Biometrics 64, 479–489
June 2008

DOI: 10.1111/j.1541-0420.2007.00895.x

Bayesian Analysis of Mass Spectrometry Proteomic Data Using Wavelet-Based Functional Mixed Models

Jeffrey S. Morris,[1,*] Philip J. Brown,[2] Richard C. Herrick,[1]
Keith A. Baggerly,[1] and Kevin R. Coombes[1]

[1]The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030-4009, U.S.A.
[2]The University of Kent, Canterbury CT2 7NZ, U.K.
[*]*email:* jeffmo@mdanderson.org

Summary. In this article, we apply the recently developed Bayesian wavelet-based functional mixed model methodology to analyze MALDI-TOF mass spectrometry proteomic data. By modeling mass spectra as functions, this approach avoids reliance on peak detection methods. The flexibility of this framework in modeling nonparametric fixed and random effect functions enables it to model the effects of multiple factors simultaneously, allowing one to perform inference on multiple factors of interest using the same model fit, while adjusting for clinical or experimental covariates that may affect both the intensities and locations of peaks in the spectra. For example, this provides a straightforward way to account for systematic block and batch effects that characterize these data. From the model output, we identify spectral regions that are differentially expressed across experimental conditions, in a way that takes both statistical and clinical significance into account and controls the Bayesian false discovery rate to a prespecified level. We apply this method to two cancer studies.

Key words: Bayesian analysis; False discovery rate; Functional data analysis; Functional mixed models; Mass spectrometry; Proteomics.

1. Introduction

Proteomic methods simultaneously detect and measure the expression of hundreds or thousands of proteins present in a biological sample, and are gaining increased attention in biomedical research. One popular proteomic method is matrix-assisted laser desorption and ionization, time-of-flight mass spectrometry (MALDI-TOF).

In a MALDI-TOF experiment, a biological sample of interest is first mixed with an energy-absorbing matrix substance, and the mixture is placed on a steel plate. A commonly used variant of MALDI-TOF, called surface-enhanced laser desorption and ionization (SELDI-TOF), incorporates additional chemistry on the surface of the metal plate to bind specific classes of proteins. The plate is then placed into a vacuum chamber, where a laser strikes the plate, desorbing ionized peptides from the sample. An electric field accelerates the particles into a potential free flight tube through which they travel at a constant velocity until striking a detector plate.

The detector plate records the abundance of particles striking it over a series of short, fixed intervals of time indexed by $\mathbf{t} = (t_1, \ldots, t_T)$, yielding the proteomic spectrum $y(t)$. Using basic physics principles, a quadratic transformation can be used to map the time axis t to a set of corresponding mass-to-charge ratios (m/z) x. Each spectrum is characterized by numerous peaks, which correspond to proteins or protein fragments (polypeptides) present in the sample. Depending on the proteomic makeup of the sample, some peptides present may fail to manifest as peaks if they are located on the shoulder of a more abundant peak (see Supplementary Figure 1 for an illustration of this phenomenon). Because most ions have equal charges (+1), the value of spectrum $y(x)$ at a peak is a rough measure of the abundance of some molecule in the sample having a molecular mass of x Daltons. The first column of Figure 1 contains two raw spectra from a MALDI-TOF instrument. In this article, we consider two example data sets from cancer studies conducted at The University of Texas M.D. Anderson Cancer Center.

Pancreatic cancer experiment: In this study, blood serum was taken from 139 pancreatic cancer patients and 117 healthy controls. The blood serum was fractionated using 25% acetonitrile elutions optimized using myoglobin, then run on a MALDI-TOF instrument to obtain a proteomic spectrum for each sample. For this analysis, we consider the region of the spectra between $x = 4{,}000$ and 40,000 Daltons, containing 12,096 observations per spectrum. These 256 samples were run in four different batches over a period of several months. More specifics of the experiment can be found in Koomen et al. (2005). Our primary goal is to identify regions of the spectra that are differentially expressed between pancreatic cancer patients and healthy controls, regions corresponding to proteins that may serve as blood serum biomarkers of pancreatic cancer.

Some recent case studies (Baggerly et al., 2003; Sorace and Zhan, 2003; Baggerly, Morris, and Coombes, 2004; Conrads

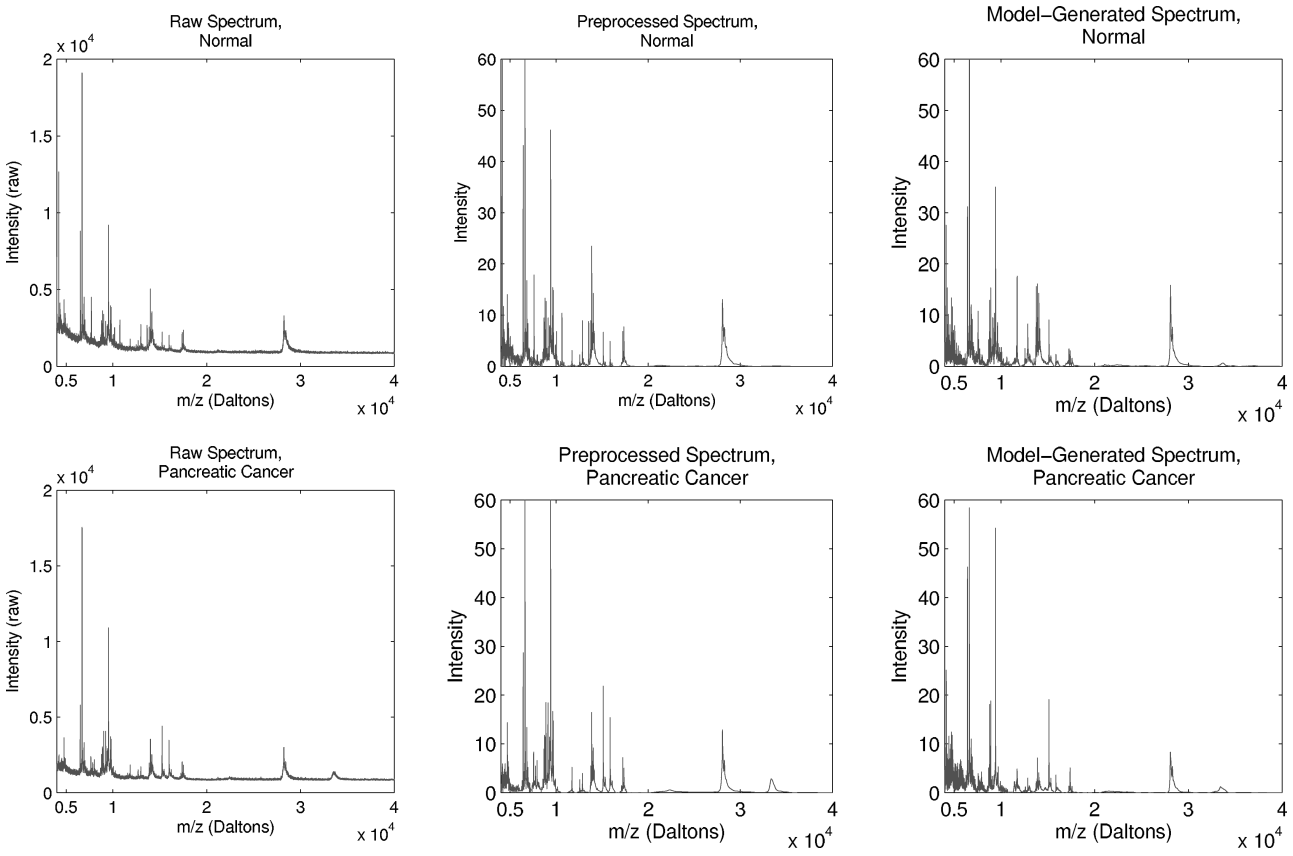

Figure 1. Sample spectra. The first column contains raw MALDI-TOF spectra from normal and pancreatic cancer patients, respectively, from the example data set. The second column shows the same spectra after preprocessing by baseline correction, normalization, and denoising. The final column contains normal and pancreatic cancer spectra randomly drawn from the posterior predictive distribution based on fitting the wavelet-based functional mixed model to the example data set. Note that the model does a good job of generating MALDI-TOF-like functions. This figure appears in color in the electronic version of this article.

and Veenstra, 2005; Coombes, Morris, 2005; Hu et al., 2005; Villanueva et al., 2005) have demonstrated that MALDI-TOF instruments can be very sensitive to experimental conditions, even varying over time within the same laboratory. These differences can manifest in systematic changes in both the intensities and locations of the peaks (i.e., both the y and x axes), and are sometimes larger in magnitude than the biological effects of interest. Thus, it is important for us to adequately model these block or batch effects if we are to properly analyze these data.

