Quantifying Peptide Signal in MALDI-TOF Mass Spectrometry Data*

Timothy W. Randolph‡§, Bree L. Mitchell¶, Dale F. McLerran||, Paul D. Lampe¶¶, and Ziding Feng

This study addressed the question of which properties in MALDI-TOF spectra are relevant to the task of identifying mass and abundance of a peptide species in human serum. Data of this type are common to biomarker studies, but significant within- and between-spectrum variabilities make quantifying biologically induced features difficult. We investigated this signal content and quantified the existence, or lack, of peptide-induced signal (as manifest in a multiresolution decomposition) by generating spectra from human serum in which the abundance of peptides of specific masses is controlled by a sequence of dilutions. The intensities of the corresponding features were directly proportional to peptide concentration. The primary goal was to exhibit some quantifiable properties of raw spectra from this application of MALDI-TOF mass spectrometry. Although no recommendations are given regarding the best method for processing these data, the results confirm the utility of a simple method, based on wavelets, for defining and quantifying features related to low abundance peptide species in a heterogeneous set of complex spectra. Estimates on lower limits of detectable peptide abundance (in the 20-nmol range) and on the number of features present in a spectrum are made possible by the controlled experimental design, the use of a large external reference data set, and dependence on relatively few modeling assumptions. Molecular & Cellular Proteomics 4:1990–1999, 2005.

Despite the apparent success of some early biomarker studies based on mass spectrometry TOF data, the use of this technology for biomarker discovery remains controversial (1–3). Some of the debate revolves around the issue of signal content and which properties of a spectrum are relevant for use in classifying samples in case/control studies. There is, at present, no consensus on a preferred method for distinguishing signal from noise or on which properties of the spectra are truly relevant for use in inferring peptide mass and abundance. This stage of analysis (often referred to as preprocessing) is important as it impacts all subsequent analysis of these data (4).

Peaks in a spectrum are typically used as indications of peptide content in the sample with relative abundance of a particular species of (ionized) peptide corresponding to the height, or volume, of a peak. Identifying peaks (or overlapping peaks) and interpreting intensities is complicated by high frequency "noise" along with global and local trends. In addition to these within-spectrum variabilities, a collection of spectra typically exhibits substantial between-sample heterogeneity in noise level and baseline intensity as well as the existence or nonexistence of biological features. Moreover there is a variable amount of mass shifting inherent in the spectrometer that results in the property that features from different spectra rarely align at exactly the same TOF value.

Studies related to biomarker discovery using mass spectrometry TOF data have been based on a variety of assumptions regarding signal-to-noise ratio, spectrum base line, and the offset of features between spectra (5–8). The analysis in Baggerly et al. (9) moreover reveals a variety of non-biological content in a set of spectra from a MALDI-TOF spectrometer. For a brief survey on current approaches to processing and analyzing TOF data, see Coombes (10).

Because of the potentially rich spectrum obtained from a single serum sample, the methods and techniques used, from the first stages of sample collection to the final stages of data analysis, are likely to have considerable impact on the outcome. The aim in this study was to provide a benchmark for understanding which of this information is biologically relevant. We emphasize that this is not purely a laboratory question: what is found in the data is dependent on the assumptions made when processing the data. One extreme is to impose strict assumptions on the amount of offset between peaks from different spectra, on modeling a global base line, on window widths in which peaks are sought, and on noise content and the amount a peak must exceed the noise before it is recorded. Tacit in these assumptions is that the notion of a peak is formally defined, but such a definition is not obvious for MALDI-TOF spectra. The other extreme is to accept every TOF measurement as informative and use all (e.g., 50,000 or more) unadjusted intensities in a comparison of samples.

In investigating properties in MALDI-TOF spectra that are relevant to the task of identifying mass and abundance of a peptide in a complex serum sample, our goals were 2-fold: 1) to exhibit the utility of a method for defining and quantifying peptide-related features that imposes minimal assumptions
about noise, base line, alignment, or normalization and yet provides an informative summary of peptide content in a heterogeneous set of spectra and 2) to provide a frame of reference for researchers who collect, process, and analyze data of this type by probing the limits of detection of biologically related signal.

