4030 Fabian Way
Palo Alto, CA 94303

MASS SPECTROMETRIC DATA PROCESSING

by Luke V. Schneider

Mass spectrometry is an analytical chemistry tool used for the separation of molecular ions by their mass, more properly their mass-to-charge ratio (m/z), and the quantification of their relative abundance in an ion stream. For convenience, “mass” will be used as shorthand for m/z herein, except where specific m/z terminology will be more helpful.

Like any tool there are correct or optimal ways to use it, and there are many incorrect ways that people have found to misuse this tool, resulting in corrupted or inferior information obtained from it in the process. In the following series of articles, we attempt to describe the best practices for the analysis of mass spectrometric data.

The mass spectrometer provides a measurement in terms of mass and abundance of the molecular ions in a stream of such ions. In any analytical method, one must distinguish the precision (the intrinsic reproducibility of the measurement) from the accuracy of the result (how well do the peak mass and abundance agree with the true values). Precision of the measurement is affected by chemical and instrument noise, fluctuations in room temperature, and by the digital sampling frequency and any non-linearities of the detector. Accuracy depends on how well the instrument has been calibrated, but may never be better (tighter) than the inherent precision of the measurement.

1. PROPERTIES OF MASS SPECTROMETRIC DATA

The primary (raw) profile spectral data produced by the mass spectrometer data acquisition software—quantized mass and abundance data pairs that represent the sampled distribution of molecular ions inside the mass analyzer—is fundamentally the same in all mass analyzers. The attributes of that raw (profile) data differ by the type of mass analyzer. These differences are highlighted in the section on Spectral Characteristics and can have a critical impact on the processing of the raw spectral data.

2. DATA ANALYSIS PROCESS

While the nature of the primary data generated is the same (mass and abundance), the goals of the analysis (i.e., how this data will be used) may be quite varied. However, the initial steps in this process are common to all downstream analyses and uses of mass spectral data. The first goal in all subsequent uses for the data is to identify within the raw (profile) data the discrete mass and abundance peaks associated with each analyte and their relative abundances that are present within the sample (i.e., converting the profile spectrum into a mass list of peaks).

As shown in Figure 2.1 below, there are several discrete steps involved in converting the profile spectrum into a mass list, and many pieces of metadata about the mass spectrum that can and should be collected during this data analysis process, which can prove useful to improve the precision of the resulting mass list. It should be emphasized that precision is all that matters in this process. Accuracy correction, which generally depends on external calibration of the mass spectrometer, is properly applied to the final mass list, not to the profile spectrum.

In these documents, we present the general alternatives used for each step of the process, along with their advantages and limitations. We also provide recommendations for the best practices that should be followed to improve the precision

and sensitivity of the resulting mass list, and to collect the relevant metadata about the spectrum that can assist in defining how well the final result is known or understood.

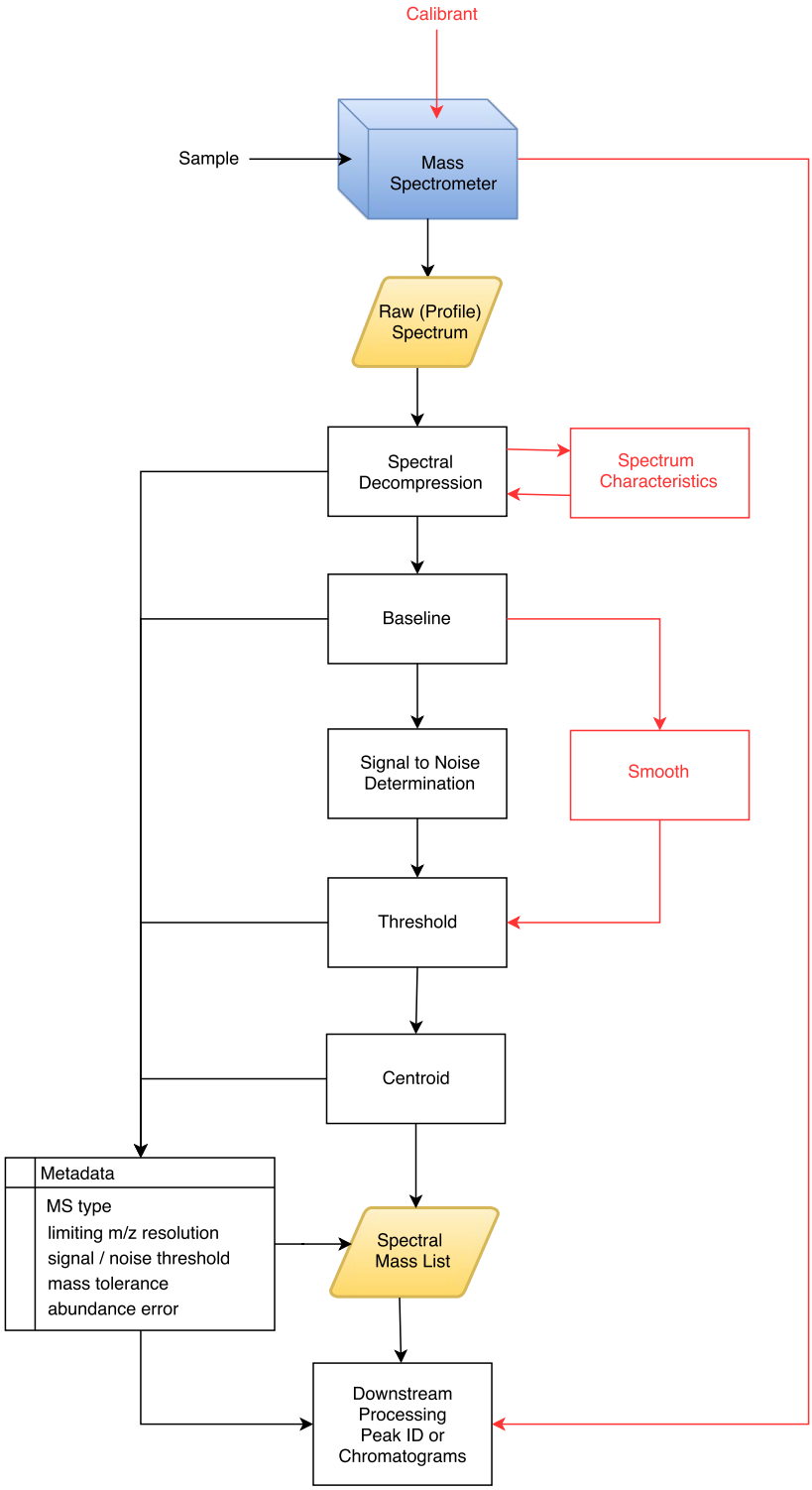


Figure 2.1. "Best Practice" steps in the process for converting raw (profile) mass spectrometric data into a mass list optimized for subsequent uses and analysis. Each of the steps of this process are detailed in subsequent

sections of this document series. Those in red are background reference documents

3. SPECTRAL DATA COMPRESSION AND DECOMPRESSION

Data Compression

Mass spectral data is stored as mass and abundance data pairs in a data file. Most instrument manufacturers automatically compress this data file to save storage space and the amount of data that must be transferred from the instrument to the analysis software. The methods employed to compress the spectral data files include:

- retention of only significant digits from values (e.g., quantifying a mass to just four instead of six or more decimal places, and rounding abundances off to the nearest integer value)
- conversion of the data to binary formats
- removal of any points with zero counts (or counts below a user-specified threshold)
- removal of any points with constant abundance between the two anchor values at the edges of the constant range.

Retained Significant Figures

Most of these data compression techniques are reversible, except for significant figure truncation and the elimination of data below a user-specified threshold value. When the relevant and appropriate number of significant figures are maintained in the mass and abundance values, the data lost is typically insignificant. However, this rounding or truncation error can still cause variation in the estimation of the intrinsic mass spacing (IMS), which must be taken into account during data decompression (see below).

Thresholding

With the exception of FT-ICR absorption spectra, which have been re-registered to a median abundance of zero counts, there is no data below zero abundance in mass spectra. Therefore, zero removal is not considered data destructive. However, when a non-zero, user-specified, minimum threshold is applied, all data between that value and the true zero is permanently lost, which creates issues for downstream spectral processing, and should always be avoided during original data acquisition and archiving.

Data Point Compression

Removing zero abundance points or using a variant of the Lempel-Ziv-Welch (LZW) compression technique[1] to remove constant values between two anchor points, are both non-destructive compression techniques utilized in mass spectrometry. However, since both methods are in common practice by different instrument vendors, it may not be clear which method has been applied on a spectrum from an unknown source (see below).

Binary File Conversion

[1] Lempel-Ziv-Welch File compression, https://en.wikipedia.org/wiki/Lempel–Ziv–Welch (accessed 23 June 2016).

Binary data can typically be converted to more readable formats using MSConvert[2] or other open-source MS file conversion software. All vendors have released proprietary .dll files into open-source use for this purpose. However, these libraries do not typically replace data points removed by zero drop or LZW compression schemes. The reversal of binary data formats is beyond the scope of this paper.

Spectral Data Decompression

Importance of Decompression

When mass spectral data files are compressed, information and statistical degrees of freedom are lost. For example, the average counts in the mass spectrum shown in Figure 3.1 changes from 9.48 in the LZW-compressed version (87,707 data points) to 3.71 in the fully-decompressed version (178,402 data points). Using the correct lower average counts as a threshold before centroiding, 67.3% more peaks are detected in this spectrum. The elimination of over 50% of the data points by LZW compression resulted in 67.3% of the peaks going undetected, because the threshold is not calculated correctly with those points missing.

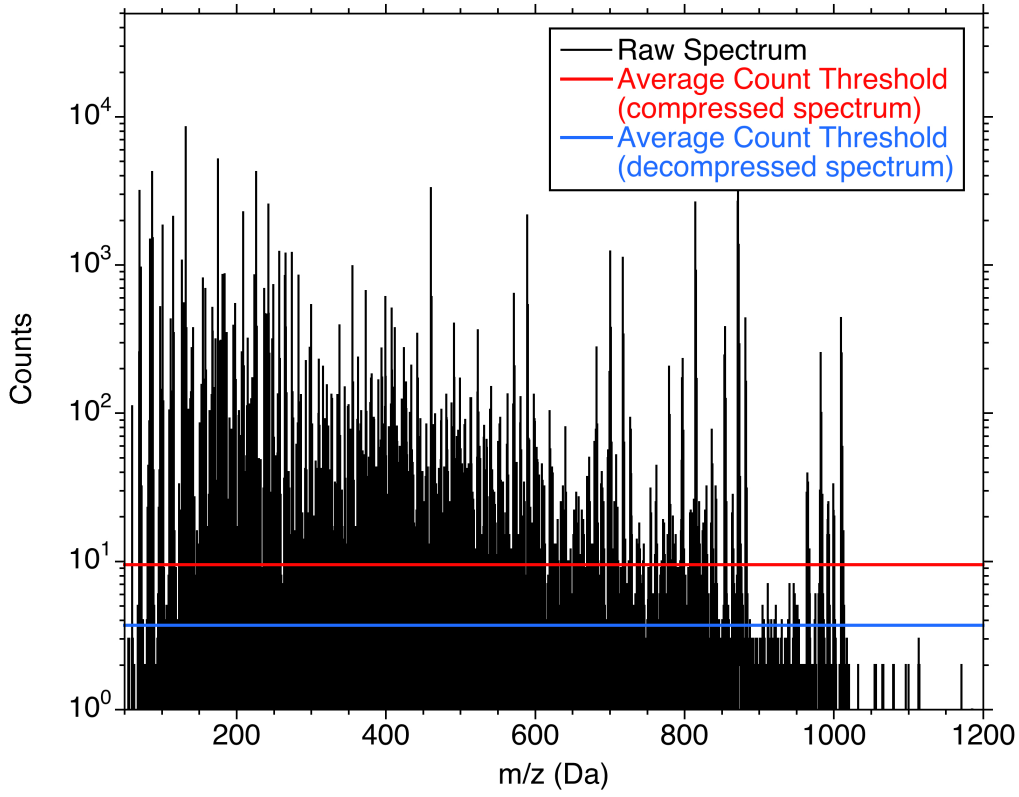


Figure 3.1. The CID fragmentation spectrum of the 2+ charge state of the peptide NQGPQESVVR from a Waters Q-TOF Premier mass spectrometer. There are 35,005 missing data points in the LZW-compressed file. When these points are returned to the spectral file, the average counts drop from 9.48 to 3.71. Using the average counts from the compressed spectrum as a peak detection threshold, 67.3% of the total peaks would remain undetected in the compressed data file.

[2] MSConvert software, http://proteowizard.sourceforge.net/tools.shtml

Intrinsic Mass Spacing

The keys to any mass spectral data decompression are to determine:

- when a data point has been removed.
- the mass value of the point(s) to be replaced.
- the abundance value of the point(s) being replaced.

The most critical step is to determine the intrinsic mass spacing for the spectrum. The masses in each type of mass analyzer are set on a periodic spacing.[3] This intrinsic mass spacing (IMS) between data points is readily determined from the mass spacing in the spectrum itself. For example, in the TOF spectrum above (Figure 3.1), the distance between mass points is expected to be constant on a mass to the half power basis (i.e., $\Delta m/z^{0.5}$). This is shown in Figure 3.2.

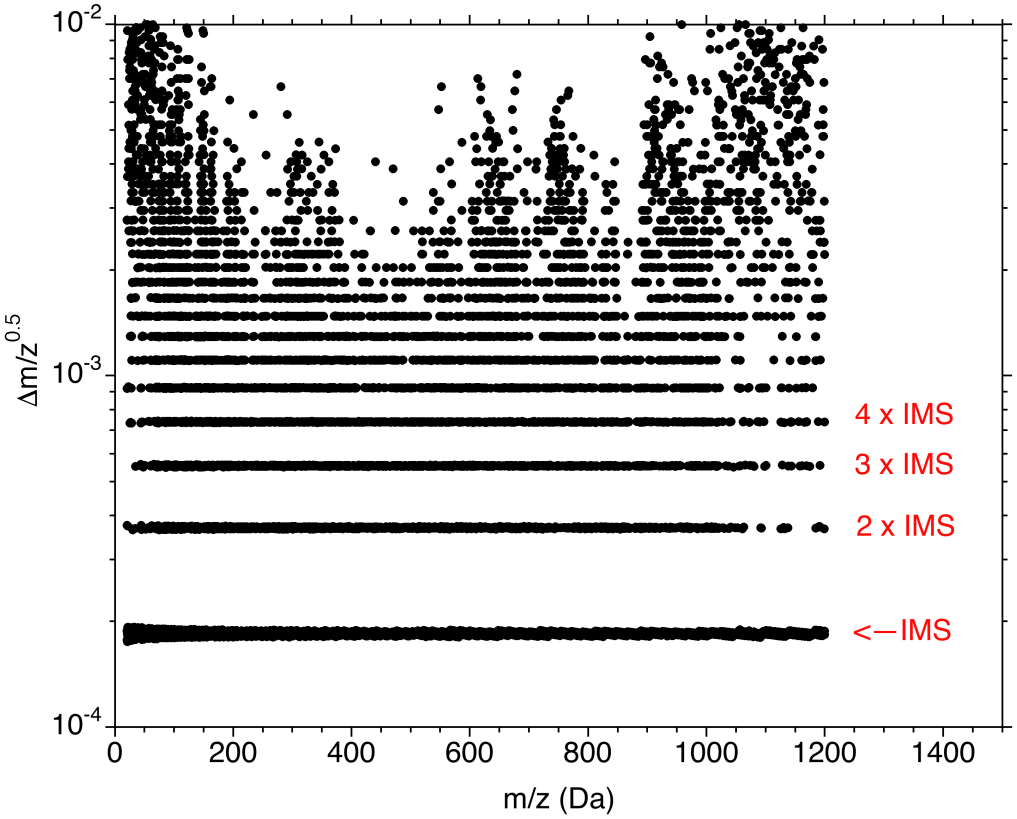


Figure 3.2. The intrinsic mass spacing of the TOF spectrum from Figure 3.1 on a $\Delta m/z^{0.5}$ basis. The average IMS is 1.85 x 10^{-4} $Da^{0.5}$, as represented by the lowest set of data points. Each line above that reflects a larger gap in mass spacing, indicating missing data points in the compressed spectrum. The multiple of the IMS (minus one) indicates the number of such missing points in that mass gap. The scatter about the average IMS is caused by truncation error in the mass values (another data compression mechanism).

Similar IMS graphs can be generated for the compressed spectral data from each mass analyzer type, but reflecting different powers for the mass calculation

[3] see section on Spectral Characteristics

(Table 3.1). The mass points in an ion trap analyzer are evenly spaced in m/z (Δm/z, Da). Orbitrap mass data is evenly spaced on the reciprocal of the square root of the m/z spacing ($\Delta m/z^{-0.5}$, $Da^{-0.5}$). FT-ICR mass data is evenly spaced on the reciprocal of the m/z spacing ($\Delta m/z^{-1}$, Da^{-1}).

Mass Analyzer	Calculation	Units
Ion trap	$\Delta m/z = \left[(m/z)_{i+1} - (m/z)_i\right]$	Da
Time-of-Flight	$\Delta m/z^{0.5} = \left[(m/z)_{i+1}^{0.5} - (m/z)_i^{0.5}\right]$	$Da^{0.5}$
Orbitrap	$\Delta m/z^{-0.5} = \left[(m/z)_i^{-0.5} - (m/z)_{i+1}^{-0.5}\right]$	$Da^{-0.5}$
FT-ICR	$\Delta m/z^{-1} = \left[(m/z)_i^{-1} - (m/z)_{i+1}^{-1}\right]$	Da^{-1}

Table 3.1: Intrinsic Mass Spacing (IMS) for different types of mass analyzers.

Mass Position of Decompressed Data Points

Armed with a knowledge of the IMS for the spectrum, it is possible to detect compression gaps by looking for any adjacent mass values that differ by more than 1.5 times the expected IMS for their position in the spectrum. Furthermore, by dividing the $\Delta m/z^x$ for each gap by the IMS calculated for the spectrum, and rounding to the nearest integer, the number of missing data points in each gap is readily determined. It is then a simple matter to add that number of mass values back into the spectrum in equally-spaced increments in the proper ($\Delta m/z^x$) domain for the mass analyzer.

Abundance Value of the Decompressed Data Points

The last piece of information needed is the correct abundance value for each decompressed point. As described above, there are two basic types of data compression schemes. The first is zero removal. The second is characterized by anchor points of equal abundance that span the missing points gap. By checking the abundance values on either side of all identified mass gaps in the spectral data, the specific type of compression can ascertained.

Once all zero abundance values have been removed from a dataset, the remaining data may by sheer chance have equal nonzero abundance values on either side of the corresponding mass gap, effectively making that gap indistinguishable from a LZW compression gap. Therefore, a zero removal compression scheme can only be robustly identified where at least one mass gap within the spectral data set contains non-equivalent abundances on either side of the data gap. An LZW type compression scheme is only robustly identified when both sides of all mass gaps have identical abundances.

Orbitrap instruments are manufactured by a single supplier, and they always utilize only the zero removal type of data compression.

4. SPECTRAL CHARACTERISTICS

A mass spectrum reports the measured mass to charge (m/z) ratio of molecular ions, generally by transforming the true detector signal data captured in the time domain, via calibration to yield final results in the m/z domain. Since the charge of any given spectral peak can only be determined from its isotopic pattern of peaks, it is common to refer to the m/z as the mass when talking about isolated peaks and mass range when talking about the range of masses that are represented in a mass spectrum.

In reality, therefore, a mass spectrum is built from the measured number of ion detection events counted in each detection bin, with a bin consisting of a slice of

detection time or frequency. The detection bins are then mapped to the mass-to-charge (m/z) or mass range through the use of mass standards (molecular ions with known m/z peaks). Since a single molecular ion may be tracked across multiple detector bins, the m/z mapping of the spectrum on the ordinate (x-axis) is at best approximate, and ultimately depends on the care taken to locate the underlying apex of the distribution of each molecular ion across multiple bins.

Detectors have duty cycles that prevent capture of all ion detection opportunities, but they can also amplify detection events. Furthermore, only molecules that are ionized can be detected. Therefore, counts along the abscissa (y-axis) must also be calibrated before they can be used to quantitate the concentration of the molecules generating them.

Ordinate (x-axis)

The size of (or spacing between) detector bins is constant along the x-axis in the acquisition time or frequency domain. However, the correlation of spacing between these time points and the mass to charge (m/z) ratio of the ions varies depending on the analyzer type. Understanding the inherent correlation between the detector bin spacing and the range of m/z represented by each detector bin is critical for proper downstream processing of the spectrum, since it affects the width of each ion peak in mass units (i.e., the mass resolving power of the spectrum at any given m/z). It also determines the limiting mass precision of any m/z or mass call, which can be no more precise than the width of a detector bin at that mass value. Mass accuracy, which is distinguished from precision in analytical chemistry, is ultimately limited by the accuracy with which detector time bins are mapped to the mass domain in the final spectrum.