Organ-by-cell line experiment: In this study, a tumor from one of two cancer cell lines was implanted into either the brain or lungs of 16 mice. The cell lines were A375P, a human melanoma cancer cell line with low metastatic potential, and PC3MM2, a highly metastatic human prostate cancer cell line. After a period of time, blood serum was extracted and then placed on a SELDI chip. This chip was run on the SELDI-TOF instrument twice, once using a low laser intensity and the other using a high laser intensity. This resulted in a total of 32 spectra, two per mouse. Here, we considered the part of the spectrum between $x = 2{,}000$ and 14,000 Daltons, a range that included 7,985 observations per spectrum.

Our primary goals are to assess whether differential protein expression, if present, is more tightly coupled to the host organ site or to the donor cell-line type, and to identify regions of the spectra differentially expressed by organ site, by cell line, and/or their interaction. Typically, spectra from different laser intensities are analyzed separately, which is inefficient because spectra from both laser intensities contain information on the same proteins. We want to perform these analyses combining information across the two laser intensities, requiring us to model an effect of laser intensity on both the location (x axis) and intensity (y axis) of the peaks, and to account for correlation between spectra obtained from the same mouse.

It is common to use a two-step approach to analyze mass spectrometry data (Baggerly et al., 2003; Yasui et al., 2003; Coombes et al., 2003; Coombes, Tsavachidis, et al., 2005; Morris et al., 2005; Randolph et al., 2005). First, some type of feature detection algorithm is applied to identify *peaks* in the spectra. A quantification is then obtained for each peak and each spectrum, for example, by taking the intensity at a local maximum or computing the area under the peak. Assuming there are p peaks and N spectra, this results in a $p \times N$ matrix of *protein expression levels* that is somewhat

analogous to the matrix of mRNA expression levels obtained after preprocessing microarray data. Second, this matrix is analyzed using methods similar to those used for microarrays to identify peaks differentially expressed across experimental conditions, while controlling the false discovery rate (FDR).

This two-step approach is intuitive because it focuses on the peaks, which are theoretically the most scientifically relevant features of the spectra, and convenient, because it can borrow from a wide array of available methods developed for microarrays. However, it also has disadvantages. First, important information can be lost in the reduction from the full spectrum to the set of detected peaks. Because group comparisons are only performed after peak detection, this approach will miss important differences in low-intensity peaks or on shoulders of peaks whenever the peak detection algorithm fails to detect them. Second, this approach affords no natural way to account for experimental effects that impact both the x and y axes of the spectra, such as block or batch effects.

An alternative to the two-step approach described above is to model the spectra as functions, in the spirit of functional data analysis (Ramsay and Silverman, 1997). D. Billheimer (unpublished manuscript) took this approach, and this is the approach we take in this article. Mass spectra are irregular functions with many peaks, and so require flexible modeling and spatially adaptive regularization to represent accurately. The wavelet-based functional mixed model introduced by Morris and Carroll (2006) possesses these properties, and in this article we use this methodology to model mass spectrometry data. In modeling the entire spectrum, this method has the potential to identify differences at locations missed by peak detection algorithms. Further, the method's flexible nonparametric representation of the fixed and random effects allows it to model the functional effects of a number of factors simultaneously, including factors of interest as well as nuisance factors related to the experimental design. As we will demonstrate, these nonparametrically modeled effects can account for differences on both the x and y axes of the spectra, allowing data to be combined across laser intensities, blocks, or other experimental factors. The output of the method can be used to identify regions of interest within the spectra in a way that takes both statistical and practical significance into account, while controlling the Bayesian FDR at a specified level.

Although the primary goal of this article is to apply an existing method to the setting of mass spectrometry, we also present some new methodological advances not found in Morris and Carroll (2006). First, we describe a systematic method for selecting an additive shrinkage constant to apply before log transformation in a way that controls the bias for fold-change estimates at peak intensities of a specified size. Second, we introduce a method for identifying regions of the spectra that are differentially expressed in a way that takes both statistical and practical significance into account, and controls the Bayesian FDR below a certain threshold. Given this threshold, we also demonstrate how to compute the corresponding estimated false negative rate, sensitivity, and specificity. These principles can be applied to any Bayesian setting yielding posterior samples of effects or of some other indicator of biomarker status. To our knowledge, this is the first presentation of FDR-based methods for the functional data setting. Third, although our method does not require that peak detection be done, we demonstrate how to perform peak detection from the method's output in case it is desired.

The remainder of the article is organized as follows. In Section 2, we describe some preprocessing steps that must be performed before analyzing MALDI-TOF data, and present a systematic method for choosing an additive shrinkage constant before log transforming the spectral intensities. Section 3 describes the wavelet-based functional mixed model, and explains how model specification should proceed for MALDI-TOF data. In Section 4, we present our Bayesian-FDR–based approach for identifying significant regions of the spectra, and describe how to perform peak detection, if desired. We present results from analysis of the example data sets in Section 5, and conclude with a discussion of the strengths and weaknesses of this approach in Section 6.

2. Preprocessing MALDI-TOF Data

A number of preprocessing steps must be performed before modeling MALDI-TOF or SELDI-TOF data, regardless of the ultimate approach used for inference. It has been shown that inadequate or ineffective preprocessing can make it difficult to extract meaningful biological information from the data (Sorace and Zhan, 2003; Baggerly et al., 2003, 2004). These steps include calibration, baseline correction, normalization, denoising, and transformation. Calibration must be done to align the peaks across different spectra. The baseline, frequently seen in MALDI-TOF and SELDI-TOF spectra, is a smooth underlying function that is thought to be largely due to a large cloud of particles striking the detector in the early part of the experiment (Malyarenko et al., 2005). This baseline artifact must be removed. Normalization refers to a constant multiplicative factor that is used to adjust for spectrum-specific factors, for example, to adjust for different amounts of total protein ionized and desorbed from the sample. Denoising is used to remove white noise, which is largely due to electronic noise from the detector, from the spectrum. In recent years, various methods have been proposed to deal with these issues. Here, we use the methods described by Coombes, Tsavachidis, et al. (2005). The first two columns of Figure 1 contain a raw spectrum and corresponding preprocessed MALDI spectrum from a cancer sample and a control sample, and demonstrate the effects of preprocessing.

It is often useful to transform the spectral intensities in order to reduce the skewness in their distribution. Some options that appear to work well include the log transformation and the cube root transformation (Coombes, Tsavachidis, et al., 2005; D. Billheimer, unpublished manuscript). Here, we choose the $\log_2$ transformation because it leads to good interpretations in terms of fold change. For example, a difference of 1 in this scale corresponds to a two-fold increase in intensity.