For this, we produced and analyzed a set of spectra from human serum samples in which peptides of known masses were added in controlled amounts. At the highest concentrations, peaks from the added peptides were obvious, whereas at the lowest concentrations no visual evidence in the spectra distinguished these mass values. Positive identification and quantification of low signal features appearing in a small percentage of the spectra was made possible by the fact that the intensity of signal induced by added peptides varied directly with the concentration of the added solution.

MATERIALS AND METHODS
Specimen Source, Sample Preparation, and Equipment

Source of Serum Samples—Five serum samples were chosen at random from serum collected as part of a randomized controlled exercise intervention study conducted in the greater Seattle area. This study (referred to as APPEAL) enrolled 200 healthy men and women aged 40–75 years with no history of invasive carcinoma who exercised less than 30 min three times per week and did not have contraindications to exercise. Participants were randomized to either a delayed exercise intervention arm where they maintained their normal activity level for 12 months and then were given a 2-month exercise intervention (non-exercisers) or to a 12-month exercise intervention (exercisers). Five replicate spectra from serum samples taken at baseline, 3 months, and 12 months for all participants produced (after some number of unreadable spectra and participant dropout) a total of 2903 spectra. These spectra were collected several months prior to the present dilution study and were used here to provide a frame of reference external to this study.

Serum Preparation—The samples were thawed, and abundant globular proteins were precipitated from a 20-μl aliquot by the addition of an equal volume of acetonitrile. Acetonitrile-treated samples were mixed at room temperature for 30 min using a multispeed vortexer at the lowest setting. Precipitated proteins were pelleted by centrifugation at 12,000 rpm for 4 min at room temperature; the supernatant was removed and stored on ice for analysis. Previously we have investigated multiple methods for prepurification and found acetonitrile precipitation to be superior for our purposes (11).

Peptide Mixture Preparation—Calibration Mixture 2 (Applied Biosystems), consisting of angiotensin I (0.065 mg/μl), adrenocorticotropic hormone (ACTH)1 clip-(1–17) (0.105 mg/μl), ACTH clip-(18–39) (0.093 mg/μl), ACTH clip-(7–38) (0.275 mg/μl), and bovine insulin (0.0502 mg/μl), was mixed in equal volume with cytochrome c (Sigma) at a concentration of 1.24 mg/μl. This mixture was diluted 1:10 with water followed by nine serial dilutions of the 1:10 standard mixture. One microliter of each dilution was mixed with 1 μl of each serum sample.

Each sample–standard mixture was spotted in quintuplicate yielding 250 spots (5 subjects × 10 dilutions × 5 spots each). In addition, each serum sample was spotted in duplicate without the addition of any standard peptide, and each standard dilution was spotted in duplicate without being added to serum. This gave a total of 280 samples whose spot positions were randomized on the plate. Because of one spotting error, one peptide-added sample was spotted on top of a serum-only sample resulting in 278 usable spots.

Mass Spectrometry—A 384-well hydrophilic MALDI-TOF plate was precrystallized with sinapinic acid matrix, a supersaturated solution of 4-hydroxy-3,5-dimethoxybenzylamine acid in acetonitrile, 0.1% TFA (50:50, v/v). Samples were mixed with sinapinic acid in a 1:3 ratio and spotted onto the plate, leaving empty all wells that border the plate edge.

The plate was analyzed in an Applied Biosystems Voyager mass spectrometer in linear mode using a 337 nm nitrogen laser. Settings were optimized using cytochrome c and were as follows: laser intensity, 1784; accelerating voltage, 25,000; grid voltage, 23,000; guide wire voltage, 1250; delay time, 325 ns; low mass gate, 1000 Da; molecular mass range, 1000–40,000 Da. Final spectra were generated by summing spectra produced from 20 laser hits at up to 20 places per spot (a total of up to 400 laser hits per summed spectra). Individual spectra had to have an intensity of 5000 hits or more to be used in the summed spectra.

The term “spectrum” refers to the graph of a function consisting of intensities (vertical axis) corresponding to ~50,000 sampled time-of-flight measurements (horizontal axis). The intensity indicates a (relative) abundance of ionized particles in the biological sample (e.g., tissue, serum, or plasma) having the corresponding time-of-flight. The time-of-flight measurements are typically converted to mass/charge values that are subsequently used to identify peptide content in the sample.