There are also several traditional mass resolution measures that are useful for characterizing what a mass analyzer is capable of separating. Resolution (a dimensionless number) is typically defined by the mass divided by the full peak width at half of its maximum height (PWHH) in m/z. The inverse of PWHH Resolution is typically presented as parts per million (ppm) precision (i.e., an approximation of the mass difference at which a mass analyzer is expected resolve two nearly isobaric species as separate peaks with half the peak heights of each being non-overlapping). The final resolution measure is the peak width at half height in Da. One of more of these measures may be constant for any given mass analyzer.

Ion Trap

The detector bin spacing in ion trap analyzers is linearly proportional to m/z. This means that every unit distance in an ion trap spectrum represents the same fraction (Δm/z slice) of the mass range of the spectrum. If there are 1000 points in the spectrum covering a 100 Da range in mass, each point represents a Δm/z slice of 0.1 Da, which corresponds to the minimum precision with which the mass of any molecular ion could be determined . If the mass range of the spectrum is reduced to 50 Da for the same sampling speed (i.e., 1000 points), then each point represents a Δm/z slice of 0.05 Da, or double the mass precision of the 100 Da range. A consequence of this direct proportionality between detector bins and m/z is that the resulting peak widths are nearly constant with m/z in ion trap spectra (Figure 4.1). However all other common measures of resolution vary with m/z in ion trap spectra.

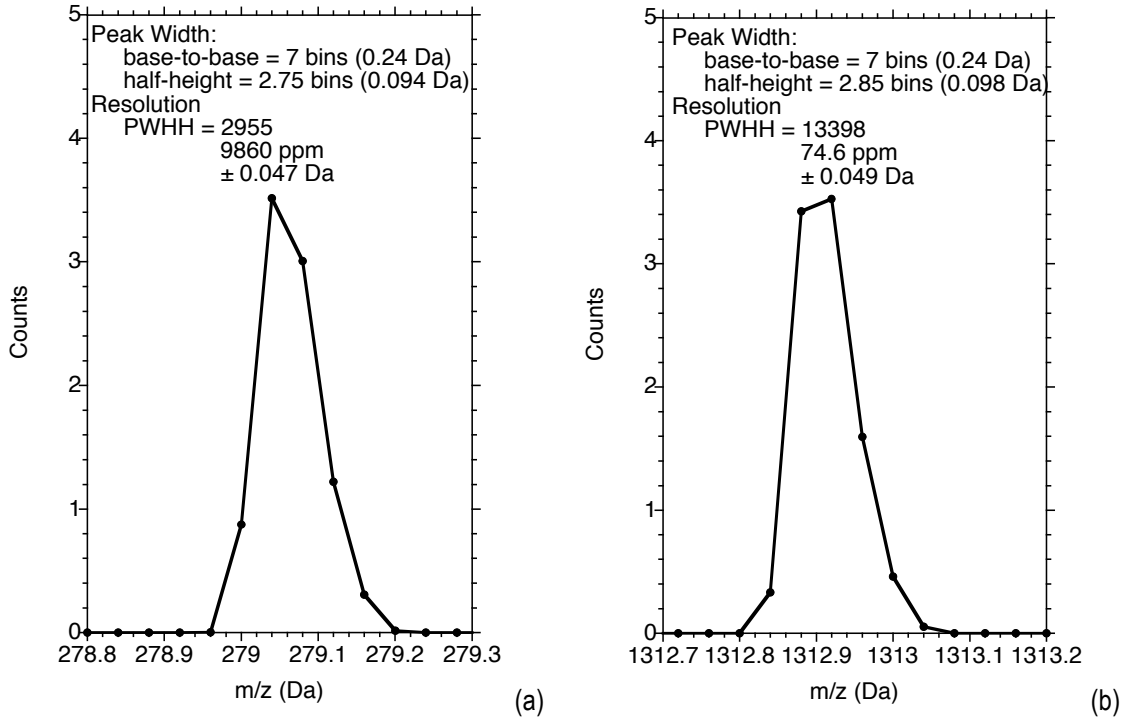


Figure 4.1: Two isolated peaks from different m/z ranges in the same ion trap spectrum of peptide fragments (a and b). Peak widths are constant in the m/z and detector bin domains, but the calculated resolutions (peak width at half height) vary in all dimensions except $\pm\Delta m/z$.

Time-of-Flight (TOF)

The detector bin spacing in TOF analyzers is proportional to the square root of the m/z ($[m/z]^{0.5}$). This means that every successive point in a TOF spectrum represents a larger fraction of the mass range than every preceding point. The limiting precision in $\Delta m/z$, therefore, depends on the mass being measured, but is constant on mass to the half power ($mass^{0.5}$). It also means that the peak width gets progressively wider with increasing m/z by all resolution measures except ppm (Figure 4.2).

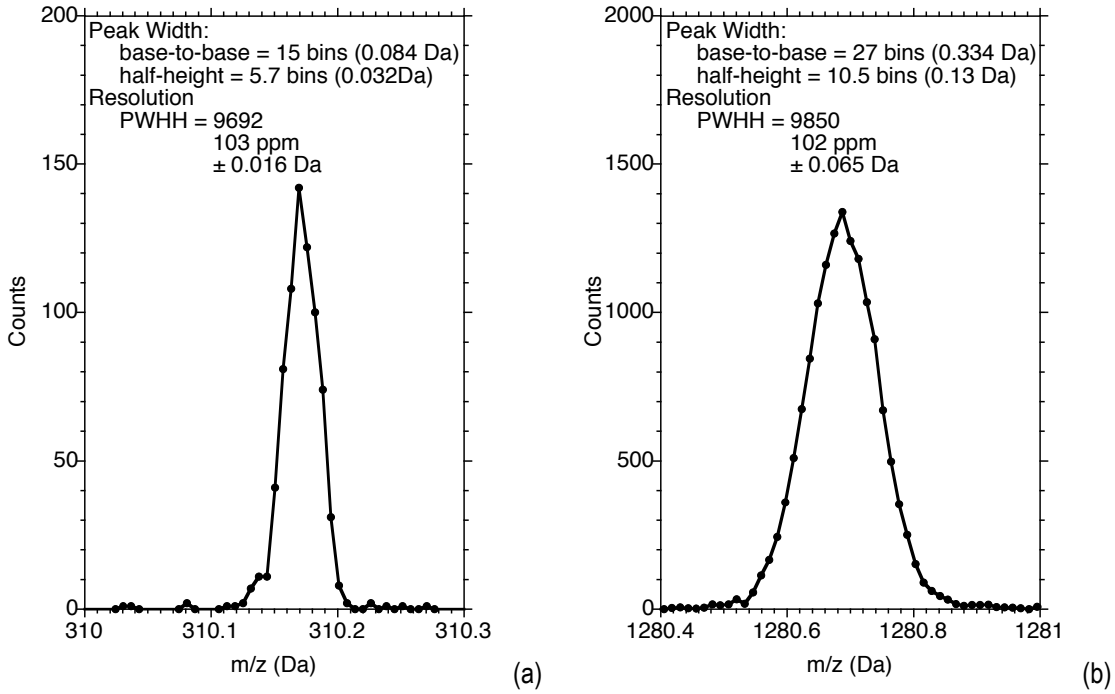


Figure 4.2: Two isolated peaks from different m/z ranges of the same TOF spectrum of peptide fragments (a and b). Peak widths increase with m/z, but the calculated resolution is constant by all measures but $\pm\Delta$m/z. The inherent limiting mass precision, however, is constant at the bin spacing of 0.0001793 $Da^{0.5}$.

Orbitrap

The bin spacing in Orbitrap analyzers is proportional to the inverse square root of the m/z ($[m/z]^{-0.5}$). This means that every successive unit distance in an Orbitrap spectrum represents a larger fraction of the mass range than every preceding unit distance. The limiting precision in Δm/z, therefore, depends on the mass being measured, but is constant versus the reciprocal of that mass to the half power ($mass^{-0.5}$). It also means that the peak width gets progressively wider with increasing mass, yet is roughly constant on the basis of number of detector bins (Figure 4.3).

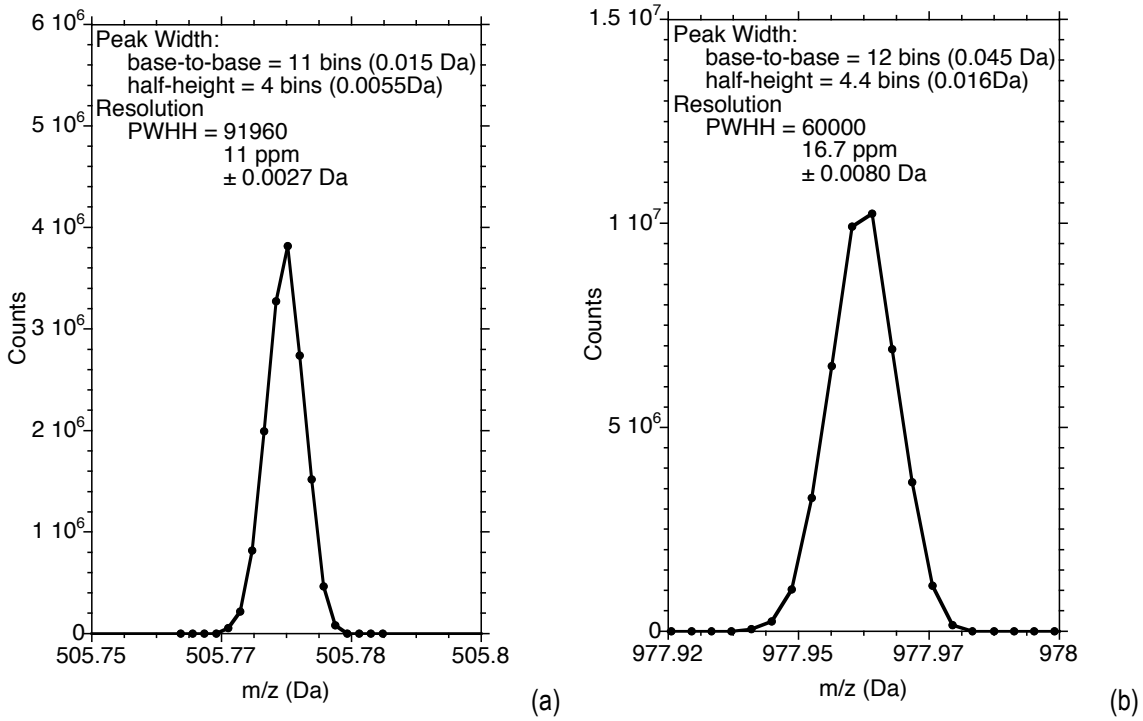


Figure 4.3: Two isolated peaks from different m/z ranges of the same Orbitrap spectrum of peptides (a and b). Peak widths increase with m/z, but are constant in terms of the number of detector bins or mass$^{-0.5}$. The calculated resolution varies by all measures. The inherent limiting mass precision, however, is constant at the bin spacing of 6.71 x 10^{-8} Da$^{-0.5}$.

Fourier Transform-Ion Cyclotron Resonance (FT-ICR)

The bin spacing in FT-ICR spectra is proportional to the inverse of the m/z ([m/z]$^{-1}$). This means that every successive unit distance in an FT-ICR spectrum represents a larger fraction of the mass range than every preceding unit distance. Therefore, the limiting precision in Δm/z, depends on the mass being measured, but is constant versus the reciprocal of that mass (mass^{-1}). It also means that the peak width gets progressively wider with increasing m/z and that there is no conventional autoscaling approach to resolution that remains roughly constant across this type of mass spectrum (Figure 4.4).

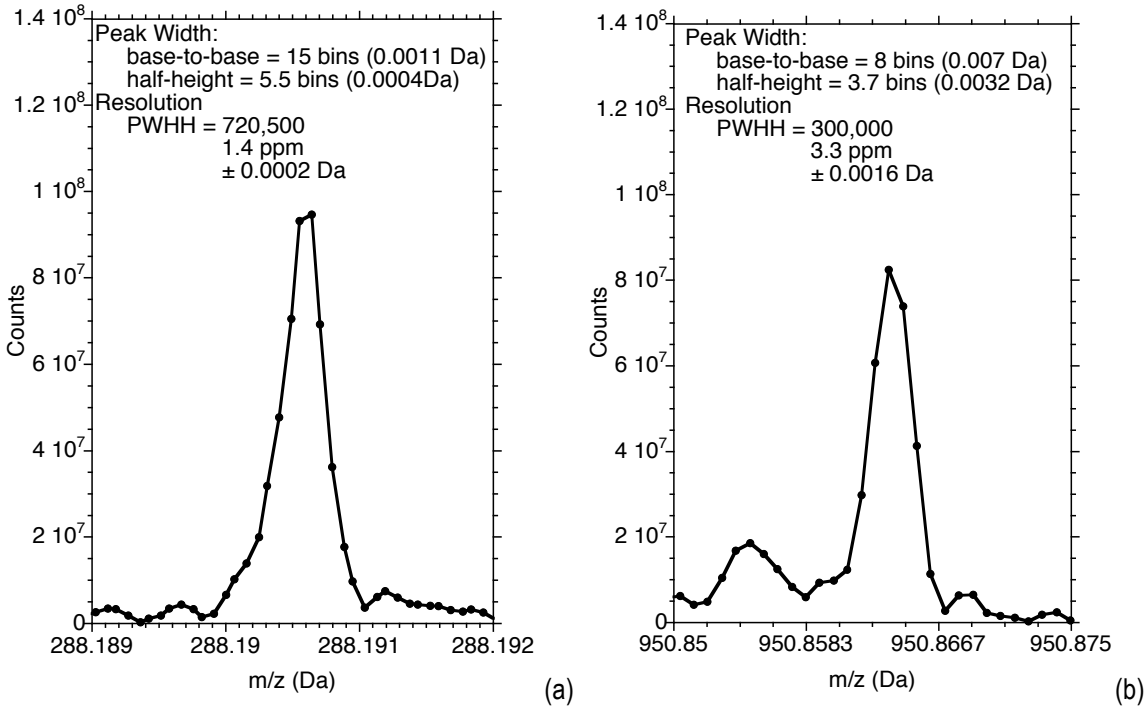


Figure 4.4: Two isolated peaks from different m/z ranges of the same FT-ICR spectrum of crude oil (a and b). Peak widths increase with m/z and decrease in number of points. The calculated resolution varies by all measures. The inherent limiting mass precision, however, is constant at a bin spacing of 6.71×10^{-8} Da^{-1}

Abscissa (y-axis)

The relative signal strength of the ions measured by the detector within each $\Delta m/z$ packet (detector bin) is the fundamental quantity represented along the abscissa of a mass spectrum. It is important to understand that different types of ions "fly" or are transported with different overall efficiencies through the mass spectrometer, and thus may generate different signal strengths at the detector. Second, the duty cycle of the analyzer (i.e., how much time it is devoted to detecting ions of a given mass) affects the signal strength registered. Third, the electronic gain of the detector (or residence time in an ICR cell) affects both the noise accumulated in the spectrum and the number of counts generated for a given molecule. Fourth, interferences due to ion density for ions circulating in an ICR cell, or due to the dead time of an MCP detector following each detection-event collision, can thwart quantitative detection of molecular ions. Finally, the ionization efficiency of the analyte itself plays a huge role in determining how the counts detected by the analyzer correlate to the actual concentration of the un-ionized parent molecule in the original sample.

Unlike most analytical techniques, both chemical and instrument noise are always positive in a mass analyzer. This severely limits the application of normal parametric statistical methods to mass spectra and makes peak detection, quantification, and discrimination very difficult. Therefore, the relative abundance of mass spectral peaks, particularly relative to stable isotope analogs (spiked or endogenous) become more useful tools for mass spectral analysis than the actual abundance of any given species.

5. SPECTRAL BASELINING

The Instrument-Specific Nature of Mass Spectral Baselines

Ion Trap and TOF Spectra

In mass analyzers using MCP detectors, the collision of an ion with the detector generates a signal, but signals may be generated by the intended analyte and by other molecular ions of the same nominal m/z (e.g., ion fragments produced during the ionization process from other analytes, ions contaminating the sample from the matrix, multiply charged or isotopic ions of other species in the sample, etc.). The detector itself also generates random noise, the amount of which depends on the operating parameters (e.g., gain and temperature) and age. Because molecular ions are destroyed at the surface of the MCP detector, there is a residue accumulation over time.

Noise (both chemical and instrumental) is always positive in a mass spectrum. This means: 1) that it is additive and raises the effective spectral baseline when multiple scans are combined (Figure 5.1), or 2) swarms of the same ions arriving at the detector simultaneously or interfering with each other in a trap or flight tube can cause the baseline around more abundant ions to rise relative to the rest of the spectrum (Figure 5.2). Therefore, the first step in any quantitative ion analysis is determining where the baseline should be established (i.e., what is the true zero from which the peak abundance should be determined?).

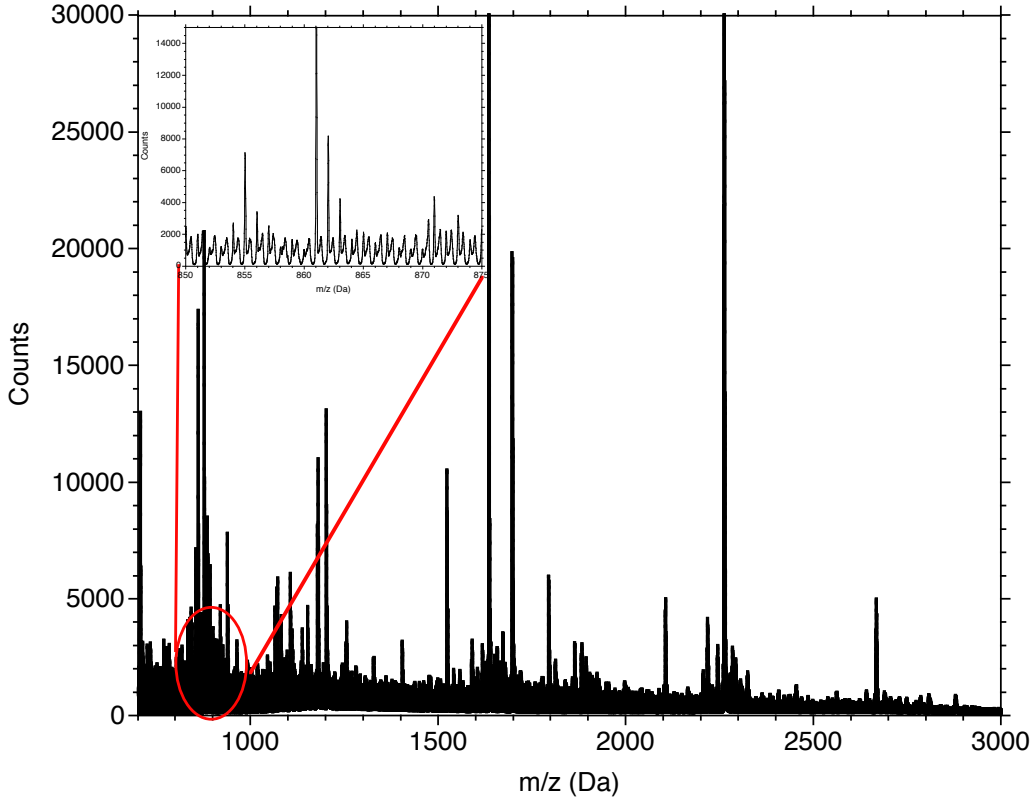


Figure 5.1. A MALDI-TOF peptide spectrum where matrix noise creates a progressive baseline offset (gap between x-axis and spectrum at 1200 Da) and superimposed matrix noise peaks (inset) towards the low mass region, which is compounded by the summing of multiple scans to create the final mass spectrum shown.

