Multiscale processing of mass spectrometry data

T. W. Randolph*

Department of Biostatistics, University of Washington, Seattle, WA, 98195

and

Y. Yasui

Department of Public Health Sciences, University of Alberta, Edmonton, AB T6G 2G3

October 10, 2005

Summary

This work addresses the problem of extracting signal content from protein mass spectrometry data. A multiscale decomposition of these spectra is used to focus on local scale-based structure by defining scale-specific features. Quantification of features is accompanied by an efficient method for calculating the location of features which avoids estimation of signal-to-noise ratios or bandwidths. Scale-based histograms serve as spectral-density-like functions indicating the regions of high density of features in the data. These regions provide bins within which features are quantified and compared across samples. As a preliminary step, the locations of prominent features within coarse-scale bins may be used for a crude registration of spectra. The multiscale decomposition, the scale-based feature definition, the calculation of feature locations and subsequent quantification of features is carried out by way of a translation-invariant wavelet analysis.

Key words: mass spectrometry, multiscale structure, feature extraction, wavelets

*email: trandolp@u.washington.edu
1. Introduction

In current research on finding biomarkers for disease, the use of high-throughput mass spectrometry produces large data sets comprised of spectra whose graphs are of the type shown in Figure 1. On the horizontal axis are mass/charge (m/z) values, or Thompson, and on the vertical axis an intensity measurement that indicates a (relative) abundance of the particle. Analysis of these data typically relies on inferring the existence of a peptide of a particular mass/charge from the existence of a spike (or similar feature) in the spectrum. This process is confounded by the variability in the m/z location and the shape/size of features when compared across samples.

[Figure 1 about here.]

This work uses a multiscale decomposition of mass spectrometry spectra and focuses on local structure occurring at various scales. The primary goal is to formally define “features” and provide an efficient means for locating and quantifying them. The definition is based on local changes (as opposed to local averages) which occur at a given scale. Once a scale (or, potentially, a sum of scales) is chosen, features are unambiguously located and quantified without estimating additional parameters.

The focus here is on local changes in the spectrum that correlate with a wavelet function of a particular scale (indexed by \( j \), corresponding to width) at a particular location (indexed by \( t \), corresponding to a m/z value). Such information is easily extracted when the spectrum is decomposed into a sum of functions, each containing signal content on a different scale. In this way, features can be defined without reference to spectrum intensity or signal-to-noise ratio and without pre-specifying a window width within which a feature
is deemed present or absent: scale-specific features are defined as (scale-based) local extrema in functions that arise from the multiresolution decomposition of the original spectrum. The \( m/z \) location of a feature is recorded for each spectrum and then using all samples a histogram of these locations is constructed. The histogram serves as a spectral-density-like function whose peaks indicate the high-density regions of features in the data.

Section 2 reviews the concept of multiscale decompositions via a translation-invariant wavelet transformation. In contrast to many wavelet-based methods the goal is not that of using the wavelet coefficients for data compression or denoising. Rather, the aim is to extract relevant features by examining the signal content in a scale-by-scale manner. This point of view is contrasted with other processing methods in Section 5. Section 3 describes how a decomposition leads defining and extracting features while in Section 4 histograms of feature locations from the entire data set are used to match features across spectra. Whether these features are biologically relevant is a question that is briefly addressed in subsection 4.1. Section 6 considers a method of crudely registering the samples based landmarks that occur in a coarse scale of the decomposition. Quantifying the intensity of a feature based on local change (hence baseline free) in described in Section 7. This is followed by a summary of the processing steps in Section 7.1. Section 8 provides specifics on the the processing of the spectra displayed in the paper.

Most of these spectra come from a set of 220 MALDI-TOF (matrix assisted laser desorption and ionization) time-of-flight (TOF) spectra obtained from serum samples in a study of dietary biomarkers (Mitchell et al., 2005). One sample from this data consists of a single mass spectrometry spectrum,
denoted by $X$, with intensity values $X(t)$, $0 \leq t \leq T$. In the present context, the argument $t$ is the time-of-flight for the particles detected in a spectrum; there are $T = 31,022$ discretely recorded times. A monotone function converts the TOF, $t$, to Thompson, via $m/z = (at + b)^2$, $t > -b/a$, corresponding to the range 1,500 to 20,000 $m/z$. Many figures show spectra plotted after conversion to the Thompson scale, but all wavelet transformations and subsequent analyses are performed on the TOF domain.