Reference Data—Table I gives the theoretical mass for each peptide that was considered in this study along with the reference time-of-flight values used in studying them. Throughout we focused on TOF, or clock tick, measurements, denoted by t. These are obtained directly from the spectrometer rather than using a calibrated axis of mass/charge ratios (denoted m/z). The reason for this is that converting from t to m/z would separate our analysis from the true measurements of the instrument by imposing a physical modeling assumption about how t relates to m/z and by approximating this relationship with a least squares fit to a calibration spectrum (i.e., estimating a and b in the model m/z = (at + b)/t, t > 0). In particular, any calibration used to produce an m/z axis from the TOF measurements merely provides a reference standard rather than absolute truth for measuring peptide masses. Therefore, all analysis was done directly on the spectrometer output of TOF measurements and unadjusted intensities; no base-line correction, denoising, normalization, or curve registration were performed. Although the peptide mixture contained six known peptides, we did not pursue properties of either angiotensin I or ACTH clip-(1–17) because they occurred in the low mass region where the spectra contain a substantial amount of matrix-contaminated signal.

Definitions: Feature and Intensity

We worked with the TOF measurements rather than calibrated m/z because, for the reasons mentioned above, it is not possible to create a perfect calibration of true m/z values. We took as a reference TOF the median of the t values at which the high concentration spectra (dilutions 1–4) obtain a maximum in a window of width 30 t that contains the visually apparent peak. This crude method of identifying the location of a peak is highly dependent on the knowledge that a local maximum in the spectrum exists in the specified interval and is the absolute maximum in this interval. However, we used this to create a “Reference TOF” value for each of the spiked-in peptides (Table I) because these prominent peaks clearly existed in the high concentration spectra, and locating their TOF values in this way is straightforward and may have the least potential for bias. When seeking more subtle features in the spectra this method breaks down. In particular, because of the high degree of irregularity of these

---

1 The abbreviation used is: ACTH, adrenocorticotropic hormone.
spectra, local maxima occur in intervals of nearly every size and location so the term “peak” is ambiguous. Therefore, a definition of feature that is more comprehensive than local maximum is needed. For this, features can be defined in terms of local changes that occur at a given scale. Specifics of this simple procedure are outlined under “Spectrum Decomposition.”

Spectrum Decomposition—Motivation for defining scale-based features comes from the fact that a spectrum \( X(t) \) of length \( T \) can be decomposed into a sum of constituent functions, \( D_j \), referred to here as ‘detail’ functions, each containing events that occur at a particular scale: at the finest scale, \( D_1 \) is the result of extracting changes in \( X \) that occur across a \( 2^1 \)-unit domain. Writing \( S_j = X - D_j \) expresses the approximation after \( D_1 \) is removed. Similarly changes in \( X \) that occur across a \( 2^2 \)-unit domain give rise to the scale \( j \) detail function \( D_j \).

Continuing through \( J \) steps (\( J \equiv \lceil \log_2 T \rceil \)) results in the decomposition

\[
X(t) = S_j(t) + \sum_{j=1}^J D_j(t), \quad 0 \leq t \leq T \tag{Eq. 1}
\]

Fig. 1 illustrates this idea with a simulated signal and five detail functions, \( D_1, \ldots, D_5 \) (plotted top to bottom) as extracted from a noisy version of the signal.

This decomposition is implemented using a translation-invariant wavelet transformation, \( W \), decomposing \( X \) into \( J \) scales of changes:

\[
d_j(t) = \sum_{m} W_j(t - m), \quad m \in \mathbb{Z}
\]

where \( d_j(t) \) is a scale \( j \) wavelet coefficient for \( X \) at time \( t \), i.e., \( d_j(t) = WX(t^j, t) \). The inverse transformation applied individually to each scale of the transformation of \( X \) gives rise to the detail functions \( D_j, j = 1, \ldots, J \). These concepts can be found in most references on wavelet analysis (see e.g. Ref. 12) and are outlined, specifically for the present purposes, in Randolph and Yasui (13).

Because a detail function, \( D_j \), arises from scale \( j \) changes in \( X \) (extracted via \( d_j \)), our interest was in the local maxima in \( D_j \). Actually our interest was in scale \( j \) local maxima rather than every blip in \( D_j \), and these are identified by finding the modulus maxima in \( D_j \). A scale \( j \) modulus maximum of a function \( f \) occurs at a point \( t_0 \), at which \( |Wj(t_0^j, t)| \) is a local maximum as a function of \( t \). The definition of features we used in studying these data is as follows.