The presence of zero intensities makes it necessary to add a small positive constant ϵ to each intensity before taking the log. This constant shrinks any fold-change estimates toward 1, with stronger shrinkage at lower intensities. Here we describe a systematic approach for choosing ϵ for a given setting. Suppose we wish to control the shrinkage factor for spectral intensities of at least γ to be no less than α, meaning that the shrunken estimate of a true fold change of δ would be at least

$\delta\alpha$. This is accomplished by choosing $\epsilon = \{(1-\alpha) * \gamma * \delta\}/\{\alpha * \delta - 1\}$. For the analyses presented in this article, we chose $\epsilon = 0.25$. This guaranteed that given a fold-change difference of 2 at spectral locations with intensities of at least 1.0, the fold-change estimate will be no less than 1.8, and at spectral intensities of 5.0 or more, the expected fold-change estimate will be no less than 1.95. Effectively, this choice leads to very little shrinkage in regions of the spectra nearby the true protein peaks, but reduces the possibility that spurious differences will be detected at very low intensities because of the log scale that was used.

3. Wavelet-Based Functional Mixed Models

In this section, we briefly overview the wavelet-based functional mixed model method introduced by Morris and Carroll (2006) and describe how to apply it to mass spectrometry data. See that paper for further details on its modeling assumptions and computational procedure.

The functional mixed model we present here is a special case of the one discussed by Morris and Carroll (2006), and is also like the functional mixed model discussed by Guo (2002). Suppose we observe N functions $Y_i(t)$, $i = 1, \ldots, N$, all defined on the closed interval $\mathcal{T} \in \Re^1$. In MALDI-TOF data, these functions are the preprocessed, log-transformed spectra on the time axis t. A functional mixed model for these data is given by

$$Y_i(t) = \sum_{j=1}^{p} X_{ij} B_j(t) + \sum_{k=1}^{m} Z_{ik} U_k(t) + E_i(t), \qquad (1)$$

where X_{ij} are covariates, $B_j(t)$ are functional fixed effects, Z_{ik} are elements of the design matrix for functional random effects $U_k(t)$, and $E_i(t)$ are residual error processes. We assume that $U_k(t)$ are independent and identically distributed (i.i.d.) mean-zero Gaussian processes with covariance surface $Q(t_1, t_2)$, and $E_i(t)$ are i.i.d. mean-zero Gaussian processes with covariance surface $S(t_1, t_2)$, with $U_k(t)$ and $E_i(t)$ assumed to be independent. A parsimonious yet flexible structure will be used to represent Q and S, as described below. One may allow different strata, $h = 1, \ldots, H$, to have their own covariance matrices Q_h and S_h by splitting the random effect functions and residual error processes into blocks, for example, to allow cancer and control spectra to have different covariance surfaces.

Covariates $\{X_{\cdot j},\ j = 1, \ldots, p\}$, discrete or continuous, are specified for any factor one wants to relate to the mass spectra. Each functional coefficient $B_j(t)$ describes the effect of the corresponding factor at location t of the spectrum. The covariates can include a column of 1's for an overall mean spectrum, continuous or discrete variables of interest, clinical or experimental covariates for which one would like to adjust, and any interactions among these factors. As in linear mixed models, absent constraints one must take care in parameterizing the $X_{\cdot j}$ so that the resulting design matrix $X = (X_{\cdot 1}, \ldots, X_{\cdot p})$ has full column rank.

When the spectra are not independent, the functional random effects provide a flexible mechanism for modeling correlation among spectra. For example, individual-level random effect functions can be specified when multiple spectra are obtained from the same individual, and additional random effect functions can be specified for other clustering units, such as blocks or laboratories when the spectra are obtained over a long period of time or at many different locations.

An important feature of this model is that it places no restrictions on the form of the fixed or random effect functions, because for MALDI-TOF data we expect their true form should be very irregular and spiky. Although their high dimensionality precludes unstructured representation, it is also important to allow flexibility in the forms of Q and S, as described below, because irregular and spiky curve-to-curve deviations imply irregularity in these matrices, as well.

It is possible to write a discrete matrix version of model (1) if we have all spectra observed on the same equally spaced grid $\mathbf{t} = (t_l;\ l = 1, \ldots, T)$, as

$$Y = XB + ZU + E. \qquad (2)$$

Each row of the $N \times T$ matrix Y contains one spectrum observed on the grid $\mathbf{t}$. The matrix X is an $N \times p$ design matrix of covariates; B is a $p \times T$ matrix whose rows contain the corresponding *fixed effect functions* on the grid $\mathbf{t}$. B_{jl} denotes the effect of the covariate in column j of X on the spectrum at clock tick t_l. The matrix U is an $m \times T$ matrix whose rows contain *random effect functions* on the grid $\mathbf{t}$, and Z is the corresponding $N \times m$ design matrix. Each row of the $N \times T$ matrix E contains the residual error process for the corresponding observed spectrum. We assume that the rows of U are i.i.d. MVN($\mathbf{0}$, Q) and the rows of E are i.i.d. MVN($\mathbf{0}$, S), independent of U, with Q and S being $T \times T$ covariance matrices that are discrete analogs of the covariance surfaces in (1), defined on the grid $\mathbf{t} \times \mathbf{t}$.

Morris and Carroll (2006) used a basis function approach to fit the model (2). They chose wavelet basis functions, which possess various properties that make them well suited for representing MALDI-TOF data. First, their compact support allows them to efficiently model the spikes in the data. Second, their whitening property allows us to make parsimonious yet flexible assumptions on the covariances Q and S. Specifically, the assumed structure requires only T parameters for each of these matrices, yet it accommodates various types of nonstationarities characteristic of MALDI-TOF data, for example, allowing the between-spectra variances and within-spectrum smoothness to vary across different regions of the spectra. This point is illustrated by Figure 1 of Morris and Carroll (2006). Third, their decomposition of the spectral energy in both the frequency and time domains makes it possible to perform *adaptive regularization* on the fixed effect functions. By adaptive regularization, we mean that the functional estimates are denoised or smoothed in a manner that tends to preserve strong peaks, which are important features that characterize these functions in MALDI-TOF applications. Finally, given spectra sampled on an equally spaced grid of length T, the special structure of the basis functions allows us to quickly compute a set of T wavelet coefficients using a pyramid-based algorithm, the discrete wavelet transform (DWT), in just $O(T)$ operations. Conversely, given the set of wavelet coefficients, the function can be constructed using the inverse discrete wavelet transform (IDWT), also in $O(T)$ operations.

The wavelet-based approach to fitting the functional mixed model involves three steps. First, the wavelet coefficients are

computed by applying the DWT to each of the N spectra. This step effectively projects the observed spectra into the space spanned by the chosen wavelet bases. Second, a Markov chain Monte Carlo (MCMC) simulation is performed to obtain posterior samples of the model parameters in a wavelet-space version of the functional mixed model. Third, the IDWT is applied to the posterior samples, yielding posterior samples of the parameters B, U, Q, and S in the data-space functional mixed model (2). These posterior samples can subsequently be used to perform any desired Bayesian inference.

Morris and Carroll (2006) made the code for performing these steps freely available at the following website: `http://biostatistics.mdanderson.org/Morris/papers.html`. This code has since been updated to effectively handle the extremely large data sets characteristic of MALDI-TOF data (100's of spectra, each on a grid of 10,000–20,000). The code is a standalone executable that runs on a Windows-based PC, and takes Matlab data files for the required input. The required input includes the matrix of log-transformed, preprocessed spectra, Y, plus a structure specifying the model used (at a minimum, X and Z, if present, must be specified), a structure specifying the wavelet basis to use, and another structure containing the MCMC specifications. Details are given in the documentation provided with the code. The output of this code is Matlab files containing the posterior samples for the model parameters resulting from the MCMC.

The third column of Figure 1 contains spectra randomly generated from the posterior predictive distribution of the wavelet-based functional mixed model fit to the pancreatic cancer example data set, and illustrates that the model is flexible enough to generate functional data characteristic of MALDI-TOF.