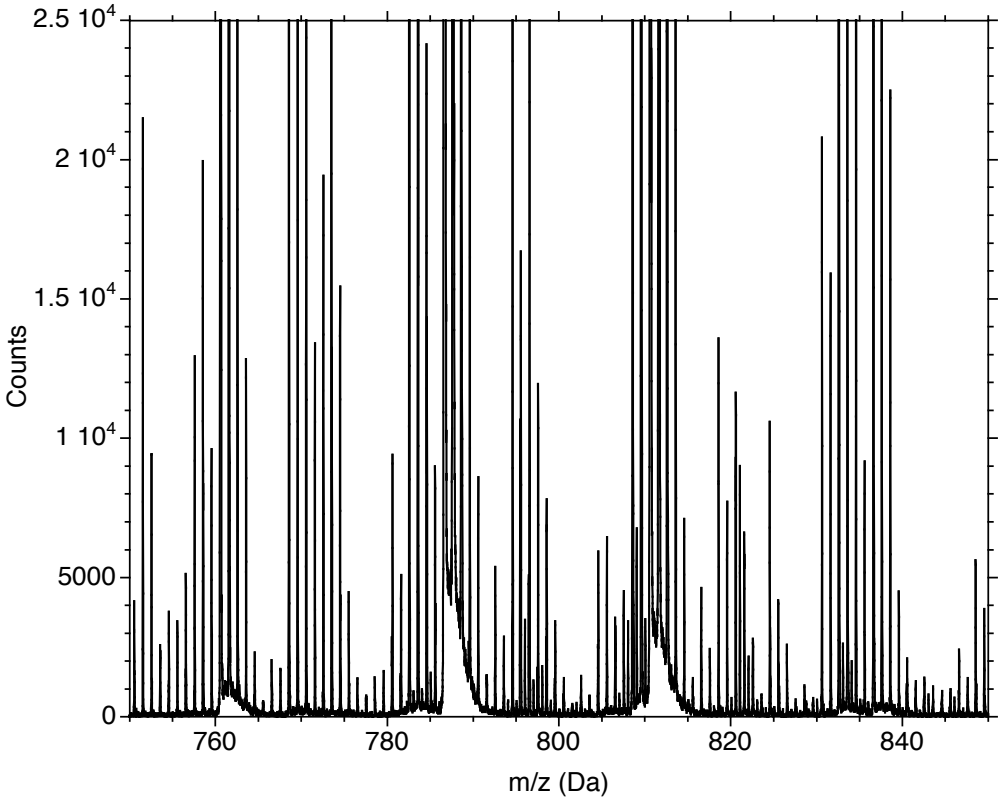


Figure 5.2. A portion of a single ESI-TOF lipidomics 3s scan from an LC/MS series where ion over-abundance causes a localized three order of magnitude baseline rise around the higher abundance peaks in the spectrum.

Orbitrap Spectra

Orbitrap mass spectra are unique in that the enhanced Fourier Transform (eFT) technique used to construct the mass spectrum effectively reduces the ion harmonics to less than 1% of the parent peak abundance.[4] Details of the eFT technique are beyond the scope of this paper. The effect, however, is that the resulting Orbitrap mass spectra appear to have a constant zero baseline.

FT-ICR Spectra

Traditional FT-ICR mass spectra are similarly produced by Fourier Transform of the ICR time domain signal, like Orbitrap spectra. The mathematics of this transform are beyond the scope of this paper, but the resulting spectra produced contain both noise and ion harmonics. Two types of FT spectra can be produced. In Magnitude spectra (Figure 5.3), the baseline floats above zero. In Absorptive spectra, the median counts are effectively subtracted from the magnitude spectrum, shifting the spectrum downward to straddle the x-axis at zero counts, with the option to trim (set to zero) negative counts from the resulting spectrum.

[4] Lange, O. et al., "Enhanced FT for Orbitrap Mass Spectrometry", *Int. J. Mass Spectrom.* ***369****: 16-22 (2014).*

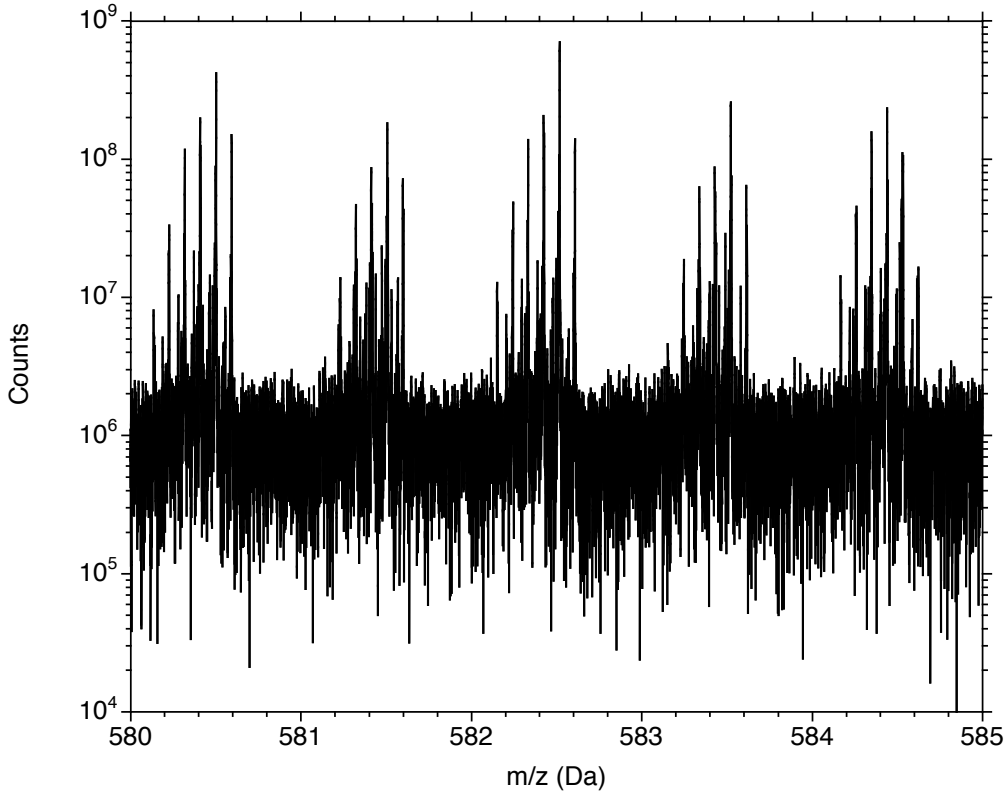


Figure 5.3. A portion of a FT-ICR magnitude spectrum of a petroleum sample (log abundance mode), showing the Fourier transform noise around some sample peaks. The median baseline of the Fourier Transform magnitude spectrum is offset from zero by ≈10^6 counts. A corresponding absorptive spectrum would effectively relocate the median baseline to zero counts, generating negative noise peaks that may or may not be trimmed from the spectrum.

Information Content in the Spectral Baseline

The baseline offset from zero and the high frequency variation about that offset provide useful metadata about the spectrum. There are various methods that can be used to estimate the baseline of a spectrum (described below). The high frequency variation about this baseline, particularly in regions devoid of real analyte peaks, provide an estimate of the general instrument and chemical noise that should overlay all peaks in the spectrum, providing useful signal-to-noise metadata.

The goal of spectral baselining is to eliminate low frequency noise (i.e., abundance variation over m/z scales larger than the width of the analyte peaks), while preserving the high frequency noise (i.e., abundance variations too narrow to be considered analyte peaks). A secondary goal is to establish the baseline in a way that the amplitude of the high frequency noise can be estimated, so that the useful signal-to-noise metadata can be applied for improved peak discrimination in downstream processing.[5]

[5] Spectral Signal-to-Noise Determination.docx

There are several options that people use in mass spectral data collection that should be avoided. Any technique that destructively removes spectral information (e.g., thresholds, smoothing, or centroiding) should be avoided as it necessarily limits metadata collection useful for further analyses.

Centroid Mode

The data acquisition software of all mass spectrometers often provide various options that unfortunately can remove data useful for establishing the true baseline and signal-to-noise limits for peak detection. For example, acquiring data in centroid mode eliminates all the natural spectral variation that can be found in regions that contain no peaks of interest, eliminating it from consideration by various baselining and signal-to-noise estimating algorithms. This may be convenient for the user because it makes the spectrum smaller, saving storage capacity and digital transfer times to move the data to other devices. It also removes options or eliminates degrees of freedom for the user in the analysis of their data, by removing all the metadata about signal-to-noise and mass and abundance precision that are useful during downstream processing. The typical assumption is that the instrument manufacturer can ascertain better than the user: what is a peak and what is noise. However, in a regulatory environment, how do you prove that is a correct assumption and why would you forever give up the ability to apply alternative or future peak detection processing options to your data?

Smoothing

Spectral smoothing is a common technique used to suppress high frequency noise. Smoothing can be applied directly to the counts of the spectrum, or to the derivatives used for peak detection through the use of higher order finite difference calculus. There are a wide variety of smoothing techniques, a discussion of which is beyond the scope of this paper. However, all smoothing effectively lowers the resolution of the spectrum by widening all peaks. The amount of resolution loss depends on the method applied. The additional challenge is how to smooth the spectrum when both peak shape and the frequency of the noise defined as x-axis points (or detector bins) vary with m/z (such as in TOF, Orbitrap and FT-ICR spectra)[6]. Fundamentally, however, all smoothing is destructive to the high frequency noise.

Thresholding

Sometimes the data acquisition software allows the user to specify a threshold below which it does not record any spectral data, as a method to limit the data storage and any subsequent data transmission requirements. Doing so also removes metadata useful for setting the true baseline and the inherent signal-to-noise in the spectrum, and permanently removes smaller real peaks below the imposed threshold that could otherwise be identified from the resulting spectrum.

Data Compression

Most instrument manufacturers automatically compress the spectral data, even in profile mode, by removing zero count points or using a variant of the Lempel-Ziv-Welch compression technique[7]. Data compression by these techniques is non-destructive, because the removed points can be replaced exactly by spectral decompression once the intrinsic data point spacing and compression type are known. However, these points must be replaced in the spectrum before any baselining method or signal-to-noise determination can be accurately applied.

---

6 Spectral Characteristics.docx [more complete reference?]

7 Lempel-Ziv-Welch File compression, https://en.wikipedia.org/wiki/Lempel–Ziv–Welch (accessed 23 June 2016).

Abundance-Based Peak Discrimination

Fundamentally, mass spectra consist of abundance values at given mass positions. However, there are additional metadata that can be brought to bear in baseline determination.[8] This includes: 1) how the intrinsic peak shape varies with m/z, which can be used to distinguish high frequency noise and low frequency baseline variation from peaks; 2) how the point spacing varies with m/z, which can be used for pattern-based noise detection algorithms; and 3) that noise is always positive (one sided) in mass spectra (i.e., there is no such thing as a negative count). This latter characteristic of mass spectra intrinsically eliminates the applicability of most parametric statistical approaches that might work well in other spectroscopy or signal analysis measurements where noise is randomly distributed about a true value.

True and False Peaks

As an example, we can centroid the TOF spectrum above (Figure 5.2) with no baseline or threshold constraints. Since this scan is part of an LC/MS run, we can also centroid the immediately preceding and following scans of that same run. If a peak is found to exist in the center scan and in at least one of the two adjacent scans within 50% abundance, that peak is considered a positive detection event. If there is a peak that appears in both the adjacent scans within 50% abundance, but not in the center scan, that peak is considered a negative detection event. Histograms of the distributions of both the positive and negative peaks found in the center scan by their abundance can then be determined (Figure 5.4).

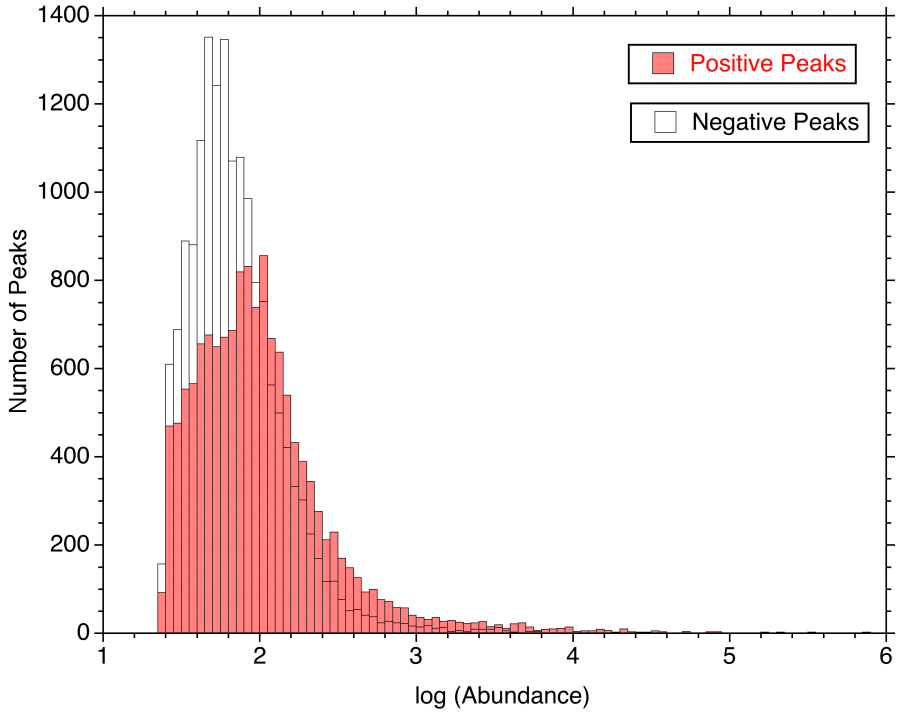


Figure 5.4. Histograms of peaks found in the spectral example of Figure 2 as distributed by their abundance. Positive peaks are those detected in either the immediately prior or following scan of the LC/MS run within 50% of the target peak abundance. Negative peaks are those found in

[8] see section on Spectral Characteristics

both the surrounding scans, but not found in the center scan within 50% abundance of either surrounding scan.

What immediately becomes apparent from these distributions is that there is very little abundance discrimination between true positive peaks and false peaks in this spectrum. Some of this overlap is due to the baseline float around peaks of high abundance (Figure 5.2), which adds counts to any noise peaks found in that local vicinity. Some of this overlap is due to low frequency non-linearities or "float" in the baseline (Figure 5.1, near 1200 Da), which also adds extra counts to peaks in these regions. Much of the overlap, however, results from noise superimposed on the sides of larger peaks, which many centroiding methods detect as a peak with its abundance estimated from its apex to zero. Similarly, noise inflation due to the summing of multiple scans can artificially raise the abundance values of all centroids drawn from zero counts.

Proper baselining minimizes or eliminates the contributions of low frequency baseline variability to the error in centroid abundances of the peaks found. More importantly, establishing a proper baseline allows the signal-to-noise limits to be accurately estimated.

Baselining Algorithms

A wide variety of algorithmic methods have been proposed and applied to the baselining challenge of mass spectra. These methods can generally be clustered into a few distinct categories: polynomial regressions of increasing orders (including the zero order fit of mean or median counts), window-based local baselining approaches (both static and dynamic), or asymmetric Whittaker smoothing and its derivatives.

Iterative Polynomial Regressions

One common approach is to find the least-squared polynomial regression of the mass domain to the abundance or log(abundance) domain. The least squares polynomial order may be determined by standard statistical methods, either χ^2 for the zero order fit (average or log[average]) or goodness of fit or incremental improvement ANOVAs for all higher order polynomials[9]. The least squares polynomial fit, however, tends to increasingly overestimate the baseline as more peaks, or peaks of greater height, appear in the spectrum (Figure 5.5).

[9] Zar, J. H., *Biostatistical Analysis*, pgs. 268-273 (Prentice Hall, 1974).

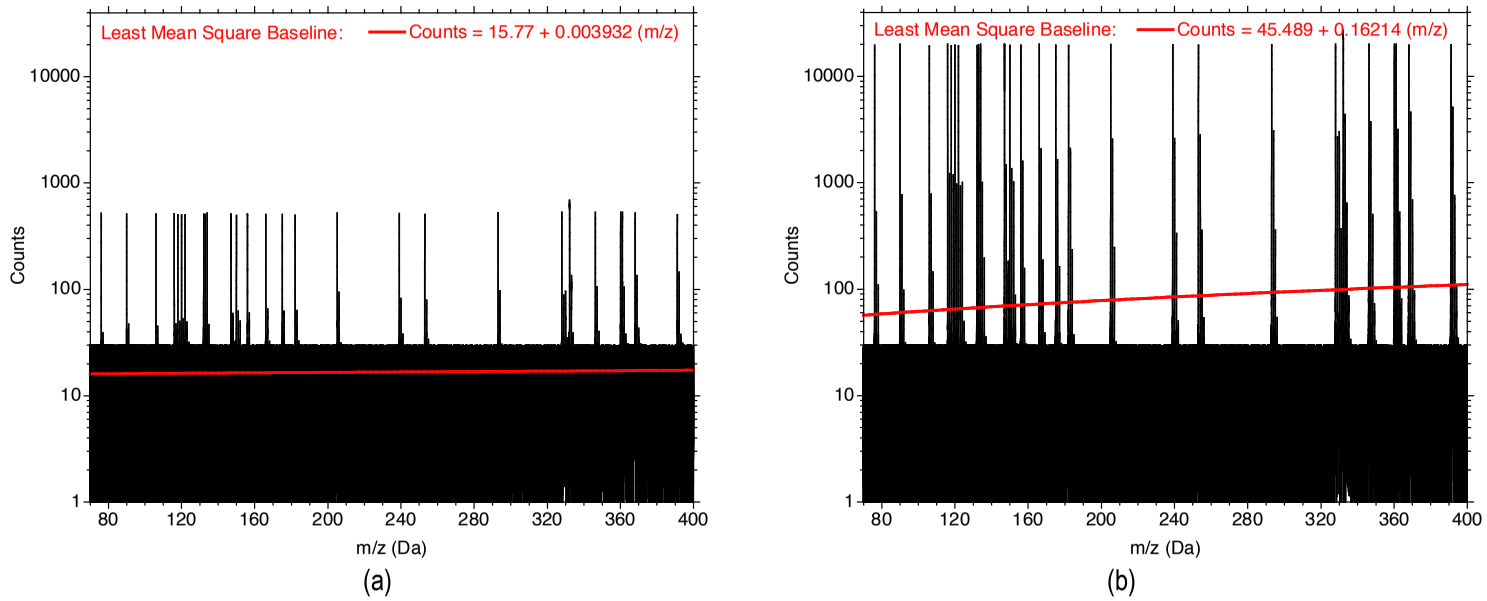


Figure 5.5. Least mean square fit of an O(1) polynomial baseline to (a) a synthetic TOF spectrum with 15 Count average noise and 500 Count high peaks, and (b) the same spectrum with 15 Count average noise and 20,000 Count high peaks.

Variations on the least mean square polynomial regression have been proposed to solve this overestimation issue, such as weighting the data inversely to the highest abundance within a mass window (i.e., discounting m/z regions that contain the larger peaks). However, this adds the user-adjustable parameter of an appropriate window size to the regression, which can be variable with m/z, depending on the mass analyzer. It is also possible to ignore the amplitude of the peaks and noise entirely by applying a least median square regression. However, least median squared regressions have multiple valid solutions, some of which are clearly suboptimal (Figure 5.6), with no statistical recourse to determine the global optimum.

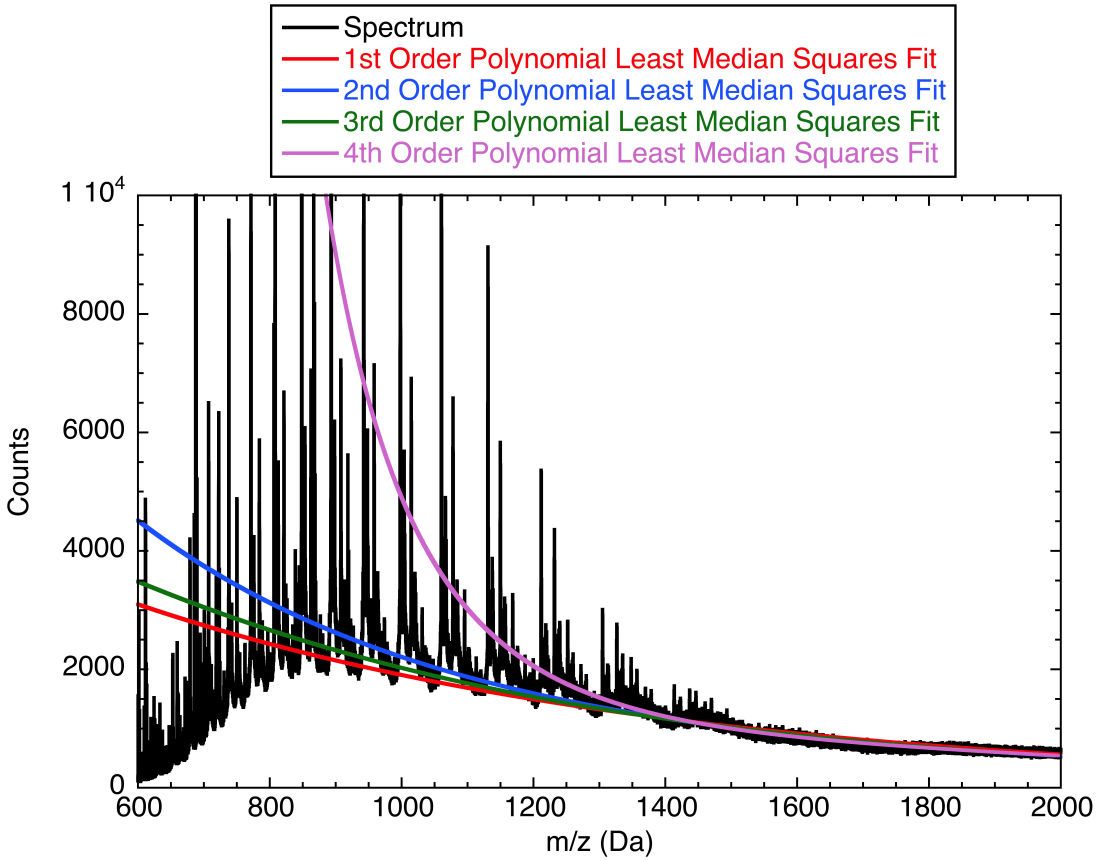


Figure 5.6. The Log_{10} Least Median Square Polynomial fits obtained using the MASS library of the R.app for the ESI-TOF spectrum of myoglobin. Those polynomials higher than O[5] converged back to the O[1] solution with no net improvement in the quality of the baseline and failed to follow the observed baseline curvature. However, in each case, the baseline solution was indeed a Least Median Squares fit with equal numbers of points above and below the fit, illustrating that multiple possible valid solutions can be obtained for higher order polynomials by the method of Least Median Squares.