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

This is illustrated in Fig. 2. The bold curve at the bottom is the detail 5 function, \( D_5 \), for a dilution 6 spectrum (bold jagged curve). The locations of the scale 5 features for this spectrum are identified by the tick marks at the bottom of the figure. Also shown are graphs of spectra from dilutions 4 through 8 from one of the serum samples. For display purposes, the curves were separated vertically.

To quantify the intensity of a feature we simply took the size of the scale \( j \) detail function at the feature location. In Fig. 2, this is the height of the smooth lower curve as measured at each tick mark. The goal in this study was not an accurate measure of abundance of the corresponding product in the serum but to quantify the relative size of a scale \( j \) feature for comparison between the various dilutions. Formally the scale \( j \) intensity of a scale \( j \) feature at \( t = t_j \) in a spectrum with a multiscale decomposition given by Equation 1 is the value \( D_j(t_j) \). A summary of the computational steps is given under “Appendix: Computational Steps.”

**Choice of Scale**—In general, the scale(s) of interest will depend on the resolution of the spectrometer and how the relevant content appears in the signal. In these data, at 2500 \( m/z \) there are approximately three TOF measurements per \( m/z \) and approximately one measurement per \( m/z \) at 25,000 \( m/z \). With higher resolution spectrometers, there may be 10-100 measurements per \( m/z \) in which case coarser scales become more relevant to characterizing peptides (although finer scales may also be used to characterize isotopic distributions). In any case, there are clearly scales that are less relevant to the analysis of biologically related signal. All subsequent analysis focuses on scale 5 features. This proved sufficient for the goals of this study, but additional analysis using other scales (or their combinations) may, in general, be informative. It is useful to note here that feature information is not restricted to a single scale as scale \( j \) features are typically present in both finer and coarser scales; see Fig. 1. In particular, if a peptide species exists in extremely high abundance so that it is detected across a wide range of TOF values (as in the largest features of Fig. 3) then a single scale will not provide an accurate quantification of its relative abundance.

**Statistical Analysis**—To define the lower limit of detection of a feature that results from the addition of peptide mixture, we consid-
RESULTS

The goal of this study was to detect low intensity peptide-induced features in spectra from complex human sera and to quantify the relative abundance of the peptide species. The concept of "feature" as given under "Definition 1" allowed us to do this without prespecifying a signal-to-noise ratio that would, in effect, predetermine a minimum abundance of peptide that is detectable. Although bypassing alignment and normalization is not recommended for careful analysis in biomarker research, the presentation here represents data direct from the instrument, unimpeded by estimates of noise, base line, curve registration, or normalization. In this study we quantified signal content that reflects local change (of any magnitude) in a spectrum that occurs at a given scale.

A novel property of this data set is that the existence of an added feature (one resulting from a product in the peptide mixture) can be positively confirmed by the property that its intensity varies as a function of the dilution level. This property allows us to distinguish between an added feature and a native feature (one present in the serum prior to adding any peptide mixture). This also makes possible an estimate on the limit of detection for this particular instrumentation, its settings, and the sample preparation. In addition to the peptides in Table I, the spectra exhibited a variety of unidentified products in the added solution; these were result of degradation of known peptides or other contaminants.

In the following summaries and displays subtle features near the detection limit provide information on the limits of this mass spectrometry platform and highlight the potential for processing the data by the methods used here. Focus is on one specific region surrounding an added feature (bovine insulin). The regions surrounding other added features (both known and unknown) exhibited similar properties.

Display of Features—Fig. 3a shows portions of 10 mean spectra, the means taken over 25 spectra at each dilution (pointwise means, $X(t) = 2^{-1} \sum_{25}^{25} X(t)$). These means are used for display only as all processing was performed individually on each of the 250 spectra. In all figures, the horizontal axis is indexed by the raw time-of-flight, or clock tick, measurements; see "Reference Data." Also for display purposes, so that separate curves can be distinguished visually, some have been manually shifted vertically. Therefore the vertical axis in Fig. 3a is not labeled with absolute intensity values, but the tick marks give a reference on relative change in intensity: the distance between tick marks on the vertical axis is 100 arbitrary units as output by the mass spectrometer. The second curve from the bottom in Fig. 3a is the mean over nine serum-only spectra in the dilution data set. Finally for reference, the APPEAL data were analyzed using the same methods; their summaries and properties are plotted alongside those from the dilution data set. To exhibit a mean spectrum for these data, a random sampling of 1000 spectra was used. This appears as the bottom curve in Fig. 3a.