4. Bayesian Inference for MALDI-TOF

Here, we describe how to perform Bayesian inference for MALDI-TOF experiments using the posterior samples output from the wavelet-based functional mixed model. First, we describe how peak detection can be done, if desired. Second, we describe how to identify significant regions of the spectra while controlling the expected Bayesian FDR, and then summarize the global properties of this significance rule.

Peak detection: A key benefit of our functional approach is that peak detection is unnecessary. For those who still wish to restrict attention to the peaks, however, it is straightforward to perform peak detection from the posterior samples output from the wavelet-based functional mixed model. Morris et al. (2005) describe a peak detection approach and demonstrate that performing peak detection on the mean spectrum results in greater sensitivity and specificity than the usual approach of performing peak detection on the individual spectra. Because the mean spectrum is easily obtainable from the functional mixed model either as a fixed effect function or as a linear combination of fixed effect functions, it is easy to adapt the procedure described in that paper to detect and quantify peaks in this setting, as well.

The mean spectrum estimate from the wavelet-based functional mixed model differs from the simple pointwise mean spectrum in several ways. First, it is denoised (adaptively regularized) as a result of the shrinkage induced by a spike-slab prior that is assumed on the wavelet coefficients for the fixed effect functions. The benefit of this denoising is that it reduces the number of small, spurious bumps that are called peaks. Second, the mean spectrum estimate will adjust for other effects in the model. For example, including a block or laser intensity effect improves the alignment across the different groups of spectra, thus sharpening the peaks in the estimated mean spectrum and making them easier to detect. Third, the denoising of the mean spectrum is affected by the random effect and residual variance structure of the functional mixed model. As can be seen by the formulas presented in Morris and Carroll (2006), both the random effect structure and residual variance directly impact the wavelet shrinkage of the fixed effect functions. This may also lead to improved denoising over the simple pointwise mean spectrum, especially in data sets with imbalanced designs in terms of the number of spectra per individual.

Morris et al. (2005) also perform a wavelet-based denoising step on the mean spectra before detecting peaks. An important difference is that the approach described in that paper uses the undecimated discrete wavelet transform (UDWT, also sometimes called the translation invariant or maximum overlap discrete wavelet transform), whereas the wavelet-based functional mixed model works with the decimated DWT (DDWT). The UDWT is translation invariant, whereas the DDWT is not. This means that arbitrary translations in the x-axes of the spectra will result in different wavelet coefficients for the DDWT, but not for the UDWT. It has been observed that the translation-invariant property can lead to better denoising, but at the cost of computational time and parsimony. The calculations for the UDWT are of order up to $O\{T \log(T)\}$ rather than $O(T)$, and the full range of translations would yield a great deal more coefficients to model, increasing the memory demands on the procedure. Although it would be possible to apply the wavelet-based functional mixed model in the UDWT context, we choose to stick with the DDWT here because with reasonably aligned spectra taken at a high sampling frequency, the differences are not great, and the increased computational and computer memory demands from the UDWT would make it more difficult to apply the method to these extremely large data sets.

Identifying significant regions of spectra: Our primary analysis goal in this article is to identify regions of the spectra that are differentially expressed across factors of interest, which can subsequently be mapped to proteins that may serve as useful biomarkers. In microarrays, two classical approaches for handling differential expression are (i) to identify all genes with a fold-change difference of at least δ and (ii) to identify genes that differ significantly across treatment groups according to a statistical hypothesis test. Option (i) is intuitive to many researchers but lacks statistical rigor because it ignores the variability in the data, and option (ii) only focuses on statistical significance, ignoring practical significance, because it is typically based on a null hypothesis of equality. In the present MALDI-TOF context, we identify differentially expressed regions of the spectra in a way that considers both statistical and practical significance, and controls the expected Bayesian FDR to be no more than α.

Suppose we are interested in identifying biomarkers that have at least a δ-fold intensity change between treatment

groups. From the MCMC, suppose we have G posterior samples of the corresponding fixed effect function $\mathbf{B}_j = [B_j(t_1), \ldots, B_j(t_T)]$ on the $\log_2$ scale, denoted by $\{\mathbf{B}_j^{(g)}, g = 1, \ldots, G\}$. From these, we compute the pointwise posterior probabilities of at least a δ-fold intensity change at each spectral location as $p_j(t_l) = \Pr\{|B_j(t_l)| > \log_2(\delta) \,|\, Y\} \approx G^{-1} \sum_{g=1}^{G} I\{|B_j^{(g)}(t_l)| > \log_2(\delta)\}$ for $(t_l, l = 1, \ldots, T)$. We replace any $p_j(t_l) = 1$ with $1 - (2 * G)^{-1}$. These posterior probabilities can also be computed for any contrast involving the fixed effect functions, $A^{(g)} = \sum_{j=1}^{p} C_j \mathbf{B}_j^{(g)}$, or similar posterior probabilities can be computed for linear combinations of spectral locations, for example, if one wanted to detect peaks and look at areas under peaks, or only consider t_l that are flagged as peaks. The quantity $1 - p_j(t_l)$ can be considered a local FDR estimate for location t_l for factor i. Global properties are described below.

Given a desired global FDR-bound α, we flag the set of locations $\psi_j = \{t_l : p_j(t_l) > \phi_\alpha\}$ as significant spectral regions for factor j. In order to obtain ϕ_α, we first sort $\{p_j(t_l), l = 1, \ldots, T\}$ in descending order to obtain $\{p_{(l)}, l = 1, \ldots, T\}$. Then $\phi_\alpha = p_{(\lambda)}$, with $\lambda = \max[l^* : (l^*)^{-1} \sum_{l=1}^{l^*} \{1 - p_{(l)}\} \leq \alpha]$. The threshold ϕ_α is a cutpoint on the posterior probabilities that controls the expected Bayesian FDR at level α, in the sense that on average we expect $\geq 100(1-\alpha)\%$ of the locations in the set ψ_j to have a true δ-fold difference in expression, as estimated by the wavelet-based functional mixed model. That is, if $\mathcal{N}(\psi_j)$ is the cardinality of the set ψ_j, defined as $\mathcal{N}(\psi_j) = \sum_{l=1}^{T} I(t_l \in \psi_j)$, then $\mathcal{N}(\psi_j)^{-1} \sum_{t_l \in \psi_j} \Pr\{|B_j(t_l)| \leq \log_2(\delta) \,|\, Y\} \leq \alpha$. If p^* factors are to be investigated simultaneously, it is possible to either use one common threshold ϕ_α or separate thresholds for each factor, $\{\phi_{j,\alpha}, j = 1, \ldots, p^*\}$. This use of Bayesian FDR is similar in principle to the approach used by Newton et al. (2004).

Given the set of locations $\psi_j = \{t_l : p_j(t_l) > \phi_\alpha\}$ flagged as *discoveries*, we can compute model-based estimates of the FDR, false negative rate, sensitivity, and specificity for detecting differentially expressed locations. Defining $\psi_j' \cup \psi_j = \mathcal{T}$, and $\mathcal{N}(\mathcal{S})$ as the cardinality of set $\mathcal{S}$, defined as above, the FDR is estimated by $\{\mathcal{N}(\psi_j)\}^{-1} \sum_{t_l \in \psi_j} \{1 - p_j(t_l)\}$, the false negative rate by $\{\mathcal{N}(\psi_j')\}^{-1} \sum_{t_l \in \psi_j'} \{p_j(t_l)\}$, sensitivity by $\{\sum_{l=1}^{T} p_j(t_l)\}^{-1} \sum_{t_l \in \psi_j} p_j(t_l)$, and specificity by $[\sum_{l=1}^{T} \{1 - p_j(t_l)\}]^{-1} \sum_{t_l \in \psi_j'} \{1 - p_j(t_l)\}$. In the idealized functional setting, ψ_j is a set of continuous regions, and the estimates given above are approximations of the continuous versions of these quantities obtained by substituting integrals over t for the summations, and defining $\mathcal{N}(\mathcal{S})$ as the Lebesgue measure of set $\mathcal{S}$. The interpretations of these quantities in the continuous case are also analogous. For example, if flagging a region ψ_j as significant corresponds to an FDR of α, then the expected proportion of the set of contiguous regions ψ_j that is truly differentially expressed at least δ-fold is $1 - \alpha$, based on Lebesgue measure.