2. Multiscale wavelet decomposition

The basic property to be exploited is that a spectrum $X$ can be decomposed into a sum of constituent functions, each containing a particular scale of events: at the $j$th dyadic scale, the scale-$j$ detail function, $D_j$, arise from changes in $X$ that occur across a $2^j$-unit domain; this is extracted from $S_{j-1}$ so that $S_{j-1} = S_j + D_j$. Continuing through $j = 1, \ldots, J$ ($J < \lfloor \log_2 T \rfloor$) gives

$$X(t) = S_J(t) + \sum_{j=1}^J D_j(t), \quad 0 \leq t \leq T. \quad (1)$$

Figure 2 exhibits some individual terms, $D_j$ (performed here by a translation invariant discrete wavelet transform and its inverse). More precisely,

**Definition.** A multiresolution, or multiscale, analysis of $L^2(\mathbb{R})$ is a sequence of closed subspaces $V_{j+1} \subset V_j$, $j \in \mathbb{Z}$, with $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$, $\bigcup_{j \in \mathbb{Z}} V_j = L^2(\mathbb{R})$. In addition, $f \in V_{j+1}$ if and only if $f(2 \cdot) \in V_j$, and $f \in V_0$ if and only if $f(\cdot - k) \in V_0$ for all $k \in \mathbb{Z}$. Each subspace $V_j$ is an approximation space of scale $j$, and there exists a scaling function $\phi$ such that the set of all translates $\{\phi_{j,k}\}_{k \in \mathbb{Z}}$, for $\phi_{j,k}(t) = 2^{j/2} \phi(2^j t - k)$, forms an orthonormal basis for $V_j$. 

[Figure 2 about here.]
Let $P_j$ denote the projection from $L^2(\mathbb{R})$ onto $V_j$. Then for $f \in L^2(\mathbb{R})$, one can write $P_{j-1}f = P_jf + g_j$ for some detail function $g_j$. This function, $g_j$ can be written in terms of the dilations and translations of some function $\psi$, so that

$$P_{j-1}f = P_jf + \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k}, \quad d_{j,k} = \langle f, \psi_{j,k} \rangle,$$

where $\langle f, g \rangle$ denotes the standard inner product on $L^2(\mathbb{R})$. The function $\psi$ is a wavelet and the corresponding wavelet family $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$ forms an orthonormal basis for $L^2(\mathbb{R})$. The wavelet coefficients $d_{j,k}$ define the discrete wavelet transform (DWT) $W: f \mapsto d_{j,k} = \langle f, \psi_{j,k} \rangle$.

In conjunction with the approximation spaces, $V_j$, the detail spaces are defined as $W_j := \text{span}\{\psi_{j,k} : k \in \mathbb{Z}\}$, so that $V_{j-1} = V_j \oplus W_j$, where $W_j$ is the orthogonal complement of $V_j$ in $V_{j-1}$. Finally, denoting by $Q_j$ the projection onto $W_j$, we can write for any $f \in L^2(\mathbb{R})$

$$P_{j-1}f = Q_jf + P_jf. \quad (2)$$

Setting $S_j := P_jf$ and $D_j := Q_jf$ makes explicit the decomposition that is described in equation (1). Heuristically, the scaling functions extract local averages in $f$, while the wavelet functions extract local differences in $f$ at given scale.