Fig. 3b shows the (raw, not smoothed) density histogram of all scale 5 feature locations for all dilutions combined. Also in

![Fig. 2. Selected spectra from four dilutions for one serum sample: dilution 4 (top jagged curve) through 8 (bottom jagged curve). For dilution 6 (bold jagged curve), the scale 5 feature locations are indicated by tick marks at the bottom of the figure. These indicate modulus maxima in its scale 5 detail function (bold curve at the bottom). The curves have been vertically separated for display.](image-url)
this figure, plotted as a light bold curve, is the density histogram of feature locations for all 2903 spectra in the APPEAL data. We note that this set of spectra, having been collected over many days, benefitted from a crude alignment that consisted of a unilateral shift of each spectrum as determined by an aligning of prominent features (see Ref. 13 for details). The aligned set of spectra is used here.

The vertical lines in Fig. 3, a and b, delimit bins in which features from the set of all 250 spectra aggregate. Bins are labeled A through H for the discussion that follows. Fig. 3c exhibits box plots of the base 2 logarithm of feature intensities from these selected bins (i.e. $\log_2(D_5(t))$ at the feature location, $t$, for each spectrum): one box plot for each dilution (labeled 1 through 10) and one for the APPEAL data (labeled 11). The bottom right panel displays intensities as recorded from the raw spectra, $X(t)$, at the feature locations in bin H.

Regional Properties—Bin F in Fig. 3a contains the bovine insulin feature that exists at various dilutions, while in bin F of Fig. 3b the density histogram (dark thin curve) shows its prevalence in all 250 spectra. In neighboring bin G is a native
feature that appears only at the low dilution levels as it is drowned out by the bin F feature at the highest dilutions (dilutions 1 through 4). Box plots for bins G and F in Fig. 3c also show this relationship. Note also that the density histogram for the APPEAL data (Fig. 3b, thick light curve) suggests evidence of a weak native feature that is offset in bin F from the added insulin feature at 13,430.

Bin D also contains an added feature near 13,280. There is no known peptide in the peptide mixture corresponding to this, and it most likely represents a degraded form of the bovine insulin. In fact, the mass difference between the added feature in bin D and bovine insulin (bin F) is −71 m/z (TOFs calculated as under ‘Definitions: Feature and Intensity’ and then converted to a m/z scale from a calibrated spectrum), and so a possible source of the peak in bin D is suggested by the fact that one of the terminal amino acids of bovine insulin is alanine (contributing mass −71 Da).

The offset in the density histogram versus the histogram for the APPPEAL data indicates that bin D is shared with a native feature (note also the mild peak in the mean curve for the APPEAL data near 13,300). In fact, the box plot for bin D exhibits a slight increase in the intensities after dilution 6. Evidence that both an added and a native feature share bin D is seen in Fig. 5, which shows that features detected at the highest dilutions are offset in bin D from the features detected at the lower dilutions. Finally the intensities in bin D do not become zero even at the highest dilutions. This is in contrast to the added feature in bin B (also probably a degraded form of insulin) whose intensities are not significantly above zero at the high dilutions. For the record, the added features in bins B and E are each −41.5 m/z units less than the added features in bins D and F, respectively.

With regard to the existence of features, the properties revealed by these methods and exhibited in Fig. 3 are not necessarily obvious by looking at the raw spectra. In particular, graphs of the raw spectra make it difficult to determine whether bins B and C contain two features that are truly distinct. They are, however, clearly distinguished in the density histogram (Fig. 3b), and the corresponding box plots in Fig. 3c confirm that they exhibit different properties: the correlation between intensity and dilution indicate that bin B contains an added feature and bin C contains a native feature.

An additional remark concerning the existence of features, as detected by this procedure, is that although the box plots for bin B indicate an added feature, the histogram alone may not definitively identify its existence. In particular, the histogram from the APPEAL data also suggests a native feature here in a small subset of spectra. Note also the box plots at the highest dilution levels record a (native?) feature. Although this may be the case, it is also possible that the histogram has recorded very low intensity noise between the bin A and bin C features, but side lobes on the detail functions (see Fig. 1) force any such noise to be recorded partway between those features.

