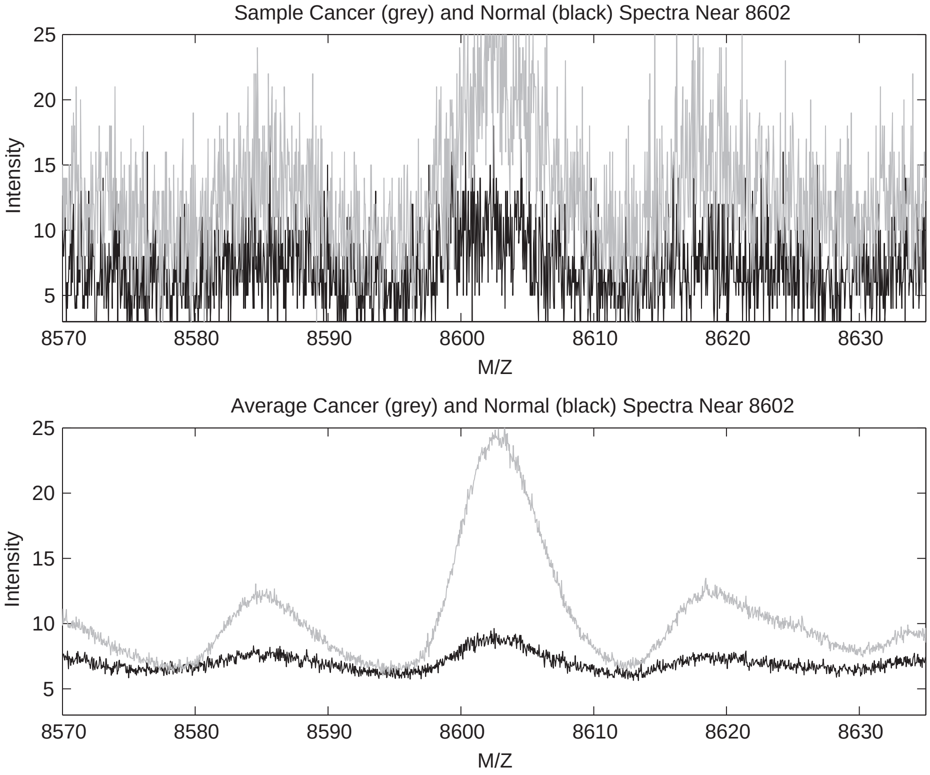

**Fig. 5.** Short Peak. Plots showing an arbitrarily chosen normal and cancer spectrum (top) and the mean spectra across 95 normal spectra and 121 ovarian cancer spectra (bottom) in the neighborhood of a significant peak at 8602 Da. This represents a protein that is more abundant in cancer patients. 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.

To illustrate this idea, we considered a publicly available dataset described in Conrads *et al*. (2004), and available at http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

This dataset contains spectra from serum samples of 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 TOF 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 Da that appeared to be more abundant in ovarian cancer patients than in healthy women. They also pointed out that this peak never achieved a (S/N) of greater than $\sim$1.5, which meant that it would be filtered out by most peak finding algorithms applied to the individual spectra. The peak clearly stands out in the mean spectrum (Fig. 5). Given our earlier observation about the noise levels, we would expect S/N to $\simeq 1.5 * \sqrt{100} \approx 15$ in the mean of either group of samples, making it easy to find.

*The mean spectrum can find peaks 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 present only in a small number of cancer samples (Coombes *et al*., 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 MD Anderson Cancer Center. The dataset 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 Da. These peaks are barely discernible in the 90th percentile spectrum, but would be clearly detected on the mean spectrum.

### 5.2 Simulation results

We ran each simulation using various choices for the S/N threshold $\phi$. Table 1 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 performance of both methods was sensitive to the choice of $\phi$. For the other simulation scenarios, we only report the results for a single choice of $\phi$, chosen as follows. We first chose $\phi$ 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 $\phi$ 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 $\phi$ used for the SUDWT ranged from 15 to 40, while the $\phi$ used for the MUDWT ranged from 2.82 to 4.