Figure 1 illustrates how the events in different mass spectrometry spectra are rarely aligned precisely, and so it is desirable to have translation invariance in the sense that if $f_\tau$ denotes translation of $f$ by $\tau$ ($f_\tau(t) := f(t - \tau)$), then $Wf_\tau(j, t) = Wf(j, t - \tau)$. Although this property does not hold for the DWT, it does hold for the continuous wavelet transform (CWT) which is defined by
using the family of all translations and positive dilations of $\psi$:

$$Wf(s, t) = \int_{\mathbb{R}} f(u) \frac{1}{\sqrt{s}} \psi \left( \frac{u - t}{s} \right) \, du, \quad f \in L^2(\mathbb{R}).$$

The DWT is a subsampling of the CWT in both scale and time: $d_{j,k} = Wf(2^j, 2^j k)$. By subsampling the CWT only at dyadic scales $s = 2^j$ while retaining all times $t$, we obtain the translation invariant wavelet transform (TIWT): $W^{TI}f(j, t) := Wf(2^j, t)$. This eliminates alignment artifacts that arise from the DWT subsampling. In particular, shifting the data by any amount will shift the TIWT detail and smooth functions by the same amount. See Percival and Walden (2000) for an illustration of the DWT’s failure in this regard. See also Mallat (1999) for an expanded discussion of these concepts.

3. Scale-\textit{j} Features

The analysis of MALDI spectra typically focuses on the concept of a “peak,” but due to the high degree of irregularity there is ambiguity about what the term means. Even in the case of a relatively smooth spectrum, one might ask whether a “bump” on the “shoulder” of a “peak” is a “feature” rather than “noise”. All of these terms are ambiguous so something more precise is needed before attempting to quantify features.

The definition of a scale-\textit{j} feature formulated here focuses on local maxima in $D_j$. Since the set of all local maxima in $D_j$ may include changes in $X$ on a scale smaller than $j$, this definition restricts attention to changes that occur on a domain proportional to the scale.

Although the concepts of a derivative and local maximum are ambiguous for the spectra studied here, a wavelet transform can be used to compute—in a stable, well-defined manner—the $n$th-order derivative of an averaging of a
signal over a domain proportional to a given scale. If the wavelet $\psi$ has $n$ vanishing moments (i.e., $\psi$ is orthogonal to polynomials of degree less than $n$) and has fast decay, then the continuous wavelet transform $Wf$ of $f$ has the property that there exists a function $\theta$ so that

$$Wf(s,t) = s^n \frac{d^n}{dt^n} (f * \bar{\theta}_s)(t)$$

where ‘$*$’ denotes convolution and $\bar{\theta}_s(t) = s^{-1/2} \theta(-t/s)$ (and $\psi = (-1)^n \frac{d^n}{dt^n}$); see Mallat (1999). The convolution $f * \bar{\theta}_s$ acts as a weighted average of $f$ with a kernel dilated by the scale $s$. We apply this concept to the TIWT and hence $W^{TI}f(j, t) = Wf(2^j, t)$ is the derivative of $f$ averaged in the neighborhood of $t$ with a kernel dilated by $2^j$. Hereafter, the TIWT of a signal $f$ is simply denoted $Wf$.

With the goal of detecting singularities and characterizing lipschitz regularity in $f$, S. Mallat defines a modulus maximum of $f$ to be a point $(s_0, t_0)$ at which $|Wf(s_0, t)|$ is locally maximum at $t = t_0$, for a given scale $s_0$ (Mallat and Hwang, 1992). To define scale-based features, we borrow this concept and apply it to the detail functions since they reflect scale-based changes in 

Definition. Let $X$ be a spectrum with a multiscale decomposition given by (1). For $j = 1, ..., J$, a scale-$j$ feature of $X$ (with respect to this decomposition) is a scale-$j$ modulus maximum in the detail function, $D_j$.

Note that the set of scale-$j$ features does not correspond to the set of local maxima in $X$, but corresponds to local changes in $X$, of scale $j$, as extracted by $D_j$. Hence it includes inflections or “shoulders” (see Figures 2 and 3). The locations of these (scale-based) local extrema are easily calculated via TIWTs.
using wavelet families having one and two vanishing moments; denote these by $W^1 D_j$ and $W^2 D_j$, respectively. In practice, these must be shifted in order for their events to align with the events in $D_j$ (see Percival and Walden (2000) for specifics); for clarity we will use the same notation to denote these shifted functions. Figure 3 illustrates this with graphs of $D_5$ and the (shifted) TIWTs of $D_5$ using both the Haar and Daubechies-4 wavelets: $W^1 D_5$ and $W^2 D_5$.