Finally box plots for bins A and H are included for reference. Features in bins A and H are apparently native, being present in all dilutions. The feature in bin H appears as a “shoulder” in the lowest dilutions, illustrating that there can exist local trends that should be accounted for in quantifying intensity. Although adjusting for base line is common, defining a base
line as subtle as that shown here would require substantial knowledge of signal content. The final two box plots in Fig. 3 show intensities using \( D_5 \) and \( X \), respectively, as recorded at the feature location in bin H. In the former, the intensities are appropriately constant across all dilutions. (Note that although the spectra in Fig. 3a have been shifted vertically for display, the raw intensities used in the final box plots were taken directly from the instrument. Also the APPEAL data are not presented in the final panel of Fig. 3c because a substantially different base line makes comparison of raw intensities difficult.)

**Data Set Properties**—Natural questions that arise when examining mass spectrometry data include the following. At what concentration is a peptide species detectable in a serum sample? How many peptide-related features are present in each spectrum? This data set cannot definitively answer these questions, but it does offer some insight.

Fig. 4 addresses the first question with box plots of intensities for each of the five peptides in Table I. Superimposed on each panel is a graph of dilution-specific intensity levels that result from an isotonic regression model. By design, the logarithm (base 2) of the intensities should decline linearly as a function of the dilution level. The dilution level that immediately precedes the lowest level set identified by the isotonic regression indicates the lowest concentration of the peptide mixture in which the given peptide can be detected and differentiated from the lower concentration(s). The amount of peptide at this dilution is highlighted in bold in Table II, which also gives the amount of added peptide at all 10 dilution levels. The lower limit of detection ranges from \(~100–900\) pg.

In general, the term 'detection limit' here only refers to the detection of a peptide whose abundance is greater than any other peptide that may share the same bin. In our analysis, this may happen if the TOF values of these peptides are closer than that which can be distinguished by the histograms because a native feature may share approximately the same location as an added feature. This was seen in bin D of Fig. 3, and a similar analysis (not shown) suggests this occurs for the two-charge cytochrome c feature. In Fig. 5a we highlight this scenario by plotting mean spectra and feature locations for dilutions 1–6 (light curves) separately from dilutions 7–10 (dark curves). In bin D of Fig. 5b the features for dilutions 1–6 (light curve) are offset by about 10 TOFs from those in dilutions 7–10. When plotted jointly, as in Fig. 3c, the density has a single mode. The vertical lines in Fig. 5b delimit bins from dilutions 7–10 only, whereas the tick marks at the bottom are from all 10 dilutions (as in Fig. 3). Bin D should be contrasted with bin E in which the features from the high versus low concentration spectra give rise to distinct bins. The difference between the locations of these features is made explicit in Fig. 5b.

For reference, the bottom row of Fig. 4 exhibits box plots and isotonic regression curves for three unidentified peptides from a wide range of TOF values (7920, 14,970, and 18,940) and intensities. Also for comparison between intensities as quantified by \( \log_2(D_5) \) versus \( \log_2(X) \) (raw intensities), the center and center right panels exhibit box plots of these two for the added peptide cytochrome c. (See also the final two subplots of Fig. 3c.) As expected, both decline approximately linearly as a function of dilution level (at least for levels 2 through 8), although the \( \log_2(D_5) \) box plots exhibit approximately constant variance across dilutions, something not seen in the \( \log_2(X) \) box plots. Because the isotonic regression used here assumes equal variances, no such regression is plotted on the box plots for the raw intensities.
The degree to which intensities \( \log_2(D_o) \) do not decline linearly (prior to leveling off) can be attributed to the fact that at dilution levels 1 and 2, where the added features are huge and have a wide base (see Fig. 3a), the scale 5 content in the signal, being based on a window of \( 2^6 \) TOF measurements, exists near the top of the peak and hence is approximately the same for both dilution levels.