We note here that the zero order polynomial solution is the same as a flat baseline (mean, weighted mean, or median) through the spectrum. This very common mass spectral baselining method is therefore considered a part of the polynomial regressions approach.

Another variation on the polynomial regression method is to remove all points above the regression line and perform the regression again, iterating fits and higher abundance point removal for a constant polynomial order until there are either too few points left to perform a

regression of that order, or a user-specified maximum number of iterations are reached.[10] However, standard parametric goodness of fit statistical methods fail to predict the optimal number of iterations. It is readily shown that the final result for any polynomial order ends up being one data point more than the order of the polynomial (a degree of freedom limitation of regression analysis). In the case of a zero order polynomial, the resulting baseline is always the lowest abundance value in the spectrum by this method, unless the number of iterations is restricted *a priori* by the user.

Rubber Band Baseline

In Rubber Band baselining[11] the spectrum is divided into even user-determined increments. The local minimum is determined within each increment and is subsequently used as a support point for the baseline. The baseline between these support points can be obtained by linear interpolation connecting the dots (i.e., a rubber band stretched over anchor points), a cubic spline, or a polynomial regression. The challenge in this method is to define a window size (x-axis increment) that is large enough to negate the impact of high frequency noise, large enough to not allow the baseline to rise into the center of real peaks, but not so large that the baseline fails to track the low frequency undulations of the spectrum.

Classic rubber band baselining is implemented with a constant $\Delta m/z$ window size, which only works for ion trap spectra. If the window is alternatively defined as a constant number of points (detector bins), then the method can be successfully adapted to Orbitrap spectra since the peak width is constant in that domain[12]. However, neither of these baselining methods are suitable for the variable window widths of either TOF or FT-ICR spectra.

The next problem is determining what measure of peak width to use for the window size: full peak width at half maximum height (PWHH), or peak width at the baseline (which becomes a "chicken-or-egg first" problem when being applied to baseline determination)? Perhaps the optimum window size lies at some multiple of the peak width? Day-to-day analyzer tuning variations can cause shifts in the intrinsic peak shape, as do any changes in the working mass range of the analyzer. Finally, how does skewness or kurtosis in the intrinsic peak shape affect the optimum window size? Since there is no statistical guidance for the optimum window size, it can only be set by user judgment and needs revision every time the mass analyzer settings are changed.

Moving Average Methods

While the rubber band method divides the spectrum into fixed increments, it is similarly possible to move the analysis window by increments through the spectrum, like a rolling ball[13], to get a local minimum, median[14], or abundance-weighted average for every detector bin (point) in the spectrum. As with the rubber band method, there is no statistical guidance for how to set the appropriate window size. Logically, it must be larger than the baseline width of any single peak or the resulting baseline will rise inside every peak. This means that the method can potentially be applied to ion trap and Orbitrap spectra, but would need m/z-adaptive peak width

---

[10] Lieber, C. A., Mahadevan-Jansen, A., "Automated Method for Subtraction of Fluorescence from Biological Raman Spectra", *Applied Spectroscopy*, **57**:1363-1367 (2003).

[11] Beleites, C., "Fitting baselines for spectra", https://cran.r-project.org/web/packages/hyperSpec/vignettes/baseline.pdf, (Mar 4, 2014).

[12] Spectral Characteristics.docx

[13] Kneen, M. A., Annegarn, H. J., "Algorithm for fitting XRF, SEM and PIXE X-ray spectra backgrounds", *Nuclear Instruments and Methods in Physics Research Section B*, **109-110**:209-213. (Apr, 1996).

[14] Friedrichs, M.S., "A model-free algorithm for the removal of baseline artifacts", J Biomol. NMR, 5:147-153 (1995).

information to be applied to TOF or FT-ICR spectra, where the peak widths are not constant on a Δm/z or bin number spacing.

The first question is: should be window be centered around the spectral peak, or skewed to the left or right from the apex, depending on the shape of the mass spectral peak? The second question is: should the baseline produced consist of the series of local minima, be inversely weighted by the maximum apex height in the window, or consist of the percentile cutoffs (e.g., median) of the abundance values in these moving windows? Finally, how are the moving average points to be connected to form a coherent baseline (e.g., regression or what order, cubic spline, etc.)? Each of these user-adjustable parameters opens up a plethora of possibilities that cannot be inferred statistically from the spectral data and, therefore, require user judgment to set.

Penalized Least Squares Smoothing

Perhaps the most common method for time series noise reduction is the Whittaker smoother.[15] Designed for evenly spaced data, the Whittaker smoother attempts to both fit a set of data (y) with a cubic spline model (μ[x]) by minimizing the least squares residual error that represents the raw data, but penalizes that model, if subsequent points of the model vary too much (i.e., the finite difference rate of change in the model shape with x is large). An arbitrary Lagrangian multiplier (λ) defines the relative contribution of the cubic spline and finite difference slope to the final model objective function (SSE):

$$SSE = (1-\lambda)\sum_i w_i (y_i - \mu_i)^2 + \lambda \sum_i \left(\delta^2 \mu_i\right)^2$$

The Whittaker objective function (SSE) for minimization consists of two parts. The first part is the standard sum of squares residual error from the regression model $(y_i - \mu_i)^2$, which results in re-creation of the spectrum as a cubic spline when λ—>0. The second part of the objective function consists of a local approximation (by difference equations) of the second derivative of the regression model $[(\delta^2\mu_i)^2 = (\mu_{i-1}-2\mu_i+\mu_{i+1})/(\Delta x)^2]$. Where the direction of previous and subsequent points is unchanged (along the trajectory of a line) the second derivative (i.e., change in slope) goes to zero.. Where the trajectory of the line changes, the square of the second derivative is always positive, irrespective of the direction of that change. The magnitude of the magnitude of the change in the second derivative increases the more that a series of model points deviates from linear. In the classic Whittaker smoother, the weighting function w_i is set to 0.5 for both positive and negative values of the residual error $(y_i - \mu_i)$.

Asymmetric Whittaker Smoothing

The Whittaker algorithm is, in essence, a data smoother that reduces the magnitude of high frequency noise in the data. To the Whittaker smoother, peaks in a mass spectrum are effectively just additional high frequency noise. However, the objective function can be made asymmetric to create a smoothed baseline, by providing different weights to the residuals that are greater than the Whittaker smoothed model than those that are less than the model points (i.e., where $w_i = p$ where the residual is positive and $w_i = 1-p$ where the residual is negative).[16] At p=0.5 all residuals contribute equally to the residual error at any value of λ. As p—>0, the Whittaker smoother moves progressively closer to the local minimum of the spectral abundances and starts to approximate a baseline (Figure 5.7).

---

[15] Whittaker model, https://en.wikipedia.org/wiki/Whittaker_model (accessed June 29, 2016).

[16] Eilers, P. H. C. and Boelens, H. F. M., "Technical Report: Baseline Correction with Asymmetric Least Squares Smoothing", Leiden University Med. Center Report, http://zanran_storage.s3.amazonaws.com/www.science.uva.nl/ContentPages/443199618.pdf (Oct 21, 2005).

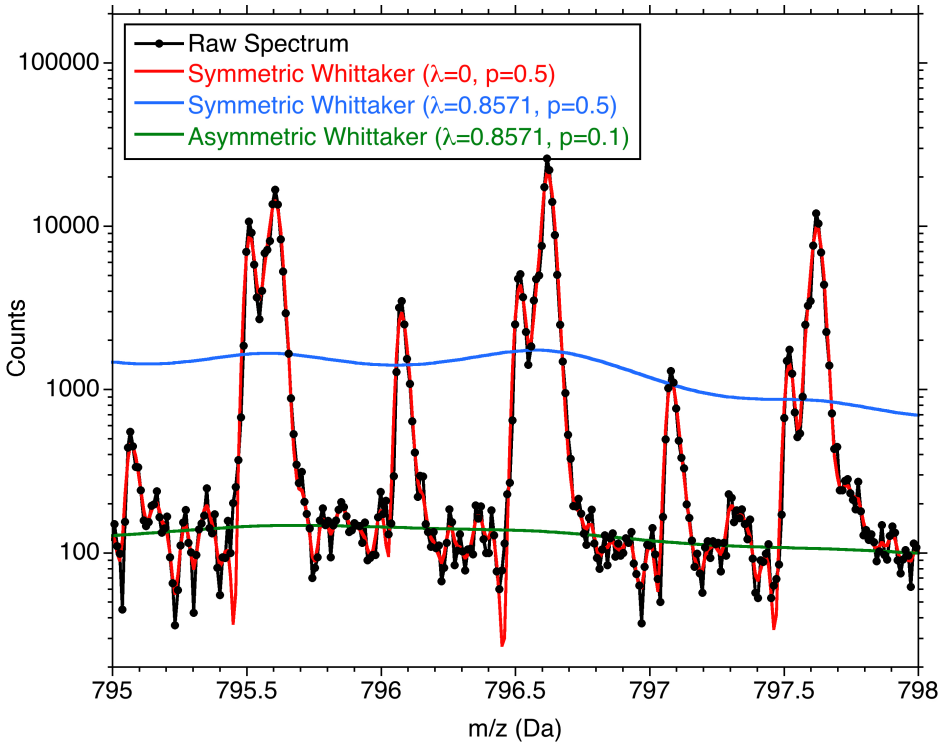


Figure 5.7. A portion of the ESI-TOF spectrum (Figure 5.2) showing the results of different parameters used in the asymmetric Whittaker smoothing function. Where λ—>0 the Whittaker smoother becomes a cubic spline approximation of the raw spectrum. At higher values of λ, the smoother approximates a local average of the spectral abundance data where positive and negative residuals are weighted equally (i.e,. p=0.5). As p is lowered, preferentially weighting against the higher abundance points, the smoothed spectrum starts to approach a spectral baseline.

The asymmetric Whittaker smoother replaces the unknown of optimum window size in the rubber band and moving average baseline methods with two more obscure parameters, the Lagrange multiplier (λ) and a residual weighting function (w_i). Unfortunately, neither adjustable parameter can be inferred directly or indirectly from the spectral data by statistical means. However, they appear to be less subject to tuning variations and mass range changes in the mass analyzer, versus the window size changes required in the previously-described methods. The asymmetric Whittaker smoother also appears to be a universally-applicable method to all mass analyzer types, since the result seems to be independent of peak shape or variation of peak shape with m/z.

It should be noted, however, that as λ—>1, the asymmetric Whittaker smoother model becomes the local weighted average abundance value of the spectrum, invariant with m/z. We have already shown (Figure 5.7) that as λ—>0 the raw spectrum is faithfully reproduced as a cubic spline). As a consequence λ <1. Furthermore, as p—>0, where p determines the residual weighting function (w_i), the asymmetric Whittaker smoother will approach the minimum iterative polynomial approximation of a constant order zero baseline at the lowest abundance value in the spectrum. As a consequence, $0 < p \leq 0.5$. While λ and p are therefore bounded, there is no further guidance on how to optimally set either user-adjustable parameter for any given spectrum via statistical means, and it is left to the user to follow their best judgment.

Combinatorial Methods

Various combinations of the above baselining approaches have also been proposed, such as the Filling Peaks method.[17] In this method, - Whittaker smoothing is applied to the spectrum using an arbitrary value for λ. The Whittaker smoothed spectrum is then subjected to rubber band minima determination with an arbitrary window size of n points. The rolling ball mean of the rubber band minima is then iteratively applied using an m times n window size (where m is an arbitrary multiple of n, $m*n <$ total points in the spectrum, and $m > 1$). In each iteration, those rubber band minima that are above the rolling ball mean are removed from the local mean calculation. The process continues for an arbitrary number of iterations. This technique contains 4 user-adjustable parameters (λ, n, $m > 1$, and number of iterations), none of which can be statistically-inferred from the spectral data.

PeakInvestigator™ Baselining

Thus, the problem remains unresolved: how to robustly baseline spectra in a way that non-destructively reads through the high frequency noise, adapts dynamically to the low frequency baseline variations, yet requires no user-adjustable parameters. Veritomyx, Inc. has developed a proprietary baselining method that accomplishes just that. The resulting PeakInvestigator™ baseline effectively provides a robust approximation to the local median counts, yet this is accomplished in a manner that avoids getting trapped in alternative local-minima solutions (as in Figure 5.6). While computationally complex, it is statistically-valid, universally-applicable, and accepts no user-adjustable parameters. Figure 5.8 shows the PI baseline automatically produced for the troublesome TOF spectrum of Figure 5.6.

The PI baseline produces a result comparable to that which might be obtained from a least median polynomial regression of optimum order and which is constrained to the global optimum result, or that might be produced with a moving median method with continuous local re-optimization of window size. However, it avoids any need for the user to set the proper polynomial order or to estimate the best window size as a function of m/z, as would be required by either of these methods (see above discussions).

PI baselining performs equally well with FT-ICR (Figure 5.9) and ion trap spectra (Figure 5.10). In the case of the FT-ICR spectrum, PI baselining was applied to both the magnitude and absorption mode versions of the same spectrum. When the resulting absorption mode baseline is scaled back to fit the magnitude spectrum, it overlaps the baseline obtained from the magnitude spectrum, suggesting that the PI baselining method is very robust to spectral amplitudes. Because Orbitrap spectra have an effective baseline of zero counts, they do not require additional baselining.

[17] Liland, K. H., "4S Peak Filling - baseline estimation by iterative mean suppression", *Methods*, **2**:135-140 (2015).

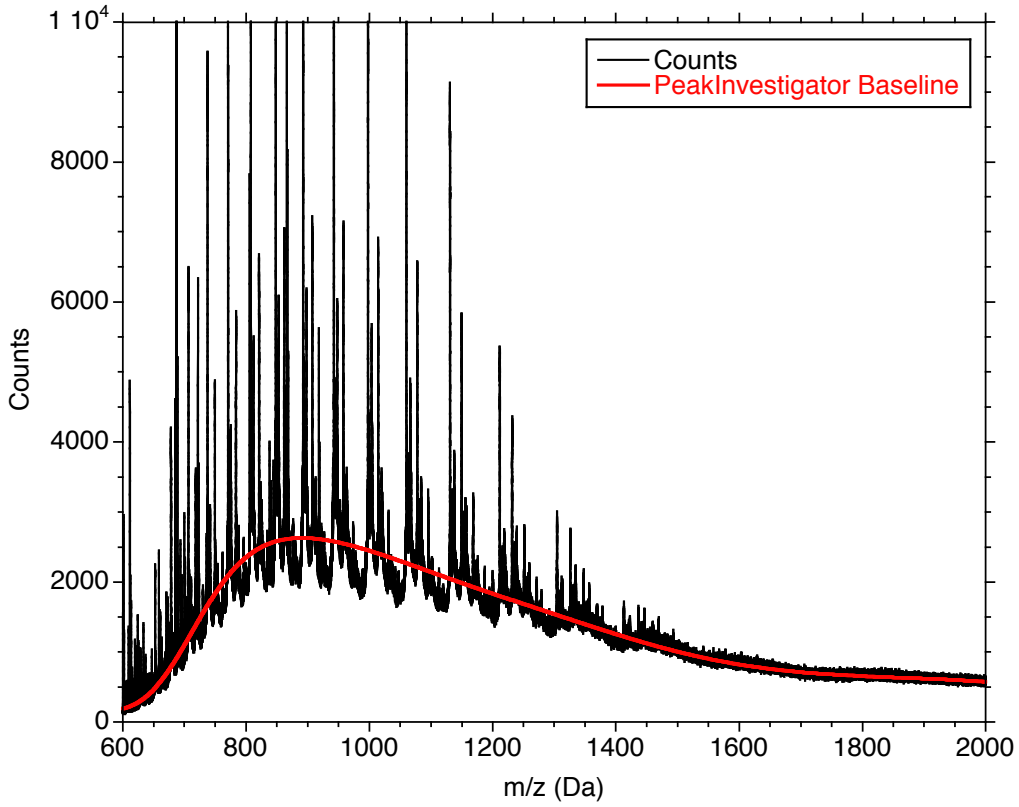


Figure 5.8. The PeakInvestigator™ baseline determined for the ESI-TOF spectrum of myoglobin shown in Figure 5.6. There are no user-adjustable parameters for the PeakInvestigator baseline. This baseline provides a local approximation to the median counts, yet avoids alternative valid solutions to the least median square regressions approach in Figure 5.6.

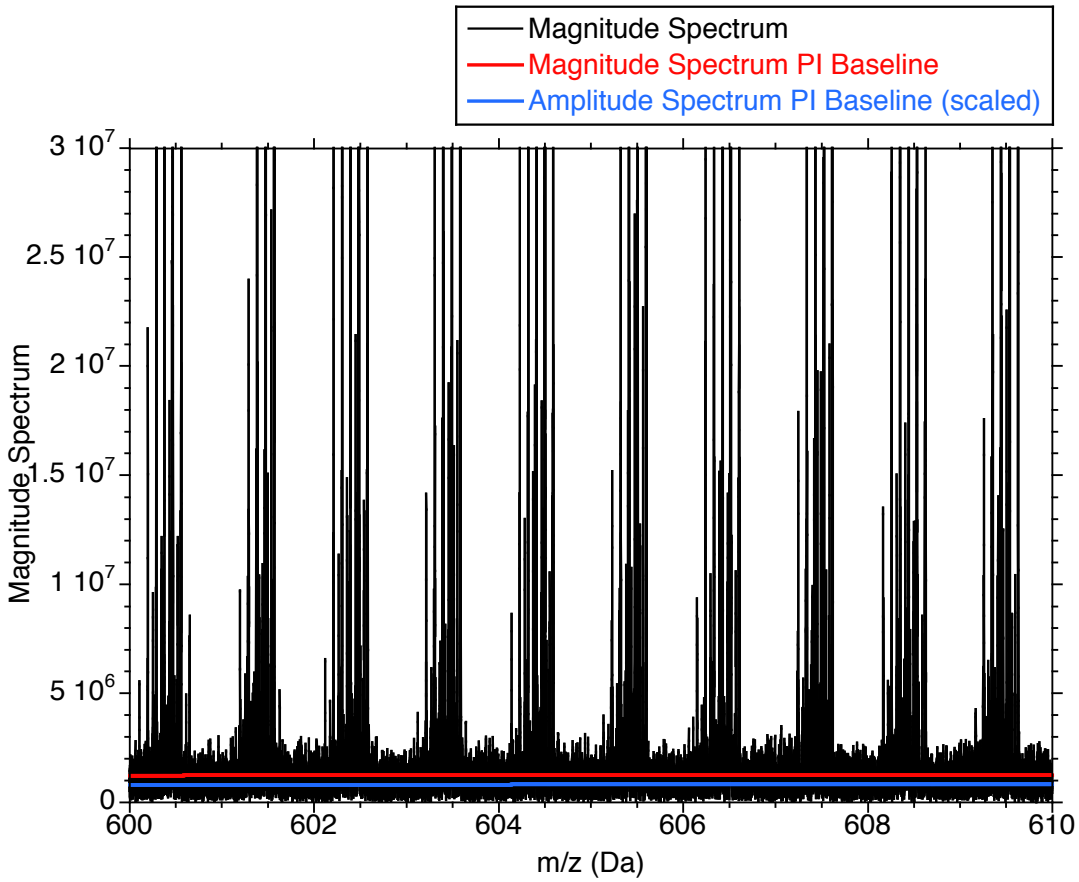


Figure 5.9. The PeakInvestigator™ baselines determined for FT-ICR spectra of petroleum samples. The figure compares the PI baseline obtained from the magnitude mode spectrum to that of the corresponding absorption mode spectrum (after scaling back to magnitude mode). The close overlap of these two baselines suggests that the method is very robust to changes in spectral amplitude over at least 5 orders of magnitude in abundance.