The existence of a scale-$j$ feature at a location $t$ is not defined in terms of the intensity of the signal at $t$. Rather, it depends on a relative change in the intensity over a region whose width depends on the scale, $j$. As such, the existence of a feature is not determined by an estimate of the signal-to-noise ratio. In the bottom of Figure 3 $X$ is superimposed on a noisy version, $X + \epsilon$, $\epsilon \sim N(0, 40^2)$. The median absolute deviation estimate for $X$ is 16.7 ($\hat{\sigma}_{MAD}$ estimated using scale-1 TIWT coefficients (Percival and Walden, 2000)) compared to 44.4 for $X + \epsilon$. For perspective on the influence that a large signal-to-noise ratio may have on the identification of feature locations, we generated 1000 noisy versions of $X$ ($\epsilon$ as above) and calculated scale-5 feature locations for each. A histogram of these locations is plotted at the bottom of Figure 3.

4. Histograms and feature locations

A histogram of feature locations from all samples is used to determine the regions in which a feature is deemed present in the data: the location of each scale-$j$ feature is recorded from each sample and a histogram of these locations is constructed from all samples. A portion of the histogram of scale-
5 feature locations from all 220 samples in the dietary study (Mitchell et al., 2005) is shown at the bottom of Figure 4. The histogram serves as a spectral-density-like function used to aggregate features across spectra into bins in which features from different spectra are matched. These bins are delimited by the tickmarks at the bottom of this graph.

[Figure 4 about here.]

The histogram in the bottom of Figure 4 is superimposed with a smoothed version, the local minima of which are used to delimit the bins. Since the goal is to delimit clusters of scale- feature, a local averaging of the histogram on a scale no larger than is performed using a scaling filter: if \( H_j \) denotes the histogram of scale- features, calculate \( S_{j-1} = P_{j-1} H_j \), as in (2). The bins are determined by finding the local minima in this smoothed version of the histogram, the minima occurring at the edges of the bins.

Although a scale- feature in a spectrum is unambiguously defined, the formation of bins for a set of spectra is subject to the method used for approximating the scale- histogram. We have, in a sense, transferred the problem of estimating a signal-to-noise ratio in individual spectra to that of estimating which regions of a histogram correspond to signal, as determined by the data set. The approach is loosely analogous to a Fourier analysis which uses estimates of a power spectrum to identify frequency content of a signal.

4.1 An illustrative example.

From Figure 4 one might infer the existence of peptides of approximately 7700, 7725 and 7750 Thompson based on the high density of features in these three regions. Whether this is accurate is unknown for these spectra so we briefly consider a set of spectra for which such knowledge is available.
These come from an experiment in which various concentrations of a solution containing known peptides were added to human serum samples (Randolph et al., 2005). Halving the concentration through ten successive stages produced spectra which, at the lowest concentration, contain little or no visual evidence of these peptides. Five replicate spectra were collected from each concentration of peptide mix as it was spiked into five different human serum samples producing 25 spectra per concentration. One of the peptides in this mix is bovine insulin (5734 Thompson, shown at \( t = 13430 \) in Figure 5).

[Figure 5 about here.]

To provide a general idea about content in this data set we calculated the mean at each TOF value, \( t \), for a variety of subsets. Specifically, denoting a set of spectra by \( \{X_i\}_{i=1}^N \), its mean spectrum is the average of the intensity measurements at each \( t \): \( M(t) = \frac{1}{N-1} \sum X_i(t) \). Figure 5(a) exhibits a portion of mean spectra for four subsets, nested according to concentrations: (i) the lowest two concentrations (50 spectra); (ii) concentrations 1 (lowest) through 5 (125 spectra); (iii) concentrations 1 through 7 (175 spectra); (iv) concentrations 1 through 9 (225 spectra).