As for the number of unambiguous features, we can informally investigate this by imposing a threshold on the size of features recorded. In particular, the choice for a threshold C in a condition such as Intensity \( > C \) can now be informed by the properties exhibited in Fig. 4. Note that the vertical axis is the base 2 logarithm of intensity as measured by a scale 5 detail function, \( D_o \). So we define a thresholded scale 5 feature as before but add the requirement that \( D_o(t_o) > 4 \) at any feature location \( t_o \). Recording only these features in an m/z range of \( \sim 2000–20,000 \) and counting the number of bins in which at least 5% of the spectra have a feature gives a total of 423. Similarly the number of bins in which at least 10% (25 and 50%) of the spectra have a thresholded feature is 386 (327 and 246, respectively).

**DISCUSSION**

This experiment provides a unique perspective on the quantitation of biologically induced signal in mass spectrometry data from complex human sera. It offers more information than can be obtained from a set of clinical samples in that even subtle features are manifest dynamically (as a function of dilution) in the spectra. The methods used in this study successfully identified and quantified features corresponding to low abundance peptide species occurring in a small subset of the spectra. We identified features from peptides added to complex serum at the 20-nmol level and distinguished between these low intensity added peptides and similar subtle features coming from peptides in the serum.

Fundamental to any analysis of mass spectrometry data is the definition of “signal” in the spectra used to infer the existence of peptides. This along with how features in the signal are quantified to reflect abundance must be carefully considered. The definitions used here (see “Definitions: Feature and Intensity”) are based on a decomposition by scale that simply extracts, rather than models, signal content. Features are not defined in terms of signal-to-noise ratios and, being defined by local changes (not local averages), are independent of the base line of a spectrum. Our presentation uses only one scale of the decomposition, however, and hence is biased toward this portion of the signal; these features occur on a scale of approximately \( 2^5 = 32 \) TOF measurements. On the other hand, nothing is lost in this decomposition, and additional scales could be included in further study. Our conclusion that this single scale reflects biologically induced signal in these data is made possible by the design of the experiment, which produced known features at various intensities. This also made possible a biologically informed judgment concerning which features are relevant in other regions of the spectra.

No claims of optimality are being made regarding this scale or these methods, but the approach taken here should be contrasted with that of defining features as “peaks,” or local maxima, in the raw spectrum where the existence and location of a peak is dependent on the window width within which the maximum is calculated. Note also that scale \( j \) features include, by definition, inflections or “shoulders” that are not local maxima but are nonetheless relevant to these data (as seen in Figs. 2, 3, and 5).

Biomarker studies using data of this type often proceed with minimal investigation into what constitutes biologically related spectrum content. Instead the focus is often on finding any properties that may classify the spectra as cases versus controls. On the other hand, some impressive work aimed at helping researchers rigorously quantify signal content has been done. In particular, Baggerly et al. (1) carefully consider reproducible signal in clinical data, whereas Morris et al. (4) introduce a prototype for simulating MALDI-TOF data. The present study, with an experimental focus on known peptide content, represents another effort in this direction. The methods used in processing this data are consistent with the idea that this study will be most valuable to other researchers if the summaries describe output obtained directly from the spectrometer rather than data that have been massaged to fit unsubstantiated modeling assumptions. Smoothing and base-line correction are, however, valuable tools for prepro-

**TABLE II**

Concentrations of peptide species at each dilution

| Dilution | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ACTH clip-(18–39) | 28,700 | 14,375 | 7,188 | 3,594 | 1,797 | 898 | 449 | 225 | 112 | 56  |
| ACTH clip-(7–38)  | 8,700  | 4,375  | 2,188 | 1,094 | 547  | 273 | 137 | 68  | 34  | 17  |
| Bovine insulin   | 15,600 | 7,813  | 3,906 | 1,953 | 977  | 488 | 244 | 122 | 61  | 31  |
| Cytochrome c     | 39,700 | 19,350 | 9,675 | 4,838 | 2,419 | 1,209 | 605 | 302 | 151 | 76  |
cessing, and so the perspective presented here may provide an experimentally informed basis for subsequent analyses that include such adjustments in preprocessing MALDI-TOF data.