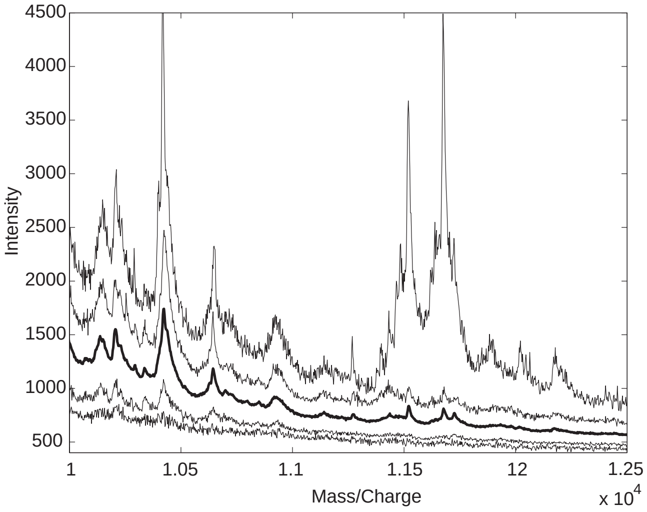

**Fig. 6.** Rare Peak. Plots of spectra obtained by pointwise application of statistical functions to 60 spectra from a study of pancreatic cancer. From top to bottom, the spectra are the maximum, 90th percentile, mean, 10th percentile and minimum. Note the occurrence of large peaks in the maximum at 11 500 and 11 600 Da that are barely present in the 90th percentile; these peaks are visible in the mean spectrum. This demonstrates that the mean spectrum can be effective for identifying peaks even when the distribution of intensities across the individual spectra is highly skewed.

**Table 1.** Mean sensitivity and FDR across 100 virtual populations for different S/N thresholds, $n = 200$, $\sigma = 66$ simulations

| 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 2 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 slightly lower for the MUDWT method in most cases, by design. The multiple match proportions ($MM_1$ and $MM_2$, not shown) were comparable between the two methods. 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 $\sigma = 66$.

Table 3 contains the sensitivities for the peaks sorted into different prevalence and abundance 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 <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.

**Table 2.** Overall results from the simulation study

| Settings | Method | Sensitivity | FDR |
|---|---|---|---|
| $n = 100$ | SUDWT | 0.75 (0.60, 0.85) | 0.09 (0.02, 0.26) |
| $\sigma = 66$ | MUDWT | 0.83 (0.75, 0.92) | 0.06 (0.00, 0.41) |
|  | Comparison | 0.97 | 0.80 |
| $n = 100$ | SUDWT | 0.58 (0.43, 0.69) | 0.25 (0.11, 0.41) |
| $\sigma = 22$ | MUDWT | 0.74 (0.61, 0.84) | 0.23 (0.10, 0.52) |
|  | Comparison | 1.00 | 0.63 |
| $n = 100$ | SUDWT | 0.70 (0.61, 0.80) | 0.08 (0.00, 0.17) |
| $\sigma = 200$ | MUDWT | 0.78 (0.69, 0.87) | 0.05 (0.00, 0.45) |
|  | Comparison | 0.97 | 0.86 |
| $n = 33$ | SUDWT | 0.73 (0.63, 0.84) | 0.09 (0.01, 0.20) |
| $\sigma = 66$ | MUDWT | 0.80 (0.74, 0.86) | 0.06 (0.00, 0.36) |
|  | Comparison | 0.99 | 0.85 |
| $n = 200$ | SUDWT | 0.75 (0.58, 0.87) | 0.12 (0.02, 0.46) |
| $\sigma = 66$ | MUDWT | 0.85 (0.75, 0.91) | 0.11 (0.00, 0.31) |
|  | Comparison | 1.00 | 0.69 |

The top element in each box is the mean quantity over the 100 virtual experiments, and the bottom interval is the range. The comparison proportion $p$ measures 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.