For MALDI-TOF data, these measures do not depend heavily on the sampling frequency of the data. To demonstrate this point, we computed the threshold ϕ_α for the pancreatic cancer example while downsampling the spectra in multiples of 2, 3, 4, 6, and 8. The results are available in Supplementary Figure 2. We found nearly identical thresholds and flagged regions in each analysis.

Although applied to the wavelet-based functional mixed model setting here, this approach for identifying cutpoints for significance and summarizing the resulting global properties can be used in any Bayesian context in which we obtain the posterior probability of "discovery" for each of a number of discrete units or continuous regions.

5. Analysis of Example Data

For both examples, we modeled the spectra on the time scale t but plotted results on the biologically meaningful mass-per-unit-charge scale (m/z, x). In our wavelet-space modeling, we chose the Daubechies wavelet with vanishing fourth moments and performed the DWT down to $J = 10$ and $J = 9$ levels for the two respective examples. Other wavelet bases were examined and yielded equivalent results. We used the modified empirical Bayes procedure described by Morris and Carroll (2006) to estimate the shrinkage hyperparameters that guide the adaptive regularization of the fixed effect functions. For each example, we ran 10 parallel chains, each consisting of 1000 iterations after a burn-in of 1000, and we kept every fifth iteration for a total of $G = 2000$ MCMC samples for our analyses. All chains appear to have converged, as indicated by trace plots. In the pancreatic cancer example, the median and 99% intervals for the Metropolis–Hastings acceptance probabilities across the roughly 12,000 covariance parameters were 0.22 and (0.11, 0.31), respectively, and for the organ-by-cell line example with roughly 8000 covariance parameters, they were 0.17 and (0.05, 0.51), respectively.

We explored the possible protein identities of any flagged regions by running the estimated m/z values of the corresponding peaks in the region through TagIdent, a searchable database (available at `http://us.expasy.org/tools/tagident.html`) that contains the molecular masses and pH for proteins observed in various species. For the organ-by-cell line example, we searched for proteins emanating from both the source (human) and the host (mouse) whose molecular masses were within the estimated mass accuracy (0.3%) of the instrument from the nearest peak or most significant location of each flagged region. This only gives an educated guess at what the protein identity of the peak could be; it is necessary to perform an additional mass spectrometry/mass spectrometry (MS/MS) experiment in order to rigorously validate the protein identity.

Pancreatic cancer example: The design matrix for this data set of $N = 256$ spectra was chosen to have $p = 5$ columns, the first column indicating cancer (=1) or normal (=−1) status, and corresponding to a functional cancer main effect $B_1(t)$ describing the difference between the mean $\log_2$ intensities of cancer and normal spectra at time t. The final four columns indicate the time blocks, and correspond to mean spectra for the respective time blocks $(B_i(t), i = 2, \ldots, 5)$. The block effects between block i and i' can be constructed by $B_i(t) - B_{i'}(t)$. No functional random effects were specified. The residual covariance matrix S was allowed to vary across cancer status.

The top two panels of Figure 2 contain posterior means and 95% credible intervals for the cancer main effect function

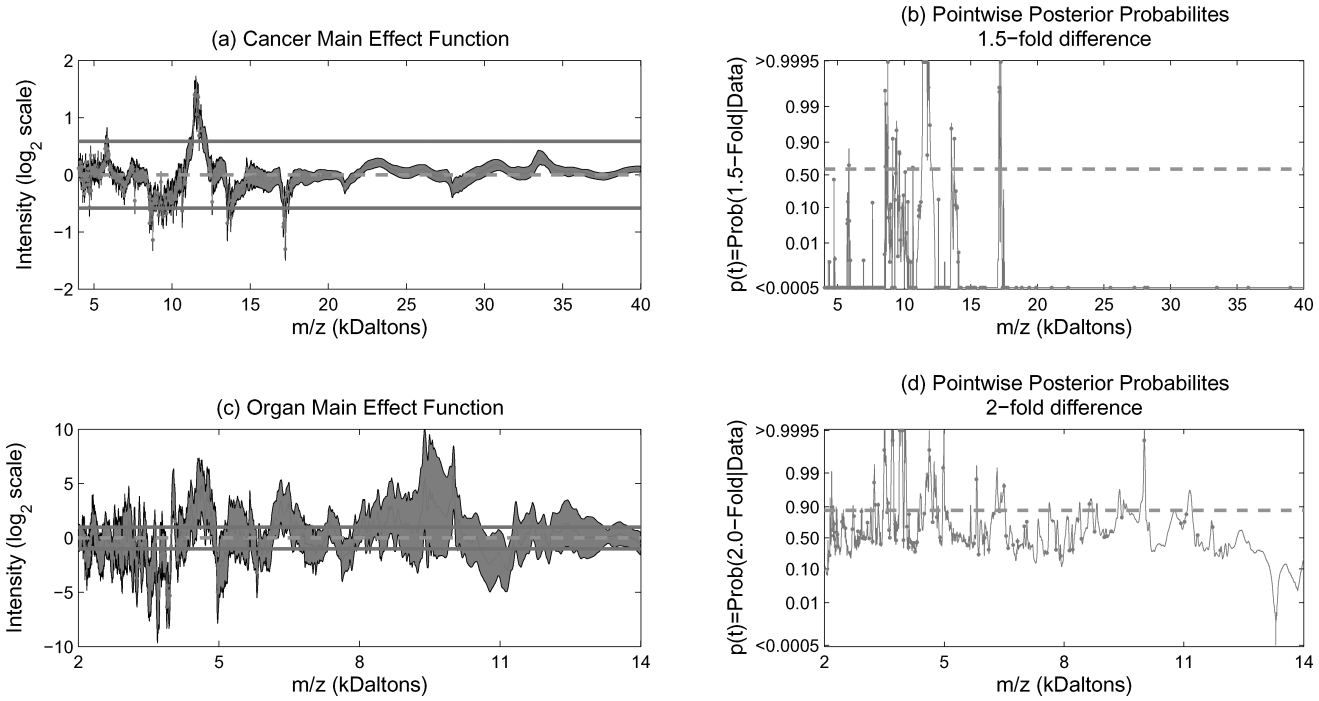

Figure 2. Fixed effect curves. (a) and (c): Posterior mean and 95% pointwise posterior credible bands for cancer main effect, pancreatic cancer example, and organ main effect, organ-by-cell line example, respectively. The horizontal lines indicate 1.5-fold and 2.0-fold differences in the two examples, respectively, and the dots indicate peaks detected using the average spectrum. (b) and (d): Pointwise posterior probabilities of (b) 1.5-fold difference in cancer/normal in pancreatic cancer example and (d) 2.0-fold difference in brain/lung in organ-by-cell line example. The dots indicate detected peaks, and the dotted lines indicate the threshold for flagging a location as significant, controlling the expected Bayesian FDR to be less than 0.10 and 0.05 in the two examples, respectively. Any spectral locations with posterior probabilities above this line were flagged as significant. This figure appears in color in the electronic version of this article.

and the corresponding pointwise posterior probabilities of at least 1.5-fold expression. The dots in the plots correspond to the 227 peaks detected on the posterior mean for the overall mean spectrum $\hat{\mu}(t) = (4G)^{-1} \sum_{g=1}^{G} \sum_{i=2}^{5} B_i^{(g)}(t)$. The horizontal dotted line in the upper right panel indicates the threshold on the posterior probabilities $\phi_{10} = 0.595$ corresponding to an expected Bayesian FDR at 0.10. This threshold yielded a false negative rate of 0.016, a sensitivity of 0.716, and a specificity of 0.996. There were a total of 506 spectral locations contained within 16 contiguous regions that were flagged as significant (i.e., appear above the 0.595 threshold in the upper right picture). Analyzing the peaks, we found 26/227 were flagged as significant. A list containing the significant regions and peaks, and a plot of the overall mean spectrum with detected peaks are available in Supplementary Tables 1–3 and Supplementary Figure 3.