The fact that the mean intensities at \( t = 13430 \) grow in correspondence with the concentration of spiked-in peptide mix suggests that the high-density region of the histogram near 13430 in Figure 5(b) does, indeed, record evidence of a protein-related feature. The feature to right of this, at 13470, presumably represents a native (not spiked in) peptide as it exists in all of the spectra from samples with nothing added as well as in the spectra from samples receiving a low concentration of peptide mix. A detailed analysis of these and other properties of this dataset appears in Randolph et al., 2005.
5. Other approaches to processing spectra

The proposed method of processing low-resolution mass spectrometry data is one attempt to overcome the obvious shortcomings of a pointwise analysis of intensities (i.e., a comparison of the values \( \{X_i(t)\}_{i=1}^{N} \) at each TOF value \( t \)) which overlooks the lack of consistent alignment inherent in this data and ignores any local structure in the spectra. Other approaches to addressing this problem are discussed briefly here.

5.1 Local differencing

A “Simple Peak Finding” (SPF) algorithm proposed and used by Coombes et al. (2003) calculates first differences between successive time points and uses these to locate local maxima and minima. It defines noise as the median absolute value of the first differences, retains local maxima whose distance to the nearest local minimum is less than the noise, and then defines as features those local maxima that are separated by fewer than \( \tau \) time points, where \( \tau \) depends on the \( m/z \) value of the feature.

This is loosely related to the proposed method since first differences between successive time points could be calculated via scale-1 wavelet coefficients using a wavelet family with one vanishing moment. The SPF definition of noise in a spectrum is the median absolute deviation often calculated via these wavelet coefficients (Percival and Walden, 2000). In contrast to the proposed method, the SPF algorithm defines features based on a single scale of local differences and estimates a noise level for each individual spectrum.

5.2 Direct identification and alignment of local maxima

In a simple approach of Yasui et al. (2003) (also Yasui et al. (2004) and Tibshirani et al. (2004)), peaks are defined by local maxima in a set
of $\pm N$ neighboring points. A refinement of this definition requires the local maximum to have intensity above a given threshold, $R$, determined by a signal-to-noise ratio. Here, noise is defined by successive application of the median smoother: calculate a median-smoothed intensity with a smoothing window of $\pm K_1$ points, then calculate a median-smoothed deviation (i.e., the noise level) from the smoothed intensity using a window of $\pm K_2$ points. The definition of a peak is a function of the parameters $(N, R, K_1, K_2)$ which the analyst can adjust according to the data and analysis goals.

After peaks are identified in each spectrum, they are aligned across spectra so that they can be compared. For this, Yasui et al. (2004) propose a method of successive peak-alignment using a window whose width corresponds to the imprecision of the $m/z$ measurements (usually $\pm 0.1$-$0.2\%$ of $m/z$). In Tibshirani et al. (2004), each MALDI spectrum is first smoothed and then peaks are aligned by an application of complete hierarchical clustering where any two peaks in each cluster lie within a specified window. Both approaches require a specification of the imprecision for the width of the $m/z$ window.

This local-maxima method was designed to achieve high sensitivity (few false negative peaks) while allowing an appreciable compromise in specificity (more false positive peaks). Indeed, it was developed jointly with a classification method for which it serves as a pre-analysis processing step. Emphasis is on sensitivity in the pre-analysis processing step since, in the classification step, appropriate methods for selecting peaks with high discriminatory power from a large set of peaks are employed.
5.3 **Smoothing versus feature extraction**

Although wavelets are often used for smoothing and denoising signals, the process of smoothing spectra and subsequently locating peaks differs from the proposed method of identifying scale-$j$ features. A wavelet smoothing procedure begins with a choice of wavelet family followed by a choice of thresholding method, each method involving parameters that control the amount variation (see, for example, the comprehensive survey of Antoniadis et al. (2001)).

One might use wavelet smoothing with the goal of recovering “visible features” in the mass spectrometry, but one must first define these terms and then choose accordingly from a myriad of methods. This is done in Morris et al. (2005) by calculating the mean spectrum which is denoised using the translation-invariant wavelet transform then defining a peak by the range between flanking local minima around a local maximum.