Abundance groups were 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 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 less difference between the SUDWT and MUDWT for the most abundant proteins, which were quite easily detected by both methods.

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 (Table 4). 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 in 86 of the 100 virtual populations; the SUDWT achieved

**Table 3.** Sensitivity by prevalence and abundance groups

| Settings | Method | Prevalence ($\pi$) | | | | Mean Log Intensity ($m$) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | <0.05 | 0.05–0.20 | 0.20–0.80 | >0.80 | <9.0 | 9.0–9.5 | 9.5–10 | >10 |
| | | (14%) | (16%) | (40%) | (30%) | (31%) | (27%) | (23%) | (19%) |
| $n = 100$ | SUDWT | 0.43 | 0.74 | 0.81 | 0.82 | 0.68 | 0.75 | 0.78 | 0.82 |
| $\sigma = 66$ | | (0.20, 0.75) | (0.50, 0.96) | (0.60, 0.93) | (0.59, 0.95) | (0.48, 0.90) | (0.53, 0.95) | (0.51, 0.94) | (0.56, 1.00) |
| | MUDWT | 0.38 | 0.74 | 0.93 | 0.97 | 0.78 | 0.84 | 0.85 | 0.88 |
| | | (0.16, 0.67) | (0.54, 0.95) | (0.78, 1.00) | (0.89, 1.00) | (0.60, 0.91) | (0.68, 0.97) | (0.68, 0.97) | (0.70, 1.00) |
| | $p$ | 0.25 | 0.49 | 1.00 | 0.99 | 0.97 | 0.89 | 0.84 | 0.78 |
| $n = 100$ | SUDWT | 0.39 | 0.62 | 0.62 | 0.60 | 0.56 | 0.58 | 0.61 | 0.61 |
| $\sigma = 22$ | | (0.09, 0.67) | (0.08, 0.85) | (0.39, 0.83) | (0.41, 0.88) | (0.37, 0.76) | (0.30, 0.89) | (0.38, 0.81) | (0.36, 0.94) |
| | MUDWT | 0.39 | 0.66 | 0.81 | 0.84 | 0.70 | 0.73 | 0.75 | 0.78 |
| | | (0.12, 0.63) | (0.42, 0.88) | (0.62, 0.94) | (0.66, 0.97) | (0.53, 0.85) | (0.50, 0.86) | (0.50, 0.86) | (0.56, 0.91) |
| | $p$ | 0.51 | 0.64 | 1.00 | 1.00 | 0.99 | 0.96 | 0.93 | 0.96 |
| $n = 100$ | SUDWT | 0.33 | 0.59 | 0.78 | 0.82 | 0.58 | 0.69 | 0.75 | 0.83 |
| $\sigma = 200$ | | (0.00, 0.65) | (0.35, 0.84) | (0.63, 0.92) | (0.67, 0.98) | (0.40, 0.75) | (0.45, 0.86) | (0.52, 0.95) | (0.61, 0.97) |
| | MUDWT | 0.30 | 0.57 | 0.89 | 0.96 | 0.73 | 0.77 | 0.80 | 0.84 |
| | | (0.07, 0.68) | (0.29, 0.84) | (0.73, 0.96) | (0.87, 1.00) | (0.55, 0.88) | (0.61, 0.91) | (0.63, 0.95) | (0.56, 1.00) |
| | $p$ | 0.40 | 0.40 | 0.95 | 1.00 | 0.98 | 0.86 | 0.78 | 0.54 |
| $n = 33$ | SUDWT | 0.32 | 0.60 | 0.82 | 0.86 | 0.63 | 0.73 | 0.78 | 0.84 |
| $\sigma = 66$ | | (0.11, 0.61) | (0.38, 0.80) | (0.66, 0.94) | (0.73, 0.98) | (0.47, 0.77) | (0.47, 0.86) | (0.56, 0.93) | (0.67, 1.00) |
| | MUDWT | 0.32 | 0.62 | 0.91 | 0.98 | 0.75 | 0.80 | 0.83 | 0.85 |
| | | (0.11, 0.57) | (0.33, 0.83) | (0.80, 1.00) | (0.90, 1.00) | (0.57, 0.88) | (0.58, 0.93) | (0.70, 0.95) | (0.68, 1.00) |
| | $p$ | 0.51 | 0.54 | 0.95 | 1.00 | 0.96 | 0.86 | 0.77 | 0.59 |
| $n = 200$ | SUDWT | 0.48 | 0.76 | 0.79 | 0.80 | 0.69 | 0.73 | 0.78 | 0.82 |
| $\sigma = 66$ | | (0.16, 0.78) | (0.50, 1.00) | (0.62, 0.93) | (0.55, 0.97) | (0.49, 0.84) | (0.52, 0.91) | (0.54, 0.94) | (0.48, 1.00) |
| | MUDWT | 0.44 | 0.81 | 0.92 | 0.97 | 0.81 | 0.85 | 0.87 | 0.90 |
| | | (0.19, 0.78) | (0.54, 0.96) | (0.85, 1.00) | (0.89, 1.00) | (0.71, 0.96) | (0.69, 0.97) | (0.68, 0.97) | (0.71, 1.00) |
| | $p$ | 0.38 | 0.70 | 0.97 | 0.98 | 0.98 | 0.89 | 0.86 | 0.79 |