**APPENDIX: COMPUTATIONAL STEPS**

The basic procedure for computing binned features and intensities for $N$ spectra is performed as follows. For computational specifics, see Randolph and Yasui (13). (i) Calculate scale $j$ feature locations for all spectra, producing $N$ vectors of ones and zeros (denoting the existence or lack of features, respectively). (ii) Sum across all $N$ spectra and divide by $N$ to create a density histogram of feature locations. (iii) Compute a local averaging smooth of the density histogram. The locations of local minima in this smooth are used to delimit bins in which feature locations aggregate in high density. (iv) Within each bin features from different spectra are treated as the same feature and quantified by the intensity of the scale $j$ detail function. If a spectrum has no features in a bin, the

**Fig. 5.**  
(a) TOF versus intensity (arbitrary units) of mean spectra for dilutions 1–6 (light) and dilutions 7–10 (dark). Vertical lines delimit bins calculated from dilutions 7–10 only. The tick marks delimit bins calculated from all dilutions.  
(b) Histograms of feature locations from dilutions 1–6 (light) and 7–10 (dark). The vertical lines are as in a.
intensity is recorded as zero. If a spectrum has more than one feature in a bin, the intensity of the largest feature is recorded.

The occurrence of a single spectrum having multiple features in a single bin arose primarily in regions of low intensity and, potentially, minimal or no signal. This led naturally to recomputing the density histogram based only on features that did not share a bin with another feature from the same spectrum. This was done followed by a recalculation of bins for this modified histogram. Intensities were then calculated as in step iv using the new bins and the original set of feature locations; if a spectrum had more than one feature in a bin, the intensity of the largest was recorded.

We note that this recalculation of bins had only a minor effect, but it can help to reduce the effect of noise-related features in regions of low signal. It is relevant to recall that the existence of a scale \( j \) feature does not depend explicitly on intensity, and hence any local change at this scale will be identified as a feature including, potentially, one in a region of a spectrum that contains no biological signal. The difference between a feature in such a region versus a region containing a true feature is reflected in consistency of location across multiple spectra (via the density histogram). Restricting attention to features that are unique to their bins provides a method of filtering out (some) noise-related features without estimating a signal-to-noise ratio.

* This work was supported by National Institutes of Health Grants K25-GM67211 (to T. W. R.), T32-CA80416 (training grant support to B. L. M.), CA116393 and R03-CA108339 (to P. D. L.), and CA086368 and P01-CA53996-24 (to Z. F.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: Dept. of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195. E-mail: trandolp@fhcrc.org.

REFERENCES
1. Baggerly, K. A., Morris, J. S., Edmonson, S. R., and Coombes, K. R. (2005) Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J. Natl. Cancer Inst. 97, 307–309
2. Diamandis, E. P. (2004) Analysis of serum proteomic patterns for early cancer diagnosis: Drawing attention to potential problems. J. Natl. Cancer Inst. 96, 353–356
3. Ransohoff, D. F. (2005) Lessons from controversy: ovarian cancer screening and serum proteomics. J. Natl. Cancer Inst. 97, 315–319
4. Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2008) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21, 1764–1775
5. Coombes, K. R., Fritsche, H. A., Clarke, C., Chen, J. N., Baggerly, K. A., Morris, J. S., Xiao, L. C., Hung, M. C., and Kuerer, H. M. (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clin. Chem. 49, 1615–1623
6. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T. (2004) Sample classification from protein mass spectrometry by ‘peak probability contrasts’. Bioinformatics 20, 3034–3044
7. Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 19, 1636–1643
8. Yasui, Y., McLerran, D., Adam, B. L., Winget, M., Thornquist, M., and Feng, Z. (2003) An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. J. Biomed. Biotechnol. 2003, 242–248
9. Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L. C., and Coombes, K. R. (2005) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. Proteomics 5, 1667–1672
10. Coombes, K. R. (2005) Analysis of mass spectrometry profiles of the serum proteome. Clin. Chem. 51, 1–2
11. Mitchell, B. L., Yasui, Y., Lampe, J. W., Gafken, P. R., and Lampe, P. D. (2005) Evaluation of MALDI-TOF MS proteomic profiling: identification of α2-hs glycoprotein b-chain as a biomarker of diet. Proteomics 5, 2238–2246
12. Percival, D. B., and Walden, A. T. (2000) Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge, UK
13. Randolph, T. W., and Yasui, Y. (2005) Multiscale processing of mass spectrometry data. Biometrics, in press
14. Schell, M. J., and Singh, F. (1997) The reduced monotonic regression method. J. Am. Stat. Assoc. 92, 128–135