Our proposed point of view is that extracting scale-based signal content (defined by local differencing, not local averaging) and using the frequency-domain-like histograms to determine high-density regions (bins) of feature locations allow for a focus on features that distinguish themselves across all spectra. Figure 5 illustrates a potential advantage to locating features prior to averaging across spectra or smoothing: the peak at 13470 is drowned out in the mean spectrum and choosing smoothing parameters that capture the wide range of feature intensities may be difficult.

For additional perspective on feature detection versus signal reconstruction, see Mallat and Hwang (1992) where the use of modulus maxima in the detection and measurement of events in signals and images is separated from the process of denoising.
6. Registration of spectra

Since some spectra may be substantially shifted relative to other spectra we describe a simple alignment procedure which begins by finding features that are prominent and consistent across spectra. This precedes the process of identifying subtle features and can increase the resolution of the histograms. Although a variety of methods could be used for performing curve registration, we choose not to attempt any subtle matching of features since the true source of features (i.e., specific peptides) is unknown. Note also that all processing is done on the TOF measurements since a registration performed on the $m/z$ axis would have to adjust not only for instrument offset, but also the quadratic warping that results from a standard calibration (recall, $m/z = (at + b)^2$ for fitted $a$ and $b$).

Choosing a relatively coarse scale, $j_r$, we say that a spectrum $X$ has a prominent scale-$j_r$ feature at $t$ if $D_{j_r}(t) > \tau$, for some threshold $\tau$. This property is independent of the baseline and is relatively insensitive to the choice of $\tau$, for $\tau$ in a mid-percentile range of detail function values. If several prominent features in a spectrum are unilaterally shifted up or down the $t$ axis relative to the centers of their scale-$j_r$ bins, then this spectrum is shifted (globally) to align these features more closely with the bin centers.

[Figure 6 about here.]

For example, within a scale-6 bin, $B_k = (7500, 7570]$, if spectrum $X$ has a scale-6 feature location at $t_k = 7520$, a within-bin feature location of $t_k^* = -15$ is recorded since $t_k$ is 15 units to the left of the center of bin $B_k$. The median value, $\bar{t} = \text{med}_k\{t_k^*\}$, among all these within-bin prominent-feature locations
determines the amount that $X$ is to be shifted to produce its registered version, $X^r$:

$$X^r(t) := X(t - \bar{t}), \quad 0 \leq t \leq T.$$  

Figure 6 exhibits ten randomly chosen spectra before and after performing the registration procedure on the set of 220 spectra. Prominent features were defined by scale $j_r = 6$ and a threshold of $\tau = 50$ (resulting in about 6% of the 350 scale-6 features being prominent in each spectrum). The locations of scale-5 features before and after alignment are also shown. Registration was performed globally and although only a small region is shown, the performance is similar in all regions. For additional perspective the scale-4 histogram is also shown.

7. Quantifying intensity.

Signal intensity has not played a direct role in the definition of features or their alignment/grouping across spectra and so quantifying the intensity of a feature has not been mentioned. This typically involves both normalization and baseline correction. The former is done to adjust for differences in total ion content between spectra; we will assume this has been done (see, for example, Yasui et al. (2003)). The latter aims to estimate a baseline contained in the spectrum so that intensity of a peak can be quantified by its height above this baseline. The baseline is typically estimated using values in a window that contains the peak; for example, using loess (Yasui et al., 2003) or an interpolation in the window after the peak is removed (Coombes et al., 2003).