The sensitivities for peaks in different prevalence and abundance 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. The comparison proportion $p$ measures 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.

**Table 4.** Interaction of prevalence and abundance

| Prevalence | Abundance (mean $\log_2$ intensity) | | | |
|---|---|---|---|---|
| ($p$) | <9.0 | 9.0–9.5 | 9.5–10 | >10 |
| <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 |

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.

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 7 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 method.

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.

## 6 DISCUSSION

The two-step analysis approach we have discussed in this paper has the advantage of reducing the dimensionality of the data in a scientifically meaningful way, since the peaks represent the proteins, the scientific units of interest in the data. Because subsequent analyses are performed only on the detected peaks, it is crucial to use effective methods for the first step of this approach, feature extraction and

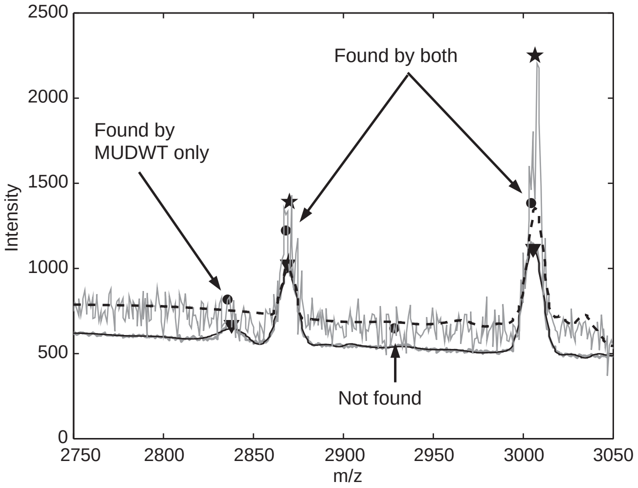

**Fig. 7.** Some Peaks. Four lines from one of the datasets from the simulation with $n = 100$ and $\sigma = 66$ are plotted here: the two 'noisy' grey lines are selected raw spectrum and the mean spectrum. The dashed line is the wavelet-smoothed version of the individual spectrum, and the solid line 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.

quantification. In this paper, we have introduced a feature extraction and quantification method (MUDWT) which appears to work very well in simulated and real data examples.

Averaging is a fundamental principle underlying many statistical methods. We put this simple idea to work in order to improve peak detection for mass spectrometry data, which to our knowledge has not been done in existing literature or software. 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, specifically when the prevalence is $<1/\sqrt{n}$, although our simulation studies suggest that this difference is small compared to the improvements seen elsewhere.