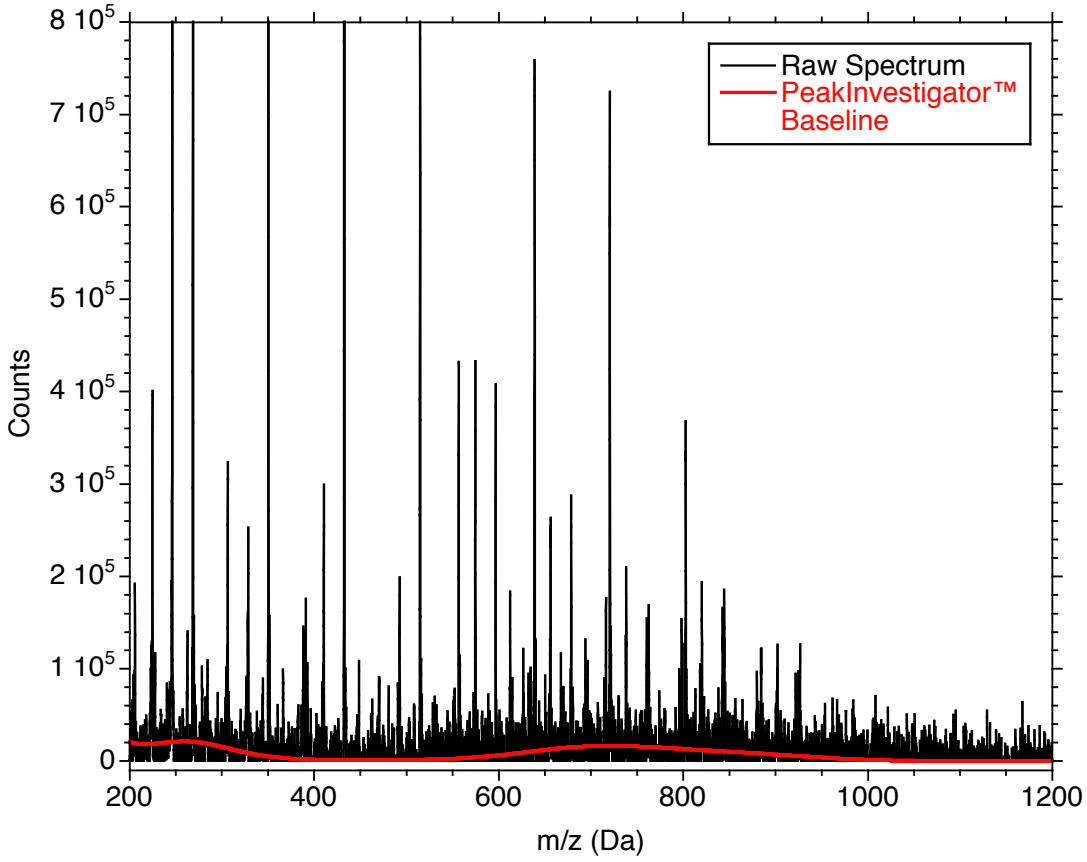


Figure 5.10. The PeakInvestigator™ baseline determined for an ion trap spectrum of lipids.

Baselining Limitations

All baselining methods, even baselining by eye, are built on the implicit assumption that the majority of spectral information is noise. If the spectral footprint of the real analyte peaks becomes the majority of the data over any large region of the mass spectrum, then the chemical and instrument noise in that region becomes drowned in a confusing jumble of overlapping analyte peaks. If the heavily-overlapped regions are small (i.e., the width of a few peaks), any robust baselining algorithm may traverse the region with only minor error. However, when such regions of ultra-high peak density start to cover 20, 50, 100, or wider Da spans of the spectrum, any baselining algorithm will increasingly start to confuse real analyte peaks with chemical or instrument noise.

This problem is more common in lower resolution mass analyzers (e.g., ion trap and unit-resolution TOF and quadrupole analyzers), where the peak width at baseline approaches 1 Da. The compositions of most ions below 1,000 Da generally produce peaks closely spaced (±0.1 Da) in a mass spectrum for singly-charged molecular ions because the maximum mass defect for any element is about 0.1 Da.[18] Therefore, singly-charged molecular ions should never produce a baselining problem at any spectral resolution above 3,000 (defined as mass divided by peak width at half height). Higher resolution mass analyzers are even more immune to this baselining limitation.

[18] Hall, M. P. et al., "'Mass defect' tags for biomolecular mass spectrometry," *J Mass Spect.,* **38**:809-816 (2003).

The problem of peak overlaps dominating the baselining algorithm primarily appears in the spectra of multiply-charged species at any resolution. Doubling the charge state halves the nominal 1 Da spacing between peaks. Triply-charged species reduces the interpeak spacing to 0.3 Da. Again, the higher the resolution, the less of a problem this poses to the baselining algorithm, but densely packed spectra containing mixtures of charge states can ruin the performance of any baselining method. This issue is generally remediated by better pre-separation of the analytes prior to mass spectral analysis.

6. SPECTRAL SIGNAL-TO-NOISE DETERMINATION

All mass spectra are confounded by chemical and instrument noise. The ability to discriminate true peak signals from this noise is a critical challenge in peak detection. The crossover between spectral data and noise, therefore, is an important piece of metadata to be determined within a spectrum. A statistically robust signal to noise threshold can be used to determine when a side peak is real or should be attributed to noise. It also aids in the establishment of thresholds for peak detection.

Chemical and instrument noise is not constant at every mass point in a spectrum. For example the matrix noise produced in MALDI ionization often underlies all analyte peaks in the low mass range of MALDI spectra (Figure 6.1). Even there, however, its contribution to the counts at any given mass varies (Figure 6.1, inset). In an FTICR spectrum, high frequency detector noise is readily visible as a jagged baseline surrounding all the analyte peaks (Figure 6.2). This variation in chemical noise about an average (or median) value is the key to establishing the spectral noise level.

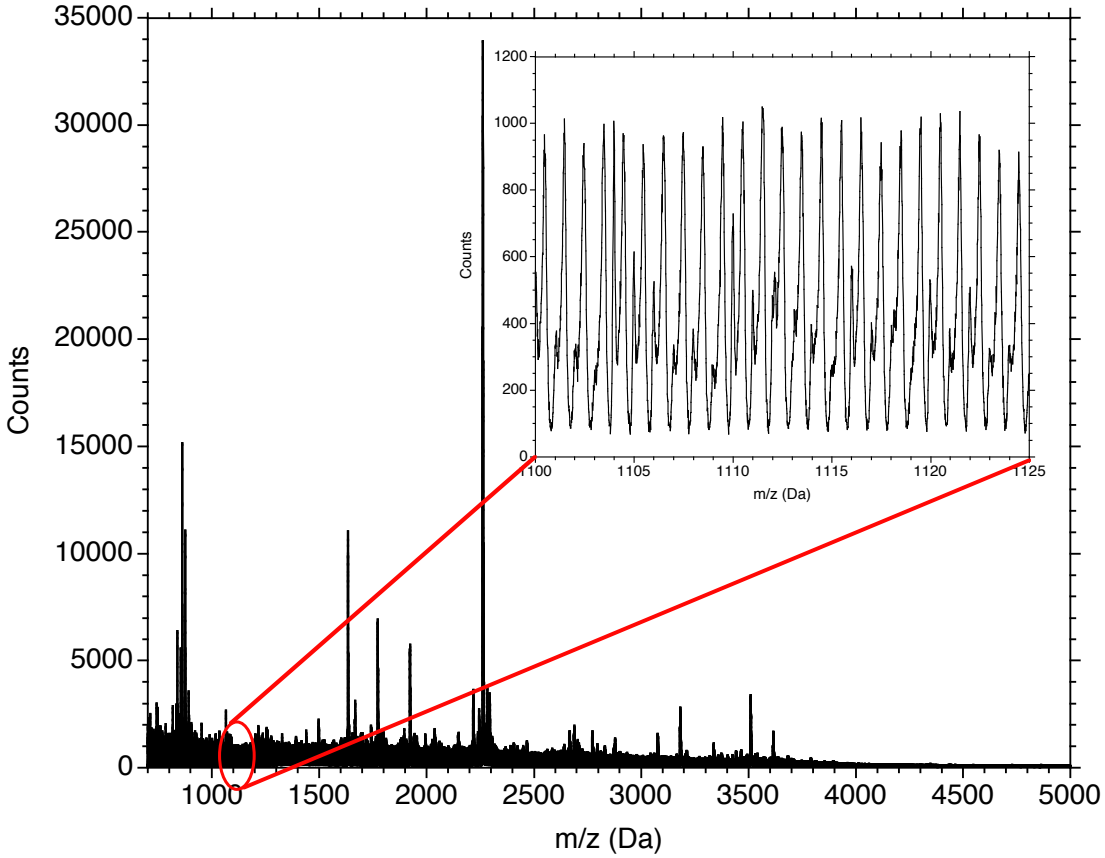


Figure 6.1. The matrix noise offset produced in a MALDI peptide spectrum. MALDI matrix noise becomes increasingly abundant at lower m/z, causing a baseline offset of the spectrum. The inset shows that the matrix contribution to the counts in any given mass point varies locally with mass, forming a repeating pattern.

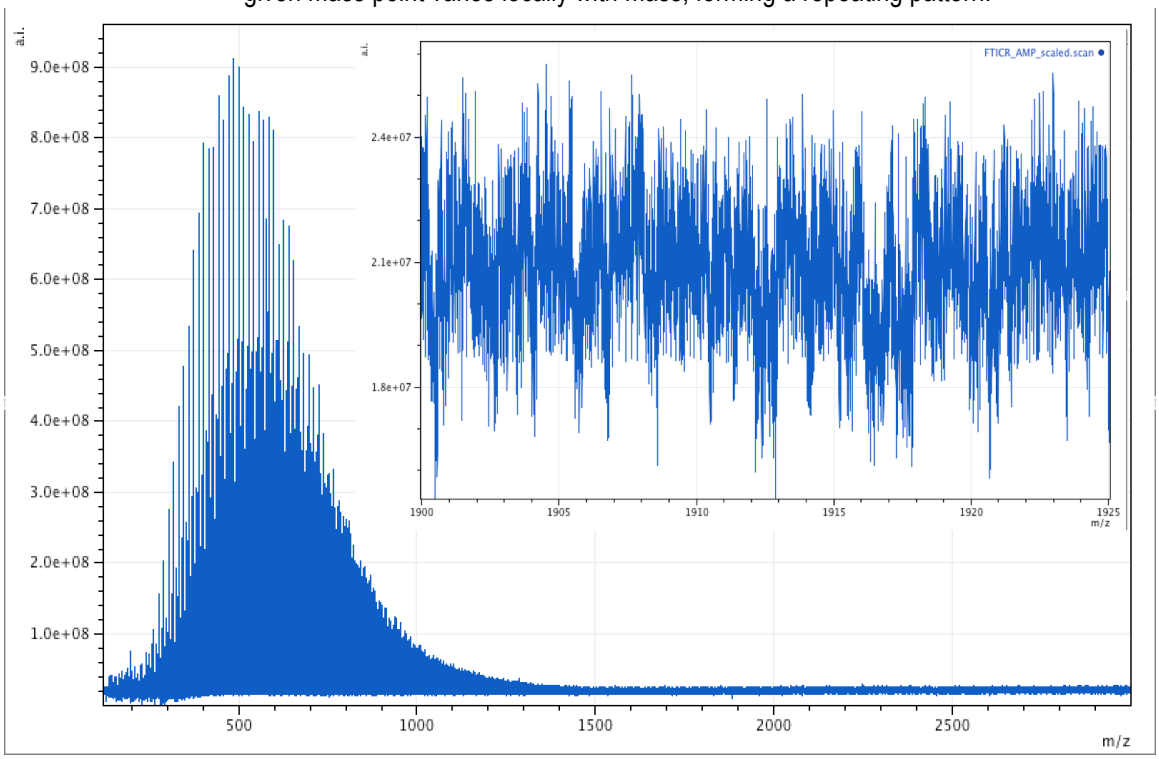


Figure 6.2. The high frequency detector noise in an FT-ICR spectrum of a petroleum sample.

Average Noise and Noise Variance

Mass spectral noise has two separable components: 1) the average noise, which lifts the baseline from zero counts because noise is always positive in a mass spectrum; and 2) the noise variance about this average. In a spectrum where the combined width of the analyte peaks represent just a small fraction of the mass range of the spectrum, the local median baseline effectively approximates the average chemical and instrument noise in the spectrum.

Assuming the analyte peaks cover a limited fraction of the total mass range, then the median baseline can be assumed to represent the average noise level in the spectrum. Since peaks generally extend above the median baseline, then any variance below the baseline typically represents mass points that accumulated less than the average noise. Those counts above the local median baseline include those mass points that accumulated more than the average noise counts, but also include the counts produced by the analyte species. Therefore, the difference between the actual counts below the baseline and the median baseline can be treated as effectively representing half of the actual noise variance.

Consequently, by subtracting the local median baseline from the spectrum, the spectrum is re-registered to zero baseline counts, eliminating the average noise offset. Those residuals that extend below zero counts represent the lower half of the noise variance and can be reflected to positive counts to provide an estimate of the level of chemical and instrument noise remaining in the baseline-subtracted spectrum.

For example, in the MALDI spectrum (Figure 6.1) the matrix offset in the low mass region can be effectively re-registered to zero counts by subtracting the local median baseline (Figure 6.3), as approximated by PeakInvestigator™. Half of the total variance in this matrix noise is then represented by the magnitude of the counts that extend below zero abundance. The other half, and the analyte peaks of interest, lie above the local median. If the pure noise variance below the median is reflected above the median (Figure 6.3), then we effectively estimate the overall chemical and instrument noise in this spectrum. This estimate is imperfect because the chemical noise in this MALDI spectrum is not randomly distributed but is in fact patterned (inset, Figures 6.1 and 6.3), but the trend in its abundance with m/z is clearly correct.

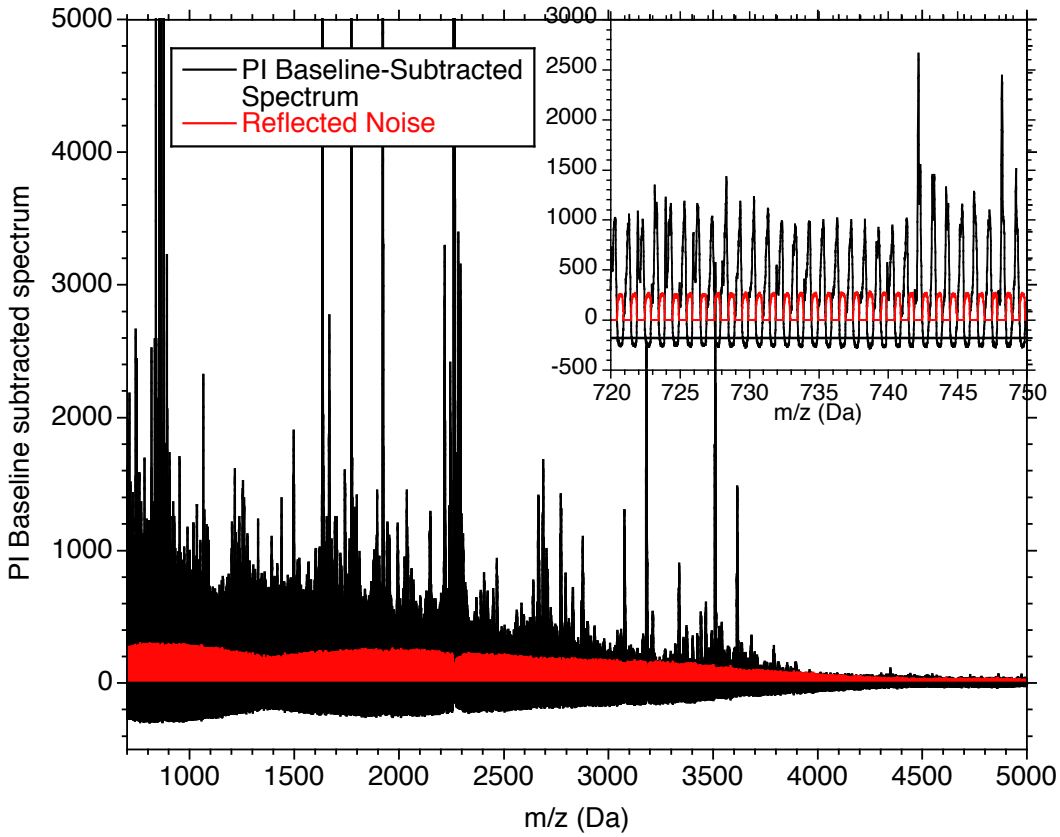


Figure 6.3. PeakInvestigator™ baseline subtraction of the MALDI-TOF spectrum re-registers the average noise counts to zero, eliminating the matrix offset created in the low mass range of the original spectrum (Figure 6.1). The peaks in the baseline-subtracted spectrum that extend below zero abundance effectively represent about half of the total variation in chemical and instrument noise in the spectrum. When these negative peaks are reflected above the median, an estimate of the noise is produced. In this case this is an under-estimate probably caused by the interlaced pattern evident in the MALDI matrix noise (inset). This matrix peak pattern is clearly narrower at the apex and wider at the base, lowering the estimate of a median baseline and biasing the s/n variance to the low side.

The spectral noise in the FT-ICR spectrum (Figure 6.2) appears to be more randomly distributed. After subtraction of the PeakInvestigator™ baseline and reflection of the residuals now below baseline, the local signal to noise is much better approximated (Figure 6.4) than that of the MALDI spectrum (Figure 6.3).

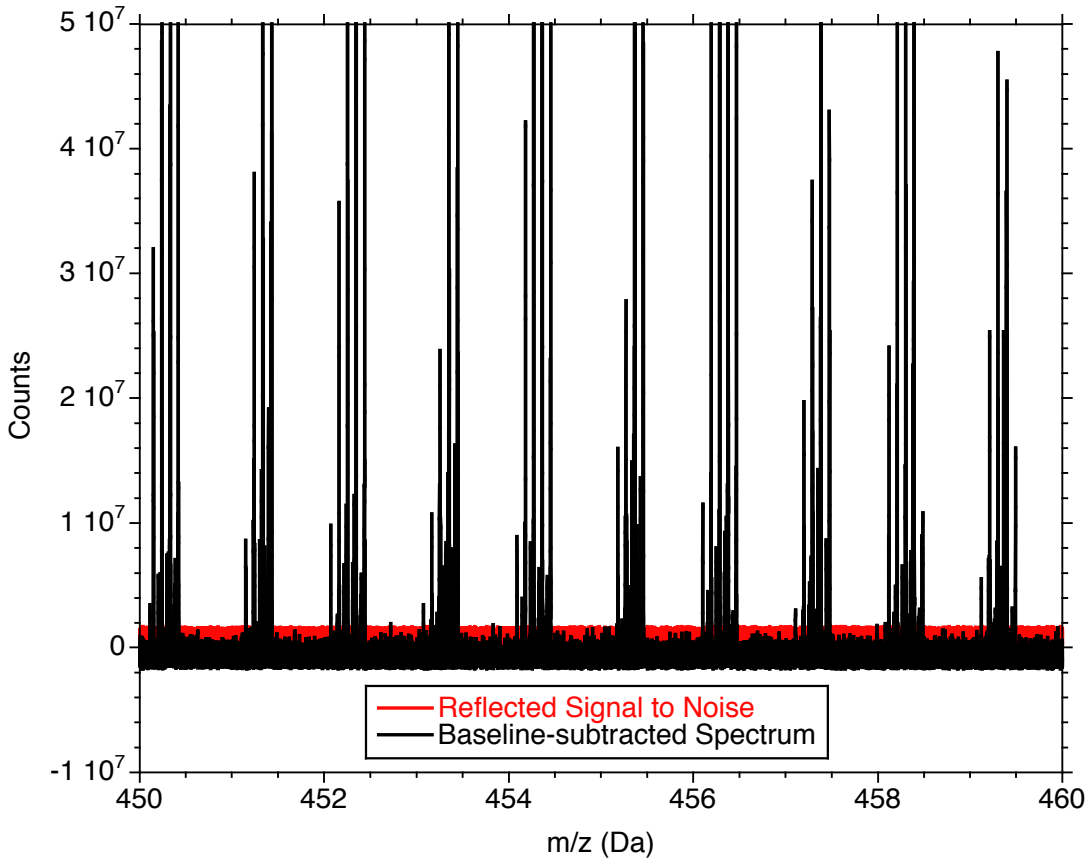


Figure 6.4. PeakInvestigator™ baseline subtraction of the FT-ICR spectrum (Figure 2) to re-register the spectrum to zero counts. The negative residuals are then inverted to estimate the upper limits of the chemical and instrument noise in this spectrum.

Statistical Signal-to-Noise Estimation

Where the noise variance is even throughout the spectrum, the statistical distribution of this variance can be determined by combining the negative and reflected noise distributions obtained from the baseline-subtracted spectrum. This is illustrated in the following examples.