The most significant effects were observed in the regions (i) (17230D, 17311D), (ii) (8730D, 8787D), and (iii) (11314D, 12037D), with maximum posterior mean fold-change differences of 1/2.46, 1/2.20, and 2.77, respectively, between cancers and normals. A fold change of δ means that cancer was overexpressed relative to normal by a factor of δ, whereas a fold change of $1/\delta$ means that normal was overexpressed relative to cancer by a factor of δ. The maximum fold-change differences for all three of these regions were located at peaks. These were also identified in Koomen et al. (2005). In that paper, they reported MS/MS results confirming the identity of (i) as a fragment of apolipoprotein A-I or apolipoprotein glutamine-I, and the cluster of 7 peaks in (iii) as serum amyloid A. Based on TagIdent, region (ii) may correspond to complement C4-A or C4-B(precursor), 8764.07D, mediators of inflammatory processes that circulate in the blood.

One peak (4284D) found to be statistically significant and highlighted by Koomen et al. (2005) had a very small fold-change estimate (1.22), and thus by design was not flagged by our analysis. Also interesting was the region (8671D, 8684D) that was on the upslope of a very abundant peak at 8688D. The peak itself was not flagged ($p = 0.186$), but this region was, with a maximum fold change of 1/1.70 at 8679 ($p = 0.968$). It is possible that this result is driven by protein at 8679D whose peak is not visible because of its proximity to the extremely abundant peak at 8688D. An MS/MS experiment would have to be done to investigate this possibility.

Plots of the block effects (see Supplementary Figure 4) demonstrate that they affect both the location and intensity of peaks, and are of a similar magnitude as the cancer main effect. Figure 3 illustrates the block effect (block 1–block 2) in the neighborhood of some prominent peaks. The nonparametrically modeled block effects were able to capture both shifts in intensity (Figure 3a) and shifts in location (Figure 3b). Note that shifts in location appear as pulses in the nonparametric block effects. These features served to calibrate the x and y axes across blocks so that they were comparable,

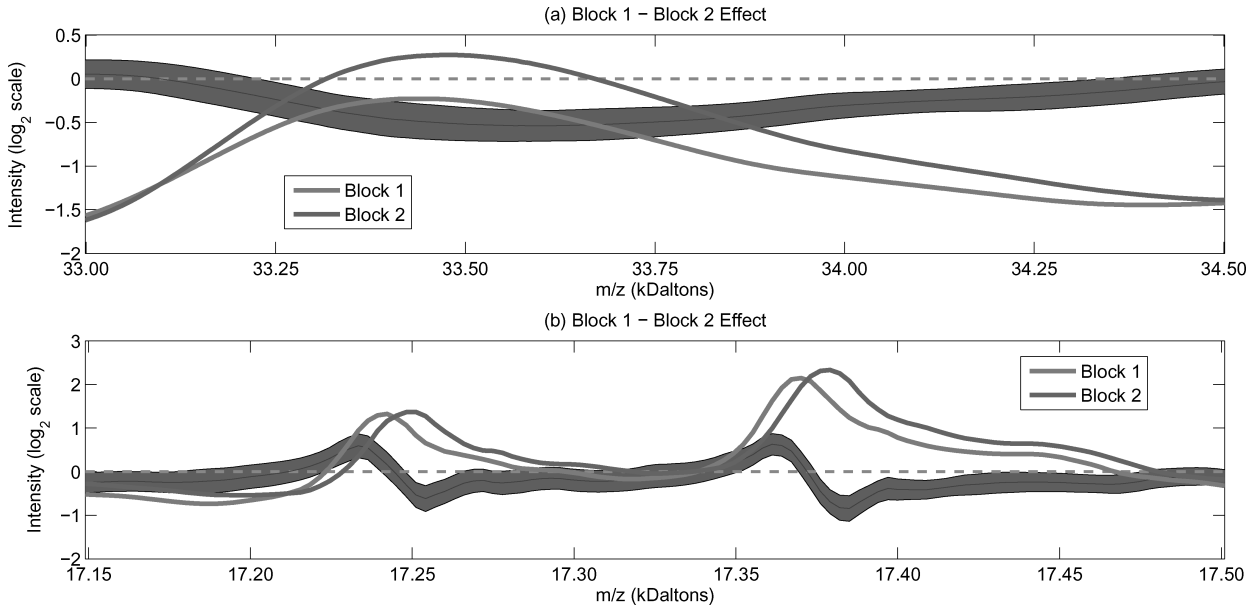

Figure 3. Block effects. Plot of the mean spectra for blocks 1 (light line) and 2 (dark line), along with the posterior mean and 95% pointwise posterior bounds for the block 1–block 2 effect (shaded area and associated lines) near (a) the peak at 33,482 and (b) the twin peaks at 17,245 and 17,376. Part (a) illustrates that the nonparametric functional effect can model changes in intensity, and (b) shows that the pulse-like features of the nonparametric effect account for systematic shifts in location. This figure appears in color in the electronic version of this article.

allowing spectra from different blocks to be pooled for a combined analysis.

Organ-by-cell line example: The design matrix for this set of $N = 32$ spectra had $p = 5$ columns. We used a cell means model for the factorial design, so the first four columns contained indicators of the four organ-by-cell line groups with corresponding mean functions $B_i(t)$, $i = 1, \ldots, 4$, ordered brain-A375P, brain-PC3MM2, lung-A375P, and lung-PC3MM2. From these, the overall mean spectrum $0.25 \sum_{i=1}^{4} B_i(t)$, the organ main effect function $B_1(t) + B_2(t) - B_3(t) - B_4(t)$, cell-line main effect function $B_1(t) - B_2(t) + B_3(t) - B_4(t)$, and the organ-by-cell line interaction function $B_1(t) - B_2(t) - B_3(t) + B_4(t)$ were constructed. Column 5 indicates whether a low (-1) or high (1) laser intensity setting was used in generating the given spectrum. The Z matrix had $m = 16$ columns, with $Z_{ik} = 1$ if spectrum i came from animal k, with corresponding mouse-level random effect functions $U_k(t)$, $k = 1, \ldots, 16$. These random effects allow our model to account for the correlation between different spectra generated from the same animal.

The bottom two panels of Figure 2 contain the posterior means and 95% credible intervals for the organ main effect function and the corresponding pointwise posterior probabilities of at least 2-fold difference, respectively. Equivalent plots for the cell-line and interaction effects are available in Supplementary Figure 6. The threshold on the posterior probabilities based on setting the expected Bayesian FDR of 0.05 was $\phi_{05} = 0.874$, which led to a false negative rate of 0.469, a sensitivity of 0.204, and a specificity of 0.987. We flagged 1393/7985 of the spectral locations in 41 contiguous regions for the organ main effect, 798/7985 in 25 contiguous regions for the cell-line main effect, and 594/7985 in 18 contiguous regions for the organ-by-cell line interaction effect. Of the 101 detected peaks, we flagged 40 as significant, 13 for organ alone, 13 for cell line, 1 for both organ and cell line, and 13 for the interaction. Table 1 contains information for the top 10 most significant regions, all of which contained locations with posterior probabilities $p_j(t_l) > 0.9995$. The complete list of significant regions and peaks is available in Supplementary Tables 4–7.