Another advantage of the MUDWT method is that peaks can be detected and quantified without having to apply peak-matching algorithms across samples. This step is necessary for any method whereby peak detection is performed on the individual spectra, since these individual peaks must somehow be combined together across spectra. This process is difficult and can lead to various errors, including combining together of adjacent peaks corresponding to different proteins into the same peak bins, as well as forming different peak bins for peaks that correspond to the same molecule, but differ slightly in $m/z$ across samples (Fig. 4). Careful choice of the tolerance parameters, $\delta_t$ and $\delta_m$, and limits on the bin widths can decrease these problems, but they cannot eliminate them. With the MUDWT, this procedure is avoided altogether because peaks are unambiguously defined across samples using the local maxima in the mean spectrum. Adjacent peaks will not be coalesced as long as they have their own 'bumps' in the average spectrum, and slight variability in the peak locations across spectra can still be accommodated by allowing the peak location in the individual spectra to vary within the interval determined by the flanking local minima in the average spectrum.

In order to implement our MUDWT method, two parameters must be set, the wavelet threshold $\eta$ and the S/N threshold $\phi$. 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 datasets. For $\eta$, 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 $\phi$ is made much smaller than 4, we have found in our simulations that the FDR 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.

We saw evidence of undersmoothing with the MUDWT when the noise level was low. To combat this problem, we suggest combining together peaks that are highly correlated with each other (e.g. $r > 0.95$) 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 will tend to combine information across peaks that emanate from the same protein, including doubly-charged ions, matrix adducts, different isotopes and other alterations that are not expected to be biologically meaningful. This further reduces the dimensionality of the data 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 think that this practice will not allow 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 for the method based on the UDWT algorithm introduced by Coombes *et al*. (2005b). 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 applying other peak detection methods to the average spectrum instead of the individual spectra. Also, while this paper focused specifically on mass spectrometry data, our mean function, UDWT-based peak detection procedure could be adapted to perform feature extraction for other types of functional data for which the features of interest are peaks.

We have also described how to design simulation studies based on the virtual mass spectrometer to compare competing methods. While applied in this paper to assess peak detection, the same procedure can be used to assess other types of performance, e.g. comparison of different baseline correction or normalization procedures or methods for detecting differentially expressed proteins. It can also be used to perform power calculations when designing studies. Since the virtual instrument is based upon the key physical principles underlying the real instrument, it yields data whose characteristics resemble actual mass spectra.

There is still more work to do to improve the virtual instrument, however. Common alterations of proteins, such as matrix and sodium adducts or neutral losses of water, ammonia or carbon, should also be incorportated 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. Malyarenko *et al*. (2005) have some insights that may be useful in that regard. The virtual instrument may also benefit by making the stochastic processes modeling the ionization/desorption processes and the ion detection more directly related to the science of the instrument.

## ACKNOWLEDGEMENTS

We thank Henry Kuerer and Mien-Chie Hung for the permission to use the nipple aspirate fluid SELDI dataset; I.J. Fidler, Stanley Hamilton, James Abbruzzesse, Donghui Li and Nancy Shih for the permission to use the pancreatic cancer MALDI–TOF dataset; and Drs Emmanuel Petricoin and Lance Liotta for making their QStar data publicly available on their website. This research was supported by NIH/NCI grants R01 CA-107304 and P50 CA070907.

## REFERENCES

Adam,B.L., Qu,Y., Davis,J.W., Ward,M.D., Clements,M.A., Cazares,L.H., Semmes,O.J., Schellhammer,P.F., Yasui,Y., Feng,Z. and Wright,G.L.Jr. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, **62**, 3609–3614.

Adam,P.J., Boyd,R., Tyson,K.L., Fletcher,G.C., Stamps,A., Hudson,L., Poyser,H.R., Redpath,N., Griffiths,M., Steers,G., *et al*. (2003) Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. *J. Biol. Chem.*, **278**, 6482–6489.