ESI-TOF Example

In this example, an ESI-TOF spectrum from an LC/MS lipidomics series is the target scan (Figure 6.5). After baseline subtraction and noise reflection, the signal-to-noise variance is produced. Here, without the interfering matrix peak pattern common to MALDI spectra (Figure 6.1), the signal-to-noise levels approximated by this method appear close to that observed by eye (Figure 6.6). The variance is also relatively even throughout the spectrum.

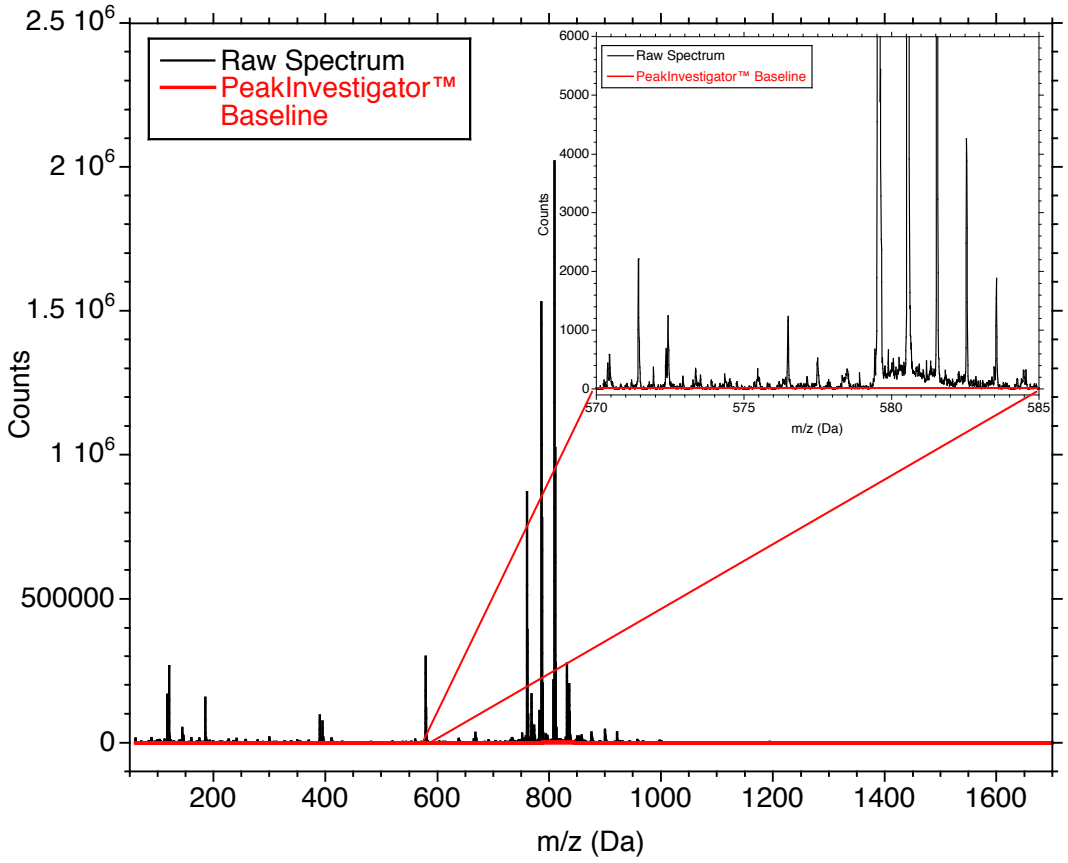


Figure 6.5. A single ESI-TOF spectrum taken at random from an LC/MS series obtained from a plasma lipidomics experiment. The inset shows the average noise offset and local noise variation around a series of analyte peaks.

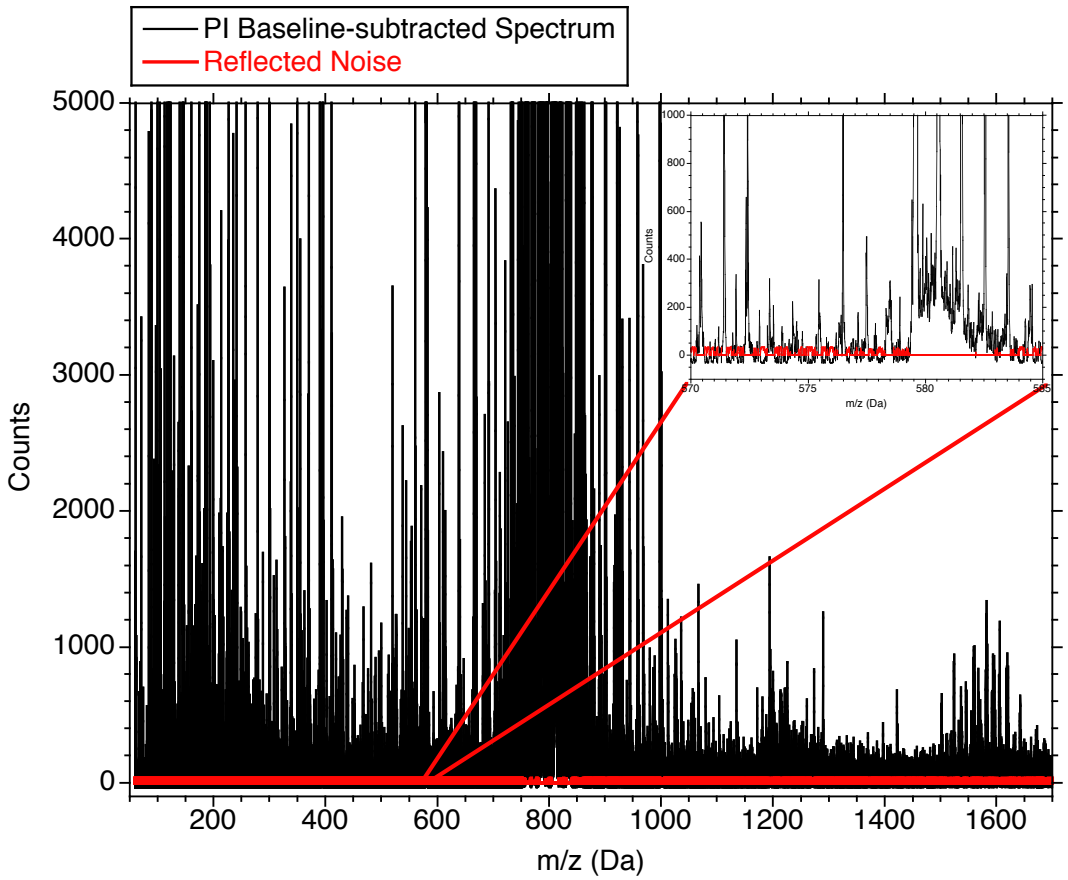


Figure 6.6. The spectrum of Figure 5 after PeakInvestigator™ baseline subtraction showing the reflection of the sub-baseline noise variance. It is observed that the noise variance thus estimated is roughly constant across the spectrum. The inset shows that the reflected noise variance agrees well with the actual spectral noise variance.

Since the noise variance is roughly uniform across the spectrum, a statistically-valid estimate of this variance can be obtained from its abundance distribution. Plotting the combined (positive and negative) noise variance abundances on a probability plot (Figure 6.7) shows that it is approximately normally distributed. Therefore, these variance data can be fit to a normal distribution, which allows the limits of chemical and instrument noise throughout the spectrum to be estimated with any statistical confidence desired.

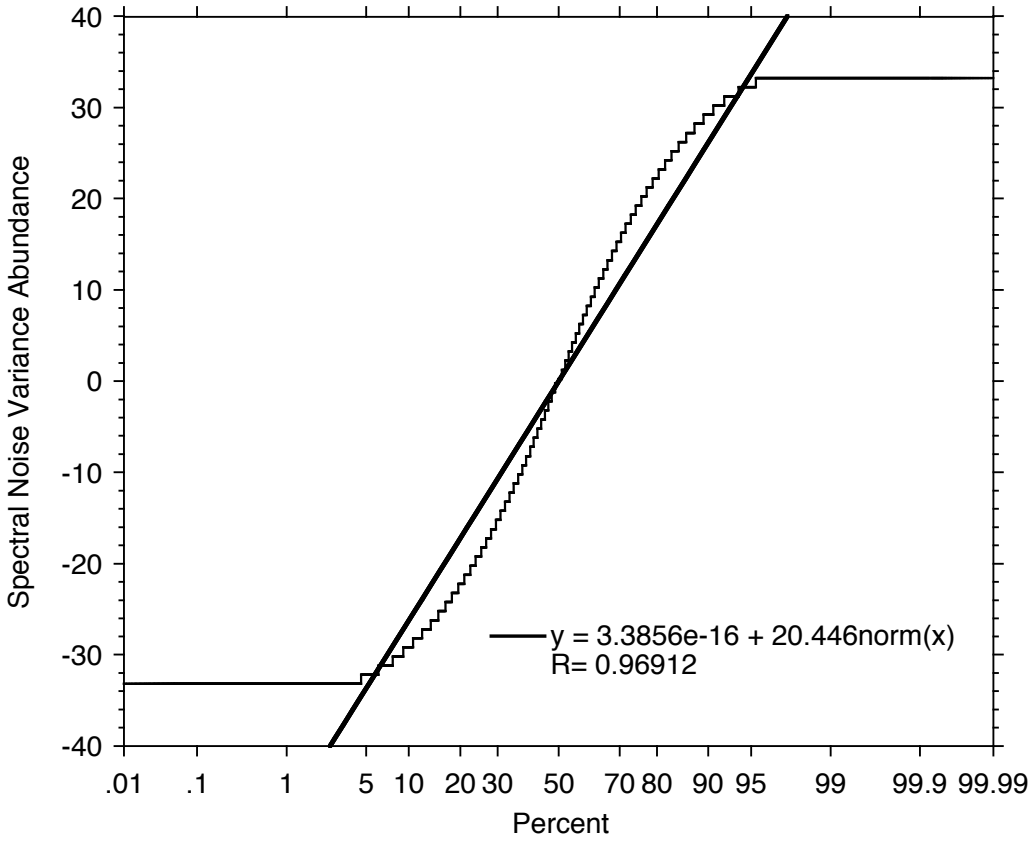


Figure 6.7. A probability plot of the ESI-TOF noise variance abundances obtained from the data of Figure 6. A normal distribution curve fit is shown that can be used to estimate the upper limit of the spectral signal-to-noise with any statistical confidence desired.

FT-ICR Example

Similar results are obtained for the FT-ICR spectrum shown above (Figure 6.2). In this FT-ICR spectrum the noise variance is seen to be approximately normally distributed (Figure 6.8) even though a slight bulge in the variance can be seen between 100 and 350 Da in the original spectrum (Figure 6.2).

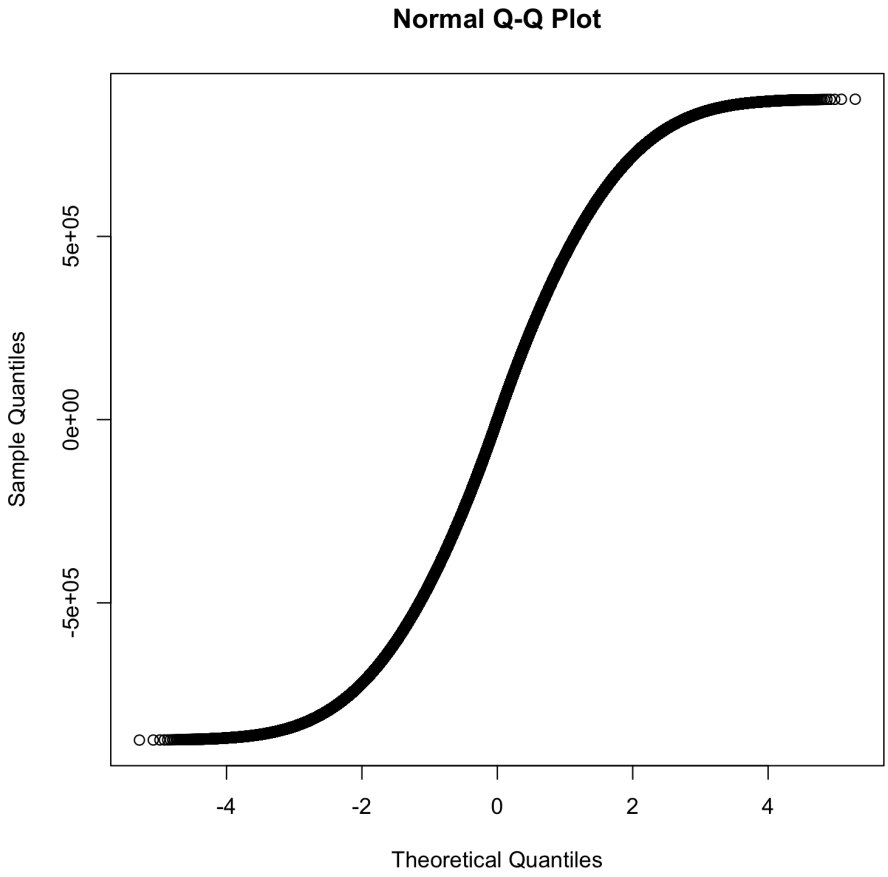


Figure 6.8. A probability plot of the FT-ICR noise variance abundances obtained from the data of Figure 6.2.

Ion Trap Example

In this example (Figure 6.9) we look at a tandem MS CID fragmentation spectrum of a peptide taken from a yeast peptidome LC/MS/MS study. The PeakInvestigator™ baseline is shown near 1 count. The resulting signal-to-noise distribution (Figure 6.10) is again seen to be nearly normally distributed by the method described above.

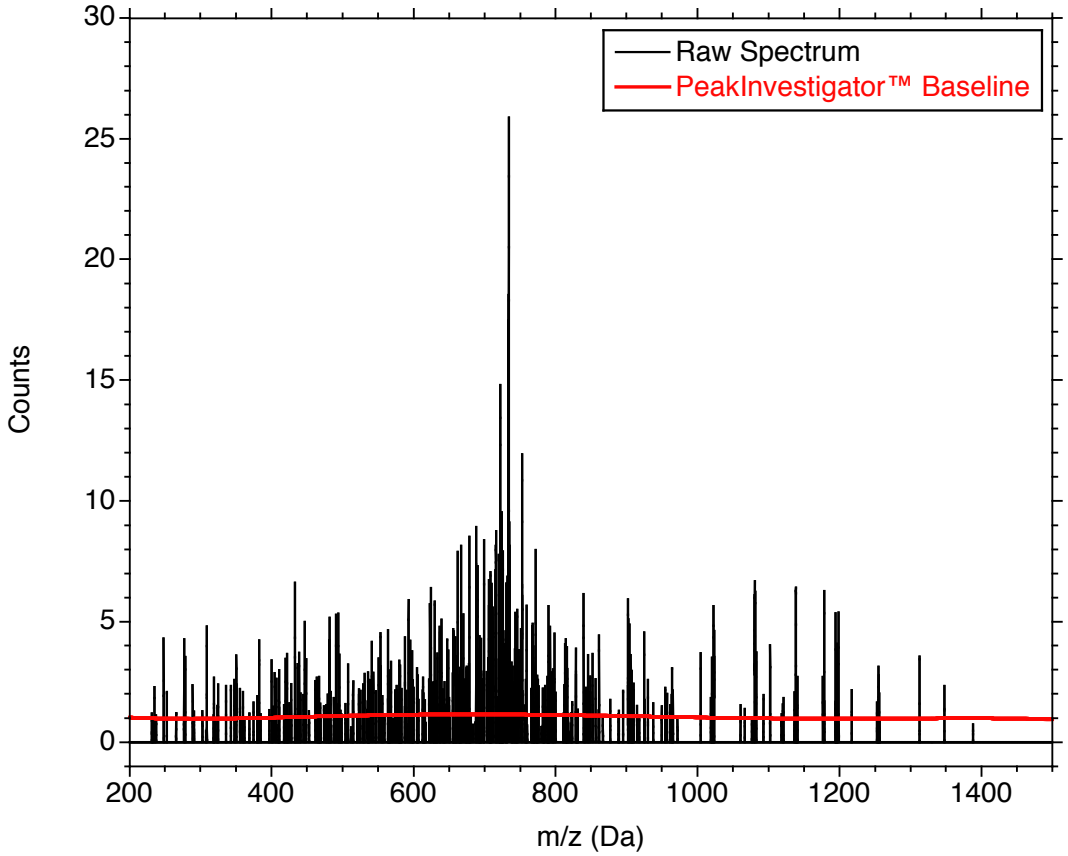


Figure 6.9. This is an example of a peptide fragmentation spectrum (MS2 scan) obtained from an ion trap analyzer. Tandem MS spectra typically have very low chemical and instrument noise, as is reflected in the PeakInvestigator™ baseline result near 1 count.

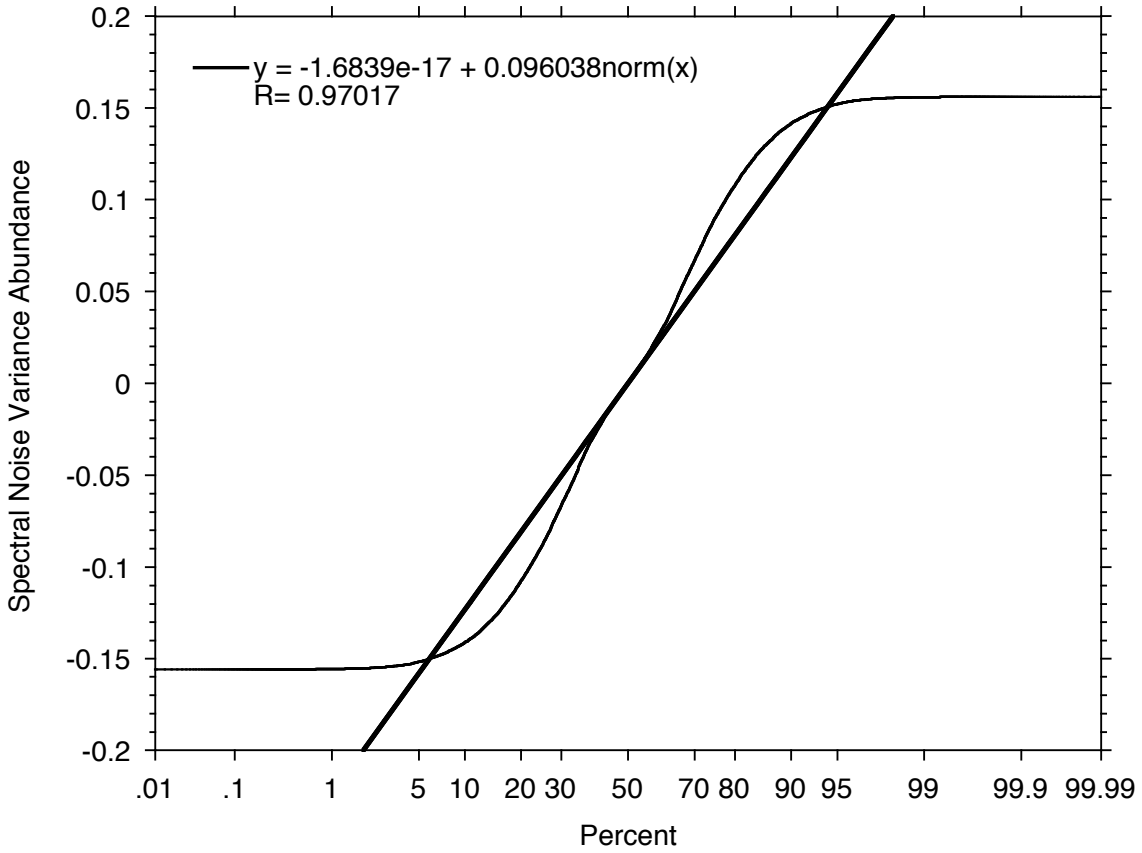


Figure 6.10. The signal-to-noise distribution variances are graphed for the below baseline variances and their above baseline reflection from the ion trap spectrum of Figure 6.9. A normal distribution curve fit is shown that can be used to estimate the upper limit of the spectral signal-to-noise with any statistical confidence desired.

Orbitrap

Orbitrap data outputs from the mass analyzer consistently provide an effective uniform baseline of zero counts.[19] Therefore, the above method for signal-to-noise estimation is not applicable to Orbitrap spectra.

7. SPECTRAL SMOOTHING

Spectral noise, particularly that higher in frequency (shorter in wavelength) than the peak width, causes problems with finite difference centroiding.[20] While baselining attempts to correct for large wavelength noise,[21] short wavelength noise is often addressed by smoothing. Given the magnitude of the noise relative to the peak height, noise can also cause false peaks to be detected by all centroiding methods, particularly when a peak model is not a perfect fit to the true peak shape. While the literature is replete with specific examples of various approaches to

[19] see section on Spectral Baselining.
[20] see section on Spectral Centroiding.
[21] see section on Spectral Baselining.

the filtering or smoothing of mass spectrometric data, as Kearnsley et al.[22] have suggested, the blind application of any of these methods to mass spectrometric data can result in significant data loss.