The strongest differences observed were between organ groups. The largest estimated fold changes were observed in the regions [3658D, 3739D] and [3866.3D, 3971.3D]. These regions each contain a peak that is strongly present in all mice with tumors injected into their brains, but absent from those injected in their lungs. The region [3866.3, 3971.3] is represented in Figure 4a and c. This region may correspond to a calcitonin gene-related peptide II precursor (CGRP-II, 3882D), a peptide in the mouse proteome that dilates blood vessels in the brain and has been observed to be abundant in the central nervous system (`http://www.expasy.org/uniprot/Q99MP3`). The region [3658D, 3739D] may correspond to a precursor of amyloid beta A4 protein in the mouse proteome (3717.1D) that "functions as a cell surface receptor and performs physiological functions on the surface of neurons relevant to neurite growth, neuronal adhesion and axonogenesis," and is "involved in cell mobility and transcription regulation through protein-protein interactions" (`http://www.expasy.org/uniprot/P12023`). Another flagged region [10912D, 11269D] may also correspond to a precursor of the same protein (11050.6D). These results may represent important responses within the hosts to the tumor implantation in their brains.

There were some significant effects that may not have been detected had we restricted our attention to the peaks. The

Table 1

Selected flagged regions from organ by cell-line example. Location of selected region (in Daltons per coulomb) is given, along with which effect was deemed significant, estimated maximum fold change difference within the region, and a description of the effect. These effects comprise all those with $p_l > 0.9995$.

Region	Effect type	max FC	Comment
3866.3–3971.3	Organ	1/93.9	Only in brain-injected mice
3658.3–3739.0	Organ	1/118.5	Only in brain-injected mice
9902.6–10044.0	Organ	46.1	Only in lung-injected mice
4762.2–4874.8	Interaction	1/13.7	PC3MM2 > A375P, especially brain
4748.2–4868.3	Cell line	1/39.7	PC3MM2 > A375P
3743.4–3565.3	Organ	1/35.0	Brain > Lung
4952.6–5008.2	Organ	1/32.8	Brain > Lung
4519.9–4697.5	Organ	27.5	Lung > Brain
5051.3–5093.3	Cell line	1/23.5	PC3MM2 > A375P
3993.4–4061.3	Organ	21.0	Lung > Brain (on upslope of peak)
10912–11269	Interaction	1/16.4	Brain > Lung for A375P only

significant organ effect in the region [3993D, 4061D], with maximum fold-change difference of 21.0, is on the upslope of a peak, but the peak value itself was not significant. Also, the region [7618D, 7650D] was flagged for an organ effect, being specific to brain-injected mice. The protein neurogranin in the human proteome, with a molecular weight of 7618.5D, is active in synaptic development and remodeling in the brain. Our mean spectrum-based peak detection procedure found no peak in the region [7618D, 7650D], so this potential discovery may have been missed had we restricted attention to the peaks.

Of the 25 regions flagged as significantly different across cell lines, 22 of them were overexpressed in the metastatic PC3MM2 cell line relative to the nonmetastatic A375P cell line. Plots of the laser intensity effect (Supplementary Figure 7) reveal systematic differences between the low and high laser intensity spectra that affect both the locations and intensities of peaks. Our nonparametric laser intensity effect

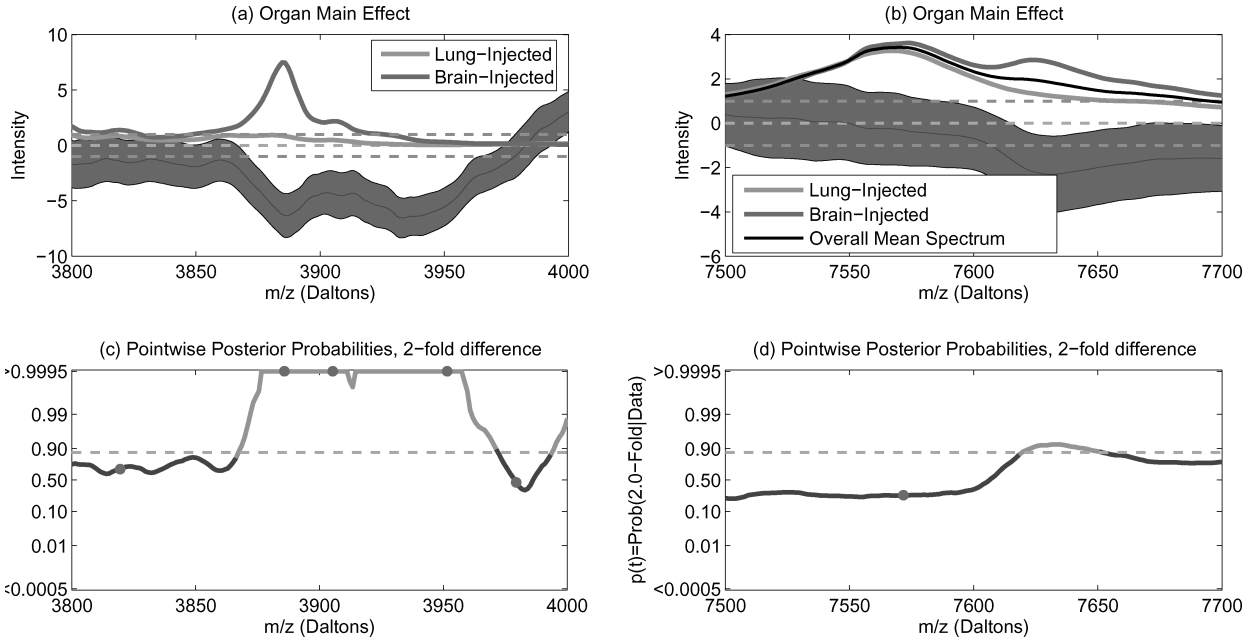

Figure 4. Select results. (a) and (b): Plot of organ main effect function in selected regions. The light and dark lines are the organ-specific mean spectra on the untransformed intensity scale, the shaded regions and associated lines are the posterior mean and pointwise 95% posterior bounds for the organ main effect on the $\log_2$ intensity scale. The dotted lines at 0 and at $+/-1$ are provided for reference. (c) and (d): Pointwise posterior probabilities of 2-fold difference in intensity. The dots indicate peaks detected in the mean spectrum, and the dotted line indicates the threshold on pointwise posterior probabilities chosen so that the expected Bayesian FDR < 0.05. The lighter regions of the plot above the dotted line indicate regions flagged as significant. This figure appears in color in the electronic version of this article.

was able to model this difference, allowing us to pool data from both laser intensities for this analysis.

6. Discussion

We have demonstrated how to use the recently developed Bayesian wavelet-based functional mixed model to analyze MALDI-TOF proteomics data. This method appears well suited for this context, for several reasons: the functional mixed model is very flexible; it is able to simultaneously model nonparametric functional effects for multiple covariates, both factors of interest and nuisance factors such as block effects. The nonparametric functional effects for nuisance factors are flexible enough to account for systematic changes in both the location and intensity of peaks in the spectra. Further, the random effect functions can be used to model correlation among spectra that might be induced by the experimental design. The wavelet-based modeling approach works well for modeling functional data with many local features such as MALDI-TOF peaks because it results in adaptive regularization of the fixed effect functions, avoids attenuation of the effects at the peaks, and is reasonably flexible in modeling the between-curve covariance structures, accommodating autocovariance structures induced by peaks and heteroscedasticity allowing different between-spectrum variances for different peaks. The method is extremely adaptive in terms of the types of functions it can represent. The example in Morris and Carroll (2006) illustrates that it can handle smooth fixed and random effect functions and spiky residual error processes, whereas our examples here demonstrate that it can also handle very spiky fixed and random effect functions as well. The use of wavelet bases and the resulting adaptive regularization are keys to this flexibility.