Baggerly,K.A., Morris,J.S., Wang,J., Gold,D., Xiao,L.C. and Coombes,K.R. (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,K.A., Morris,J.S. and Coombes,K.R. (2004) Reproducibility of SELDI–TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, **20**, 777–785.

Beavis,R.C. and Chait,B.T. (1991) Velocity distributions of intact high mass polypeptide molecule ions produced by matrix assisted laser desorption. *Chem. Phys. Lett.*, **181**, 479–484.

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, pp. 545–548.

Conrads,T.P., Fusaro,V.A., Ross,S., Johann,D., Rajapakse,V., Hitt,B.A., Steinberg,S.M., Kohn,E.C., Fishman,D.A., Whitely,G. *et al*. (2004) High-resolution serum proteomic features for ovarian cancer detection. *Endocr. Relat. Cancer*, **11**, 163–178.

Coombes,K.R., Fritsche,H.A.Jr., Clarke,C., Chen,J.N., Baggerly,K.A., Morris,J.S., Xiao,L.C., Hung,M.C. and Kuerer,H.M. (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.

Coombes,K.R., Wang,J. and Baggerly,K.A. (2004) A statistical method for finding biomarkers from microarray data, with application to prostate cancer. *M.D. Anderson Biostatistics Technical Report UTMDABTR-007-04*, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Coombes,K.R., Kooman,J.M., Baggerly,K.A., Morris,J.S. and Kobayashi,R. (2005a) Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics*, in press.

Coombes,K.R., Tsavachidis,S., Morris,J.S., Baggerly,K.A., Hung,M.C. and Kuerer,H.M. (2005b) 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. *Proteomics*, in press.

Eilers,P.H. (2003) A perfect smoother. *Anal. Chem.*, **75**, 3631–3636.

Eilers,P.H. (2004) Parametric time warping. *Anal. Chem.*, **76**, 404–411.

Fung,E.T. and Enderwick,C. (2002) ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques* (Suppl.) S34–S41.

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. *J. Mass Spectrom.*, **34**, 467–477.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning*. Springer, New York.

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. *Int. J. Mass Spectrom.*, **226**, 239–248.

Malyarenko,D.I., Cooke,W.E., Adam,B.L., Gunjan,M., Chen,H., Tracy,E.R., Trosset,M.W., Sasinowski,M., Semmes,O.J. and Manos,D.M. (2005) Enhancement of sensitivity and resolution of surface-enhanced lazer desorption/ionization time-of-flight mass spectrometric records for serum peptides using time series analysis techniques. *Clin. Chem.*, **51**, 65–74.

Morris,J.S. and Carroll,R.J. (2004) Wavelet-based functional mixed models. *M.D. Anderson Biostatistics Technical Report, UTMDABTR-006-04*, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Paweletz,C.P., Gillespie,J.W., Ornstein,D.K., Simone,N.L., Brown,M.R., Cole,K.A., Wang,Q.H., Huang,J., Hu,N., Yip,T.T. *et al*. (2000) Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Dev. Res.*, **49**, 34–42.

Paweletz,C.P., Trock,B., Pennanen,M., Tsangaris,T., Magnant,C., Liotta,L.A. and Petricoin,E.F., III (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**, 301–307.

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 Int.*, **65**, 323–332.

Sorace,J.M. 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,Z.L., 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. *Int. J. Mol. Med.*, **9**, 341–347.

Yasui,Y., McLerran,D., Adam,B.L., Winget,M., Thornquist,M. and Feng,Z. (2003a) An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Biomed. Biotechnol.*, **2003**, 242–248.

Yasui,Y., Pepe,M., Thompson,M.L., Adam,B.L., Wright,G.L.Jr., Qu,Y., Potter,J.D., 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,T.A., Johanson,R.A., Cantor,A.B., Clark,R.A. and Tockman,M.S. (2003) Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer*, **40**, 267–279.