Savitzky−Golay Smoothing

Savitzky−Golay filters[23, 24] are probably the most popular mass spectral digital smoothing method. In this method, successive sub-sets of adjacent data points ($y_{j\pm h}$) are fitted with low degree polynomials and convolved to create a single optimal set of abundances ($\hat{Y}_j$), by least squares minimization of the Savitzky-Golay coefficients (C_h, Equation 7.1). The resulting smoothed spectrum ($\hat{Y}_j$) essentially becomes a weighted average of all neighboring abundance values in the original spectrum ($y_{j\pm h}$).

$$\hat{Y}_j = \frac{1}{N} \sum_{h=-k}^{k} C_h \, y_{j+h} \tag{7.1}$$

Examples of Savitzky-Golay filtering applied to a TOF data file is shown in Figure 7.1. In general, the lower the order the more smoothing is accomplished. A characteristic of the Savitzky-Golay filter is that the window width (i.e. the number of points) used in the polynomial fit should be just larger than the order of the polynomial for best results. Abrupt changes in spectrum abundance (before and after peaks) tend to cause discontinuities in the Savitzky-Golay smoothed spectrum, as is evident from the deviations below the spectrum for the smoothed spectra in Figure 7.1. These deviations are amplified on the log(abundance) scale used in the figure. The effect of these deviations on the resulting centroids is generally mitigated by thresholding of the resulting false negative peaks these deviations generate during the centroiding process.

However, as with all smoothing methods the raw signal is distorted in the convolution process. The peak height is reduced and the half-width of the peak is increased (Figure 7.1). Thus applying this smoothing method can force nearly-isobaric partially overlapped peaks to be irresolvably merged.

---

[22] Kearnsley, A. J., Wallace, W. E., Bernal, J., and Guttman, C. M., "A numerical method for mass spectral data analysis," *Appl. Math. Lett.,* **18**:1412-1417 (2005).

[23] Savitzky, A.; Golay, M. J. E., *"Smoothing and Differentiation of Data by Simplified Least Squares Procedures,"* Analytical Chemistry, ***36*** *(8): 1627–39 (1964).*

[24] *Steinier, J., Termonia, Y.. and Deltour, J., "Smoothing and differentiation of data by simplified least square procedure," Analytical Chemistry,* ***44*** *(11): 1906–9 (1972).*

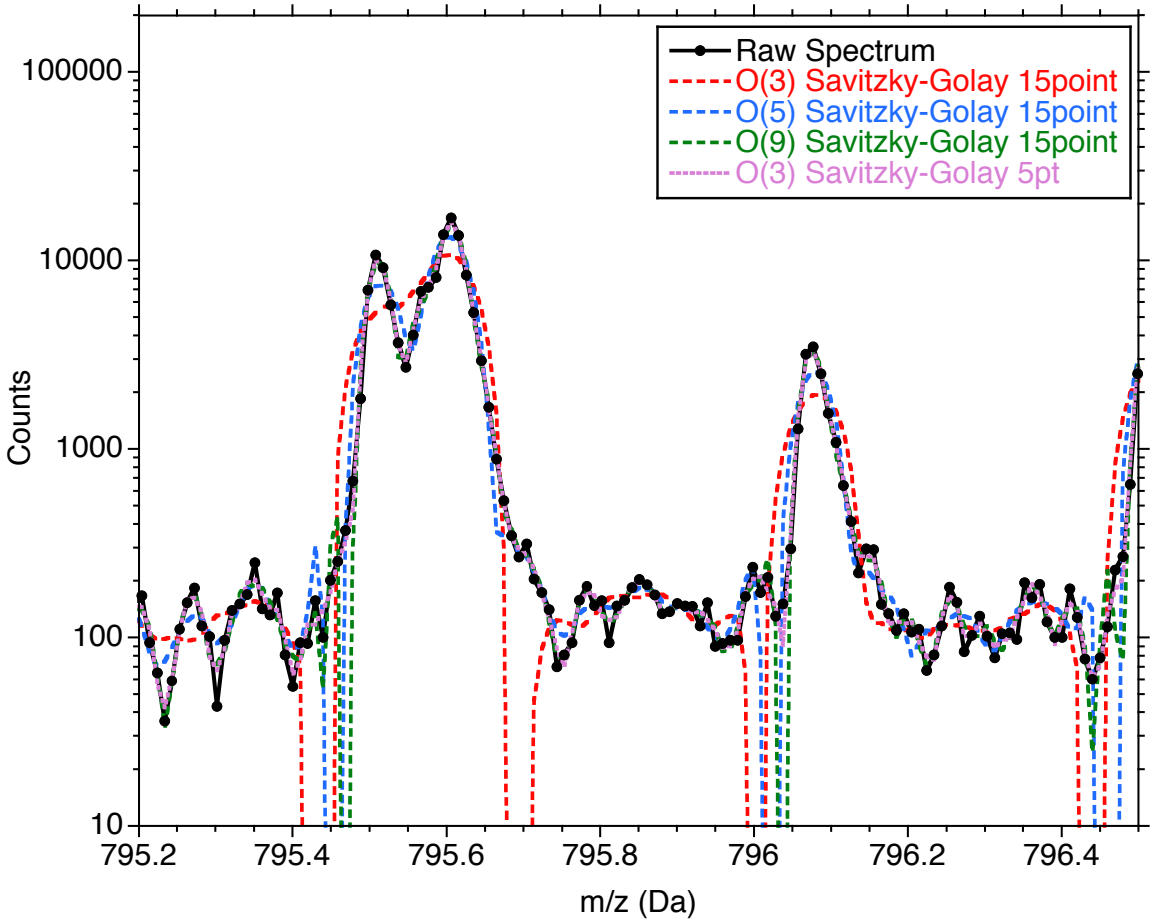


Figure 7.1. Savitzky-Golay smoothing of different polynomial orders and with different numbers of neighboring points applied to a TOF lipidomics spectrum. The discontinuities of the smoothed spectrum are characteristically seen immediately before and after a peak and deviate strongly below the spectrum. These can give rise to false positive detections, but the resulting peaks are generally below the threshold limits for reporting.

Whittaker Smoothing

Whittaker realized that smoothing is a balance between absolute fidelity of a model to the data versus simply minimizing deviations in the model curve fit to that data (Equation 7.2).[25] Designed for evenly spaced data, the Whittaker smoother attempts to fit a set of abundance data (y) with a cubic spline model (μ) by minimizing the least squares residual error versus the raw data, but penalizes the model when subsequent points within the model vary too much. With λ=0, it gives the cubic spline solution that maximizes fidelity to the raw data. As λ increases the model is smoothed out, until it eventually smooths over all the peaks as λ approaches 1.

[25] Whittaker, "On a new method of graduation." *Proceedings of the Edinburgh Mathematical Society*, **41**, 63-73, (1923).

$$SSE = (1-\lambda)\sum_i (y_i - \mu_i) + \lambda \sum_i \left(\delta^2 \mu_i\right)^2 \quad (7.2)$$

The Whittaker objective function for the sum of squares error (SSE) for minimization consists of two parts. The first part is the standard SSE from the regression model $(y_i - \mu_i)^2$, which when λ—> 0 would result in the re-creation of the spectrum as a cubic spline. The second part of the objective function consists of a local approximation (by finite difference methods) of the second derivative of the regression model $[(\delta^2\mu_i)^2 = (\mu_{i-1}-2\mu_i+\mu_{i+1})/(\Delta x)^2]$. Where the direction of points in a series is unchanged (along the trajectory of a line), the second derivative goes to zero. However, the square of the second derivative is always positive at inflection points, and the magnitude of the inflection increases the more that a series of model points deviates from linear. The user-adjustable parameter (λ) is used to weight the two terms of the Whittaker objective function to control the local responsiveness of the regression to inflections in the data (y_i). The proper value for λ can only be determined empirically (Figure 7.2).

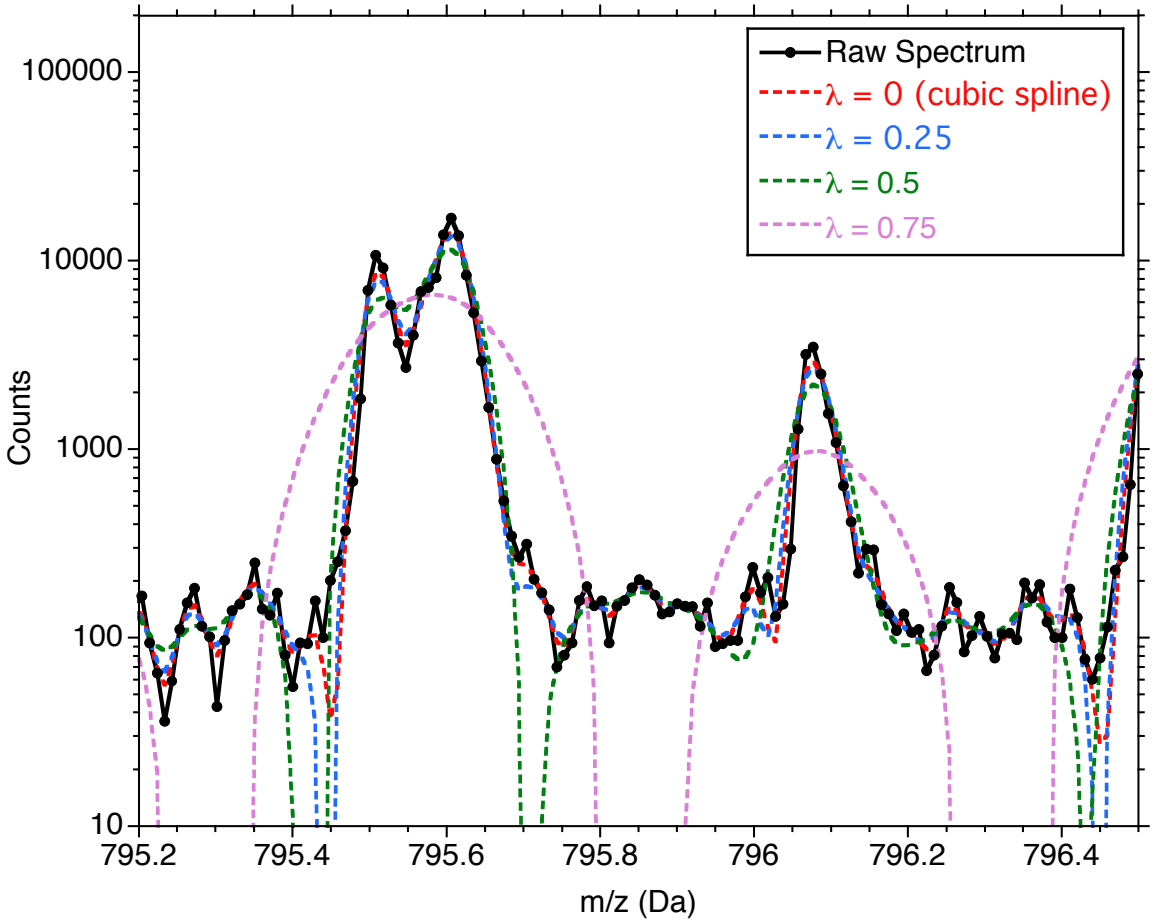


Figure 7.2. Whittaker smoothing applied to the Spectrum of Figure 7.1 with different values for λ, which varies between a minimum of 0 and a maximum of 1. Note the increased width and depth of the discontinuities at both the start and end of each peak in the smoothed spectrum, as a function of increasing λ. Values of λ>0.75 generate negative peaks in the region of interest.

Fourier Filtering

Fourier filtering is often employed to smooth data when the superimposed noise is of a different frequency than that of the signal itself. In these situations, it can be an effective means to sharpen the peaks for better discrimination by removing confounding noise that is mismatched versus the frequency of the true peaks. However, the technique can be difficult to apply generically to mass spectra for a variety of reasons.

First, the peak shape must be uniform, and must be uniformly sampled across the mass domain of the spectra, otherwise no single set of sine wave harmonics will adequately describe every peak. This is generally straightforward for ion trap spectra, but requires m/z^x transformation of TOF, Orbitrap, and FTICR spectra into alternative evenly-spaced mass domains, in order to produce uniform peak shapes and a constant waveform sampling rate in these spectra, to support Fourier filtering.

Second, the peaks of interest in the spectrum are often not evenly spaced (periodic), particularly in multi-charged spectra where the spacing between members of each isotopic series depends on the reciprocal of the net charge ($z^{-1)}$. This is further complicated by the necessary conversions of the mass domain to facilitate uniform peak shapes in TOF, Orbitrap, and FTICR spectra.

Fourier transformation starts with the assumption that any spectrum can be modeled as an infinite series of sine waves of different frequencies. The resolution power of this mathematical transform, however, depends on the number of reinforcing repeats of that peak shape pattern contained across the mass spectrum. This is illustrated in the Fourier filtering of a MALDI-TOF spectrum of Na+ PEG adducts Figure 7.3. Figure 7.3a shows the Fourier Transform of this spectrum with the various peak harmonics clearly identified. By removing all the high frequency noise (≥ 10 Da^{-1} corresponding to frequencies less than 0.1 $Da^{0.5}$) the reverse transform can be produced (Figure 7.3b). However, in this case, the resolution of the filtered Fourier reverse transform is lower than that of the original spectrum.

The full MALDI-TOF spectrum extended for nearly 5,000 Da with a PEG Na+-adduct isotopic pattern every 44 Da for a total of 113 repeating signals, yet there is still a net resolution loss by filtering. This points to the second issue with Fourier filtering, that there must be enough peak repetition on a specific frequency to be able to extract a signal, otherwise resolution is lost with the transform. In the successful literature example of Fourier filtering[26] the high frequency MALDI-matrix noise is suppressed (filtered) from a MALDI-TOF spectrum. That matrix noise generates a characteristic peak pattern within every mass unit of the spectrum, dwindling in abundance slowly over several hundred Da and having a characteristic mass spacing that was off-frequency (i.e., with a different mass defect) from that of the few analyte peaks.

[26] Kast, J. et al., "Noise filtering techniques for electrospray quadrupole time of flight mass spectra," *J. Am. Soc. Mass Spectrom.,* **14**:766-776 (2003).

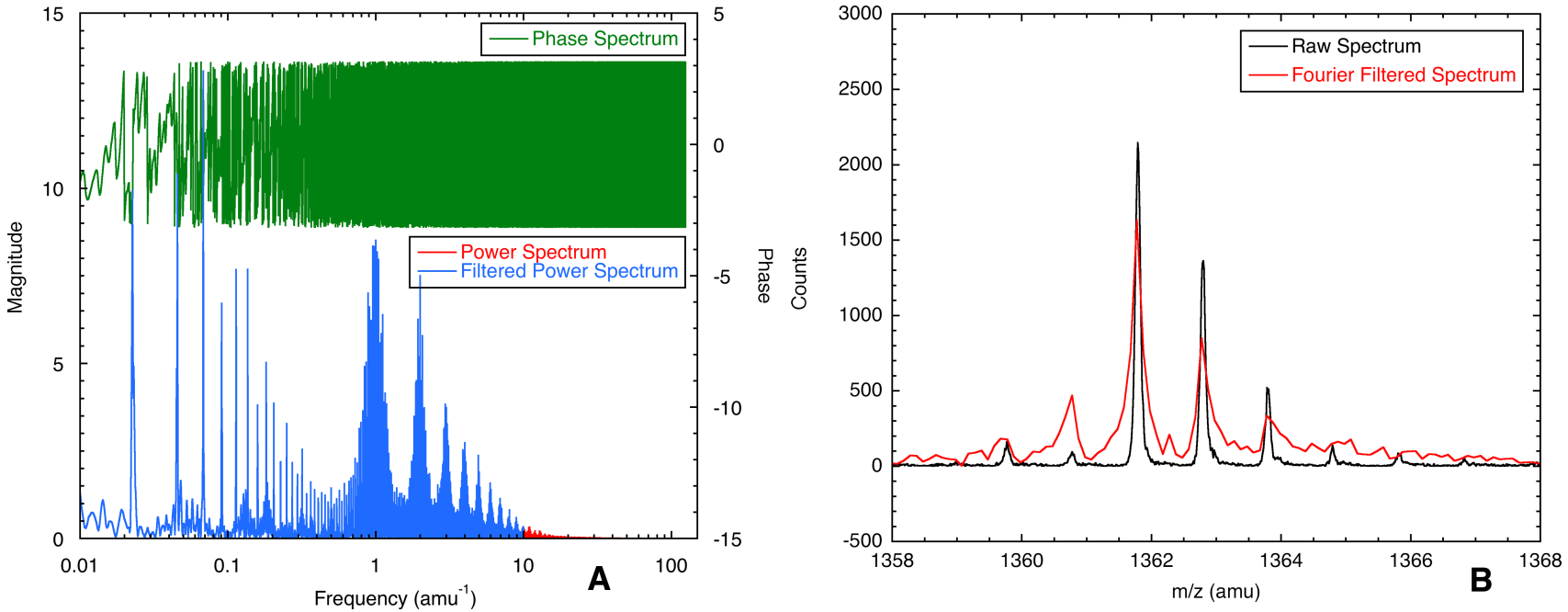


Figure 7.3. The Fourier transform (both magnitude and phase) of a MALDI-TOF poly(ethylene glycol) Na+ adduct spectrum is shown in panel A. After filtering the Fourier transform for those features less than 0.1 amu wide (i.e., frequencies >10 amu^{-1}), the inverse transform is re-registered to the original spectral data in panel B.

Wavelet Denoising

Wavelet theory solves the insufficient repetition (sampling sufficiency) issue of Fourier filtering by assuming something about the underlying waveform (i.e., by choosing a mother wavelet). The basic assumption in wavelet denoising is that the peak shape can be fully represented by a few harmonics of the mother wavelet. Ergo, no infinite series is required (as in Fourier transforms) and very few data points in localized regions of the mass spectra are needed to calculate the adjustable parameters of a relatively few harmonic terms of the mother wavelet. The art of wavelet denoising arises in guessing a mother wavelet that fully represents the peak shape, and in finding the right series of harmonics to represent the signal of interest.

Wavelet methods in mass spectrometric analysis were originally applied to ion cyclotron resonance (ICR) analyzers as an alternative to Fourier transforms.[27] However, wavelet denoising has more recently been applied to the analysis of TOF spectra. [28, 29, 30] As with Fourier filtering, the peak shape must be constant across the mass domain. Therefore, wavelets could be applied directly to ion trap spectra, but only to m/z^x transformed versions of TOF, Orbitrap or FTICR spectra.

In Figure 7.4 we illustrate the practical aspects of wavelet denoising with a simple spline wavelet denoising algorithm[31] applied to a Jeffamine polymer (an amino-terminated PEO/PPO copolymer) isotopic series from a MALDI-TOF spectrum. The first step in wavelet analysis, after choosing the form of the mother wavelet is to establish the appropriate order of the wavelet transform, specifically a high enough order that adequately reproduces the parent signal. Figure 7.4a shows that an order 5 spline wavelet is the minimum necessary to reproduce the parent spectrum. The second step (denoising) involves compression of the wavelet transform coefficient matrix by eliminating (i.e., setting to zero) the lesser coefficients of the transform matrix. Any harmonics of the mother wavelet with zero coefficients no longer contribute any signal to the inverse wavelet transform. The final step is the inverse wavelet transform. We illustrate these last two steps by mapping the inverse transforms for different noise cutoffs over the original data (Figure 7.4b).

---

[27] Shew, S. I., "Method and apparatus for determining relative ion abundances in mass spectrometry utilizing wavelet transforms," US5436447 (July 25, 1995).

[28] Morris, J. S., et al., "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics,* **21**:1764-1775 (2005).

[29] Lange, E. et al., "High-accuracy peak picking of proteomics data using wavelet techniques," *Pacific Symposium on Biocomputing,* **11**:243-254 (2006).

[30] Nafati, M. et al., "Multi-scale data reduction algorithm of proteomic mass spectrum," *Internet J. Acad. Physician Assistants,* **5**(1) (2006).

[31] Mathematica Wavelet Package. Wolfram Research.