Having identified features without modelling a baseline, it is natural to quantify intensity similarly. In particular, recall that a scale-$j$ detail function,
\(D_j\), is based on local changes and contains no trend. Therefore, define the intensity of a scale-\(j\) feature in \(X\) located at \(t_k\) (in bin \(B_k\)) as \(I_X(k) := D_j(t_k)\). If \(X\) has no scale-\(j\) features in bin \(B_k\), set \(I_X(k) = 0\). If \(X\) has multiple scale-\(j\) feature locations in \(B_k\), use the largest intensity in \(B_k\): \(I_X(k) = \max\{D_j(t) : t \in \{t_{k_1}, ..., t_{k_\ell}\}\}\), where \(t_{k_i}\) denote feature locations in \(B_k\). Although the value of \(D_j\) at a feature location is not an accurate measure of peptide abundance, it may provide a fairly robust quantification of relative intensity (Randolph et al., 2005).

7.1 Summary of the procedure

For each spectrum \(X\), form the multiscale decomposition given in (1). For scale(s) \(j \in \{1, ..., J\}\) calculate:

i) **Scale-\(j\) feature locations.**

- Using TIWTs with wavelets having 1 and 2 vanishing moments (\(W^1\) and \(W^2\), respectively), form the \(j\)th scale of the circularly shifted transforms of \(D_j\): \(W^1_j D_j\) and \(W^2_j D_j\)
- Feature locations: set \(T_j(t) = \begin{cases} 1, & W^1_j D_j(t) = 0 \text{ and } W^2_j D_j(t) > 0 \\ 0, & \text{otherwise} \end{cases}\)

ii) **Scale-\(j\) histogram and bins.** Each spectrum \(X_i\) \((i = 1, ..., N)\) now has a corresponding vector of scale-\(j\) feature locations \(T_{ij}\).

- Sum across samples to form a histogram: \(H_j = \sum_{i=1}^{N} T_{ij}\)
- Use local minima in a smoothed version of \(H_j\) to determine bins \(B_k, k = 1, ..., K\), in which the scale-\(j\) feature locations from each spectra are recorded.
iii) **Intensity of scale-\(j\) features within each bin.** Within each bin, \(B_k\), record an intensity \(I_X(k)\) for spectrum \(X\) based on its signal content at the scale-\(j\) feature locations \(t_k \in B_k\).

0) **Optional Registration.** Begin by performing steps (i) and (ii) on a single coarse scale, \(j_r\). Use a threshold \(\tau\) to locate prominent features, \(\{t \in T_{j_r} : D_{j_r}(t) > \tau\}\). Shift each spectrum by an amount equal to the median offset \(\bar{t}\) of the prominent scale-\(j_r\) feature locations from the centers of their bins: \(X^T(t) = X(t - \bar{t})\). Using this set of registered spectra, proceed with (i)–(iii).

The occurrence of a single spectrum having multiple features in a single bin may arise in regions of low intensity and, potentially, minimal or no signal. This suggests a second iteration of step (ii): recompute the density histogram and bins based only on features which do not share a bin with another feature from the same spectrum. The heuristic here is that multiple scale-\(j\) features within a scale-\(j\) bin indicate “scale-\(j\) noise.” This was used to produce Figure 4.

8. **Remarks on implementation**

For all figures, the Haar wavelet was used to produce scale-\(j\) detail functions, \(D_j\), for each spectrum. Scale-\(j\) features were located using Haar (one vanishing moment) and Daubechies-4 (two vanishing moments) wavelets. Although the Haar wavelet is not well-adapted for approximating smooth functions, the fact that it extracts first-order changes and has minimal support makes it the natural choice for the purposes used here (see Section 5.3). To produce the local-averaging smooth of the scale-\(j\) histogram \(\mathcal{H}_j\), the LA(8) scaling
function of scale \( j - 1 \) was used. For definitions of these wavelet families, see Percival and Walden (2000). All calculations and graphics were done using Matlab (Mathworks Inc., 2004) and the (freely available) WMTSA Wavelet Toolkit (Cornish et al., 2003). Code is available at www.tibs.org/biometrics.