Before applying this method to mass spectra, it is important to perform adequate preprocessing, at a minimum to remove the baseline artifact and align the spectra. Variable baselines, if not removed, can add extra noise variability to the data, making it more difficult to identify meaningful differences. Misalignment in the spectra will cause the fixed effect functions to be less peaked, and will increase the variability across spectra, also decreasing the power for detecting differentially expressed regions of the spectra. In our experience, spectra obtained on a given day in a given laboratory tend to be quite well aligned with each other, and require no further alignment. It is still a good idea to use a heat map of the spectra to check, and then to perform some type of function registration if the alignment is off. Spectra from different laboratories or obtained at different times, however, are frequently severely misaligned. These systematic misalignments appear to be handled quite well in the functional mixed model framework by including nonparametric block and laboratory effects in the functional mixed model, as long as these block effects are not completely confounded with other effects in the model. If there is complete confounding due to poor experimental design, however, then there is little that any statistical analysis can do to factor out the confounding effects (Baggerly et al., 2004, 2005; Coombes, Morris, et al., 2005).

We applied this method to two cancer proteomic studies, and identified spectral regions that were differentially expressed and may correspond to potential biomarkers. Many of these regions contained peaks, but several may not have been found had attention been restricted to peaks alone. Another benefit of our approach is that both statistical and practical significance were considered in identifying potential biomarkers.

In the pancreatic cancer example, this method was able to model nonparametric block effects that served to calibrate the x and y axes across blocks, making spectra from the different time blocks comparable and enabling them to be pooled for a common analysis. In a similar fashion, the incorporation of the nonparametric laser intensity effect in the organ-by-cell line example allowed us to account for systematic differences in spectral intensity and peak locations between the high and low laser intensity spectra. Along with the nonparametric random effects accounting for the correlation between spectra from the same animal, this allowed us to pool data across laser intensities for a common analysis, potentially increasing our power for detecting differentially expressed proteins.

Although the method is complex, it is relatively straightforward to implement using the code freely available at `http://biostatistics.mdanderson.org/Morris/papers.html`. The user only needs to construct a matrix Y containing the preprocessed spectral intensities for the N spectra in the study and specify the design matrices X and Z. Starting values, empirical Bayes and vague proper priors, and proposal variances are all automatically computed by the program and can be used without any user input. Default choices for wavelet basis and levels of decomposition are also automatically computed and can be used, if desired. The code yields posterior samples and summary statistics for all quantities in the functional mixed model, from which Bayesian inference can be conducted in a straightforward fashion. The method is computationally intensive, but the code has been optimized to be able to handle very large data sets, and parallel processing can further speed the computations when it is available. For example, on average each chain of 2000 MCMC iterations for our pancreatic cancer example with 256 spectra and 12,096 observations per spectra took under an hour to run. In our analysis, we ran 10 of these chains in parallel using `Condor` (`http://www.cs.wisc.edu/condor`), a parallel processing freeware that shared the job among roughly 10 Pentium IV computers in a Windows network. Computational issues are discussed in more detail in Herrick and Morris (2006).

We described a systematic method for choosing the additive shrinkage constant before log transformation in settings when estimating fold changes is the question of interest. This method has applications outside of this work, in the setting of microarrays and in other measurement technologies. We presented a new method for identifying significant regions of a curve that takes both statistical and practical significance into account, while controlling the Bayesian FDR at a prespecified level. We discussed how to assess the properties of these discovery rules in terms of FDR and false negative rates, sensitivity and specificity. These approaches can also be applied outside the context of this article to any situation for which posterior samples of fold changes are given, including Bayesian models for microarray data, as well as other Bayesian settings.

Wavelet-based functional mixed models show great promise for the analysis of MALDI-TOF proteomic data. This approach may also prove useful for analyzing data from other biomedical platforms that generate irregular functional data.

7. Supplementary Materials

Web Tables and Figures referenced in Sections 1, 4, and 5 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank Jim Abbruzzesse, Nancy Shih, Stan Hamilton, Donghui Li, John Koomen, and Ryuji Kobayashi for the data sets used in this article. We also thank the associate editor and two referees, whose insightful comments have led to a much improved article. This work was supported by a grant from the National Cancer Institute (CA-107304), and the UK Department of Trade and Industry Texas-UK Collaborative Initiative in Bioscience.

References

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 mass spectrometry patterns in serum: Comparing proteomic data sets from different experiments. *Bioinformatics* **20,** 777–785.

Baggerly, K. A., Morris, J. S., Edmonson, S., and Coombes, K. R. (2005). Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute* **97,** 307–309.

Conrads, T. P. and Veenstra, T. D. (2005). What have we learned from proteomic studies of serum? *Expert Review of Proteomics* **2,** 279–281.

Coombes, K. R., Fritsche, H. A., Jr., Clarke, C., Cheng, 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 using surface enhanced laser desorption and ionization. *Clinical Chemistry* **49,** 1615–1623.

Coombes, K. R., Morris, J. S., Hu, J., Edmonson, S. R., and Baggerly, K. A. (2005). Serum proteomics profiling: A young technology begins to mature. *Nature Biotechnology* **23,** 291–292.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., and Kobayashi, R. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra using the undecimated discrete wavelet transform. *Proteomics* **41,** 4107–4117.

Guo, W. (2002). Functional mixed effects models. *Biometrics* **58,** 121–128.

Herrick, R. C. and Morris, J. S. (2006). Wavelet-based functional mixed models analysis: Computational considerations. Joint Statistical Meetings 2006 Proceedings. ASA Section on Statistical Computing.

Hu, J., Coombes, K. R., Morris, J. S., and Baggerly, K. A. (2005). The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales. *Briefings in Genomics and Proteomics* **3,** 322–331.

Koomen, J. M., Shih, L. N., Coombes, K. R., Li, D., Xiao, L. C., Fidler, I. J., Abbruzzese, J. L., and Kobayashi, R. (2005). Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clinical Cancer Research* **11,** 1110–1118.

Malyarenko, D. I., Cooke, W. E., Adam, B. L., Malik, G., Chen, H., Tracy, E. R., Trosset, M. W., Sasinowski, M., Semmes, O. J., and Manos, D. M. (2005). Enhancement of sensitivity and resolution of SELDI TOF-MS records for serum peptides using time series analysis techniques. *Clinical Chemistry* **51,** 65–74.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* **68,** 179–199.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. *Bioinformatics* **21,** 1764–1775.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

Randolph, T. W., Mitchell, B. L., McLerran, D. F., Lampe, P. D., and Feng, Z. (2005). Quantifying peptide signal in MALDI-TOF mass spectrometry data. *Molecular and Cellular Proteomics* **4,** 1990–1999.

Sorace, J. M. and Zhan, M. (2003). A data review and reassessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **9,** 4–24.

Villanueva, J., Philip, J., Chaparro, C. A., Li, Y., Toledo-Crow, R., DeNoyer, L., Fleisher, M., Robbins, R. J., and Tempst, P. (2005). Correcting common errors in identifying cancer-specific peptide signatures. *Journal of Proteome Research* **4,** 1060–1072.

Yasui, T., Pepe, M., Thompson, M. L., Adam, B. L., Wright, G. L., Jr., Qu, Y., Potter, J. D., Winget, M., Thornquist, M., and Feng, Z. (2003). A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4,** 449–463.

Received March 2006. *Revised June* 2007.
Accepted June 2007.