With regard to the choice of scale, features corresponding to higher mass peptides may appear at larger scales than low mass peptides and so analysis may benefit from consideration of additional, or a combination of, scales. However, many scales are less informative: scale \( j = 10 \), say, corresponds to changes in spectra that occur across 1024 TOF measurements which, depending on the resolution of the instrument, may be too wide to capture localized peptide information. On the other hand, scales \( j = 1 \) or 2 are potentially finer than the accuracy of the detector and, consequently, exhibit properties similar to that which results from a set of pure noise spectra. The histograms at various scales suggest that most relevant features are described by a small subset of scales. This is consistent with the laboratory-based study (Randolph et al., 2005). Note also that features are manifest across a range of wavelet scales (see Figure 2) making the identification of features less sensitive to the specific choice of scale than if this property were absent.

9. Discussion

The utility of mass spectrometry spectra in laboratory or clinical studies relies on a useful quantification of the data. Specifically, what features indicate the existence of a peptide? How can these be extracted it from a set of spectra in which there is significant between-sample variability in background noise, the measurement of intensity and in the \( m/z \)-value at which a feature is recorded? The point of view taken here is that signal content in mass
spectrometry data can be identified and quantified without first smoothing the spectra, estimating signal-to-noise ratios or modelling a baseline. Some a priori information about the signal of interest exists—it is not periodic, it is localized within a range of bandwidths and local changes such as peaks and shoulders are relevant—and leads to an unambiguous definition of “features” to be quantified. This approach is not claimed to be optimal with respect to biological content, but some definition is required before rigorous analysis can proceed. These ideas are an attempt to do this while minimizing the number of the modelling assumptions. Questions regarding the precise nature of peptide-induced signal must be ultimately be addressed by laboratory-based study, possibly in conjunction with computer-based analysis such as a simulation tool that generates virtual spectra based on the underlying physics of the instrument (Morris et al., 2005).

10. Acknowledgements

The authors thank D. McLerran, Z. Feng and M. McIntosh for useful discussions. Author T. R. was supported by National Institutes of Health grant K25-GM67211 and Y. Y. was partially supported by Grant U01CA086368 from the National Cancer Institute, Grant P01CA053996 from the National Cancer Institute, and the Canada Research Chair Program.

References


Sample classification from protein mass spectrometry by peak probability contrasts. *Bioinformatics Advance Access* June.


Figure 1. Sample of a MALDI-TOF mass spectrometry spectrum, and a closer view of it along with three other samples. (Spectra are vertically separated for display.)
Figure 2. A simulated signal, $Y$, superimposed on a noisy version, $X = Y + \epsilon$, along with detail functions $D_j$, $j = 1, \ldots, 5$, from a multiscale decomposition of $X$. 
Figure 3. The top figure shows one spectrum, $X$; underneath it are graphs of its $D_5$ detail function (darkest curve) along with $W^1 D_5$ and $W^2 D_5$ (dotted curve). The 20 tickmarks at the bottom indicate locations of the scale-5 features. The bottom figure shows $X$ superimposed on a noisy version, $X + \epsilon$, $\epsilon \sim N(0, 40^2)$. Also shown is a histogram of feature locations of 1000 noisy versions of $X$. The tickmarks at the bottom reproduce those in the top figure.
Figure 4. The top graph shows four spectra. The four sets of tickmarks indicate the scale-5 feature locations in each. The bottom graph is a histogram of the scale-5 feature locations from the entire data set. The vertical lines delimit the bins in which the scale-5 features are grouped as determined by scale-based local minima in a locally-averaged histogram (dark curve).
Figure 5. (a) Mean spectra for four nested subsets of spectra from a dilution experiment: (i) concentrations 1 (lowest) and 2; (ii) concentrations 1 through 5; (iii) concentrations 1 through 7; (iv) concentrations 1 through 9. (b) Histogram of scale-5 feature locations of 225 spectra from varying concentrations of a peptide. The horizontal axis is indexed by raw time-of-flight.
Figure 6. (a): Ten spectra prior to registration with their scale-6 feature locations (tickmarks) and the scale-6 bins (vertical lines); (b): Same ten spectra after registration now shown with scale-5 bins (vertical lines) and their scale-5 feature locations (tickmarks). The lightest tickmarks at the bottom indicate the scale-5 feature locations prior to registration; (c): Histogram of scale-4 feature locations from all 220 spectra.