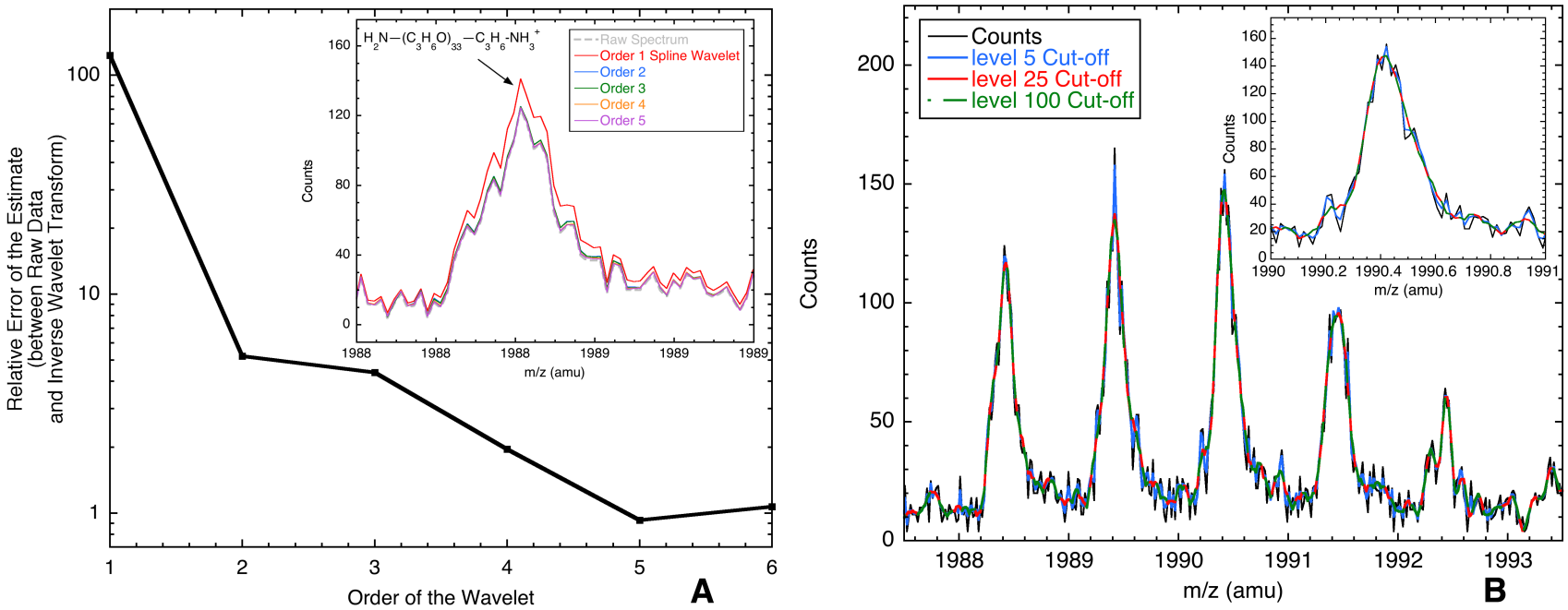


Figure 7.4. The application of spline wavelet denoising to Jeffamine D-2000 isotope series from a MALDI-TOF spectrum. Panel A shows how increasing the order of the wavelet transform improves the fit of inverse transform to the original data with an order 5 spline providing the best fit. The inset shows the actual superimposition of the inverse wavelet transforms on the original spectrum of the monoisotopic peak shown. Panel B shows how the signal can be denoised by progressively compressing lower value coefficients of the wavelet transform matrix before taking the inverse transform as applied to the isotopic series (the inset shows one of the [^{13}C]-isotopic peaks. In panel B the coefficients below the cut-off values shown were set to zero before taking the inverse transform.

While this technique shows promise for removing the high frequency noise, the user must define a suitable mother wavelet and determine which components of the transform matrix to eliminate, with nothing to guide them but trial and error. Unfortunately, the spline mother wavelet used in this example failed in other examples containing multiply charged higher molecular weight polymers and proteins in an ESI spectrum on the same mass analyzer. There are also rigorous constraints on functions that can be used as mother wavelets (i.e., they must be infinitely differentiable). Furthermore, every time the mass range of the analyzer or any of its tuning parameters are changed, these user-adjustable parameters must be re-optimized.

The Downside of Smoothing

All of these smoothing techniques are designed to remove the noise that exists at higher frequencies (lower wavelengths) than that of the analyte peak widths. However, all of these methods are data destructive, eliminating metadata about the spectrum that can be useful for downstream spectral analysis or peak quality measurements. Finally, any medium wavelength (lower frequency noise) not eliminated by baselining or below the detection threshold will still be present in the spectrum, confounding the centroiding results (Figure 7.5).

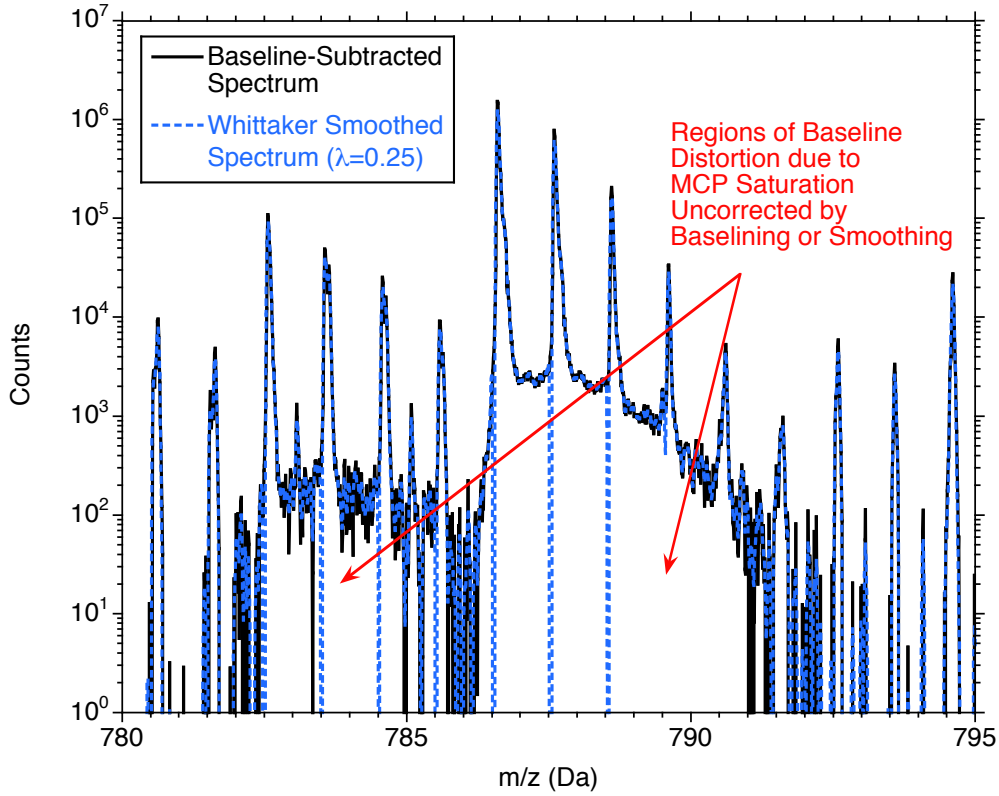


Figure 7.5. Residual medium frequency noise (due to detector saturation) in the Whittaker-smoothed TOF spectrum (Figure 7.2). Baseline subtraction reduces the lowest frequency noise. The baseline-subtracted spectrum was smoothed with the Whittaker smoother (λ=0.25) to suppress the high frequency noise. However, the medium frequency noise (that between the frequency of real peaks and the low frequency noise of baseline shifts or "float" due to ion saturation of the MCP detector) is still evident in the regions indicated.

8. THRESHOLDING

Thresholding is a procedure to discriminate spectral information that contains sufficiently strong and differentiated signal levels to be accepted as containing reliable peak information, versus the remaining spectral information that cannot be adequately differentiated from noise. The threshold is always based on abundance, but the threshold does not have to be constant with m/z and may vary locally (if the software supports that). The difference between a threshold and a baseline is that a baseline attempts to estimate the average level of noise superimposed on the spectrum, whereas a threshold attempts to estimate the interface between the noise and the smallest reliably detectable peak.

Some of the same methods used for spectral baselining can also be applied to the problem of thresholding with slight changes in parameters[32]. For example, the Asymmetric Whitaker smoother can be weighted to the data above the cubic spline rather than below. Yet all the same problems of estimating proper user-adjustable parameters remain, but in the thresholding application are made more complex because there is no upper bound to limit the result (i.e. the peaks rise in the same direction as the noise variance, whereas with baselining the noise variance is determined in the opposite direction of the peaks). For example, a least squares regression may bisect the spectral noise or rise above the smaller isotopic peaks depending on the relative peak abundance compared to background noise, and iterative polynomial fits with subtraction of the points below the regression will ultimately move the threshold to the apexes of the highest n peaks, where n is one more than the order of the polynomial regression.

Using the Signal-to-Noise Estimate for Thresholding

A Synthetic Example

It is possible, however, to use the estimated maximum signal-to-noise level (determined by subtraction of the median baseline and reflection of the negative residuals)[33] as a threshold for peak detection. We illustrate this by the following example. In this example a synthetic TOF spectrum was constructed containing the full isotopic patterns of known composition peaks with monoisotopic peaks of 500 counts (Figure 8.1). Noise was superimposed over this spectrum at an average of 15 counts (varying randomly between 0 and 30 counts).

[32] Spectral Baselining.docx
[33] Spectral Signal-to-Noise Determination.docx

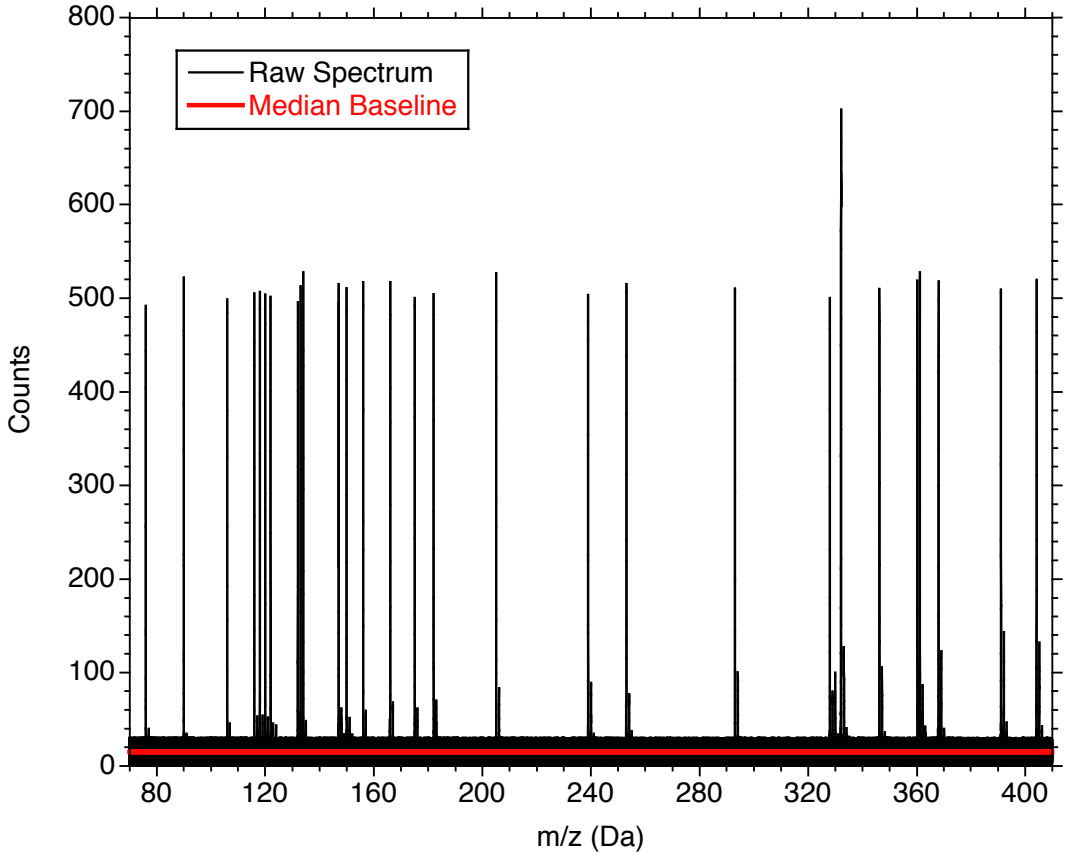


Figure 8.1. A synthetic TOF spectrum produced from 31 known chemical species and their predicted isotopic abundances. The monoisotopic peak of each species was set to 500 counts. A 15 count average noise was superimposed over the spectrum, varying between 0 to 30 counts randomly by mass bin. The 15 Count median baseline is also shown.

Following the procedure outlined previously for estimating the spectral signal-to-noise variance,[34] the median baseline was subtracted and the noise variance below zero counts was reflected into the positive count domain. The distribution of noise variance was then modeled as a normal distribution with 1 sigma = 8.4045 Counts.

The baseline-subtracted spectrum was then centroided using mMass v5.3.0[35] with constant absolute abundance thresholds proportional to different multiples of the modeled noise variance (n times sigma). Since all peaks present in this spectrum are known with certainty, it is possible to determine the Positive Predictive Value (PPV, Equation 8.1), the False Negative Rate (FNR, Equation 8.2), and the Fraction of Observable Peaks (Equation 8.3) that were detected at each threshold level (Figure 8.2). Observable peaks are defined as those known peaks with theoretical abundances above the threshold level.

In all cases the allowable mass error of the centroid was ± 2.5 times the intrinsic mass spacing of the spectrum (±0.0005 $Da^{0.5}$). Duplicate detection events (i.e., where a

[34] Spectral Signal-to-Noise Estimation.docx

[35] M. Strohalm, Kavan, D., Novák, P., Volný, M., Havlíček, V., "mMass-Open Source Mass Spectrometry Tool v5.3.0," *Anal Chem,* **82**:4648-4651 (2010).

secondary noise peak was superimposed near the apex of the known peak) were eliminated with the secondary centroid classified as a False Positive detection event. False Negative detection events were established based on the failure to detect any known peak whose theoretical abundance was above the threshold. No distinction was made between noise being added to the abundance of a peak and adventitious detection of pure noise within the mass tolerance of a theoretical peak but not riding on the sides or apex of the theoretical peak.

$$PPV = \frac{True\,Positives}{(True\,Positives + False\,Positives)} \tag{8.1}$$

$$FNR = \frac{False\,Negatives}{(True\,Positives + False\,Negatives)} \tag{8.2}$$

$$Fraction\,of\,Observable\,Peaks = \frac{True\,Positives + False\,Negatives}{Total\,Number\,of\,Observable\,Theoretical\,Peaks} \tag{8.3}$$

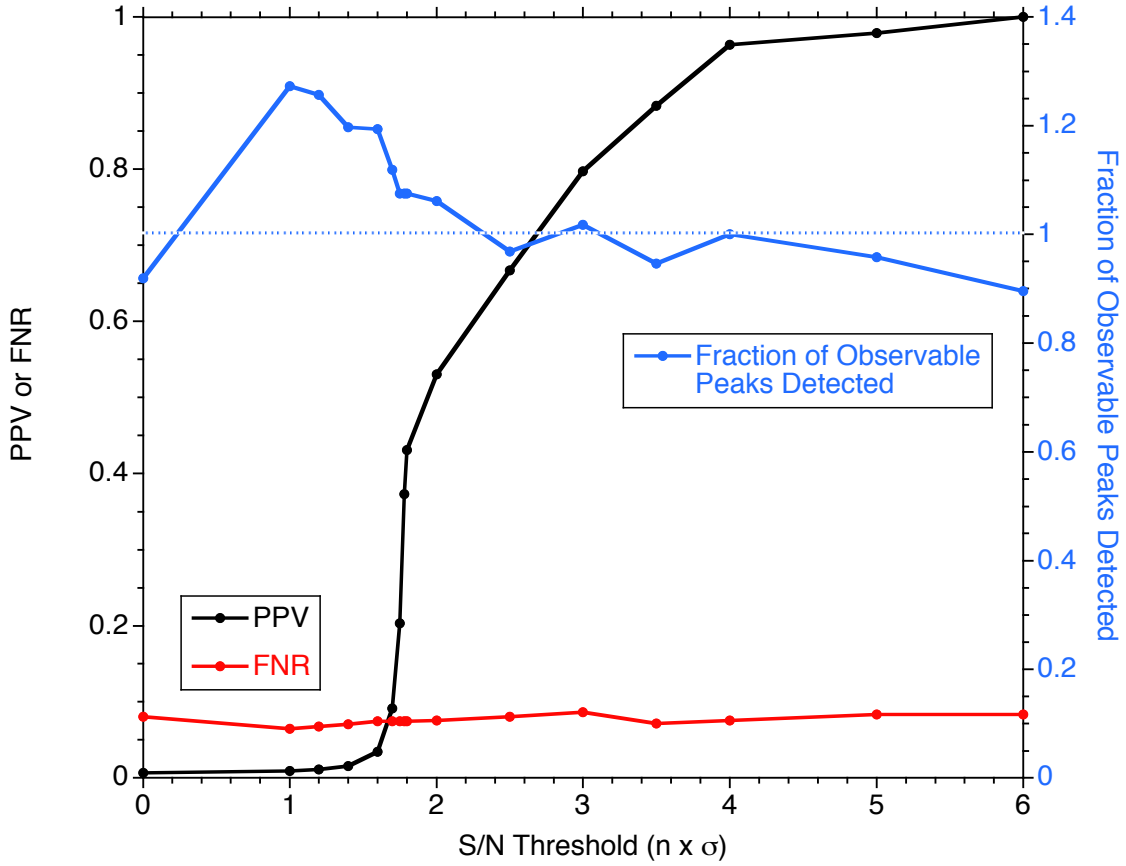


Figure 8.2. The Positive Predictive Value (PPV) and False Negative Rate (FNR) determined as a function of threshold value. The threshold value is cited in multiples of the estimated noise variance (σ). One σ implies 84.13% of the estimated noise is below the detection threshold. At 2 and 3 σ, respectively 97.72% and 99.87% of the noise is below the threshold. The Fraction of Observable Peaks (i.e., peaks detected versus the known synthetic peaks with abundances above the detection threshold) detected at each threshold is also shown.

High numbers of false (noise) peaks are detected until the threshold gets high enough to clear the inter-peak noise just before the 2σ threshold. These False Positives adversely affect the PPV in this region as well. These noise peaks sometimes align within the mass tolerance of observable peaks that would normally go undetected since they were below the detection threshold. When this happens they are counted as True Positives, even though they may really be noise. These residual noise peaks account for the nearly 20% over-detection of theoretical peaks in this threshold region. While perfect PPV is asymptotically approached at thresholds between 4 and 6σ, the false negative detection rate is nearly constant (around 10%) at all threshold values. This is because each peak contains some superimposed noise on its sides and apex. This noise can shift the apparent peak outside the mass tolerance window of the theoretical observable peak, resulting in a False Positive. Alternatively, if the noise variance is below the median baseline, it can lower the measured apex of smaller isotopic peaks below the threshold of detection.

The synthetic spectrum shown above provides a useful demonstration of the basic signal-to-noise thresholding technique. It shows that by statistically eliminating increasing amounts of the spectral noise from between 97.7% (2σ) to effectively 100% (4σ) that the relative proportion of true peak identifications improves dramatically, with little loss of real information content. However, it is a contrived example where all details of the problem are known *a priori* and accounted for in the solution.

Application to a Real Spectrum

In the following example the same principles are applied to a real TOF spectrum. This scan (Figure 8.3) was selected randomly from an LC/MS plasma lipidomics run. The PeakInvestigator™ approximation of the median baseline was applied (as shown) and the spectral noise estimated by reflecting the noise variance calculated below this baseline above the baseline.

A more detailed analysis of the spectrum in different mass regions (Figures 8.4a, b, and c) suggests that the bulk of the spectral noise will fall below a 2-4σ S/N threshold. However, there are some regions of localized baseline variations that are not adequately modeled by the baseline and corresponding s/n threshold near 122 Da (Figure 8.4a), 788 and 812 Da (Figure 8.4b). These shorter wavelength localized baseline variations are commonly associated with detector saturation events in TOF and ion trap spectra. The wavelengths being only slightly longer than that of the peaks, is ignored by the PeakInvestigator™ baselining algorithm, which focuses on the longest wavelength baseline variations. Since they are above the baseline they also would not be reflected in any S/N threshold.

By subtracting the baseline, the S/N threshold becomes constant across the mass range (Figure 8.3), and the baseline-subtracted spectrum can then be centroided by mMass since that algorithm only accepts flat threshold values. It should be noted, however, that the s/n threshold may be variable with m/z in some spectra, but variable thresholds are not accommodated in most centroiding software programs. In this case after baseline subtraction a constant threshold can be assumed and mMass centroiding was performed on the baseline-subtracted scan with progressive threshold levels corresponding to multiples of the standard deviation of the noise variance. This process was repeated for each of the scans immediately preceding and following this center scan in the LC/MS series. Any differences between the actual median noise level and associated noise variance between the three scans are thus eliminated and the nσ thresholds of each can be directly compared, even if the absolute values of the thresholds may be different.

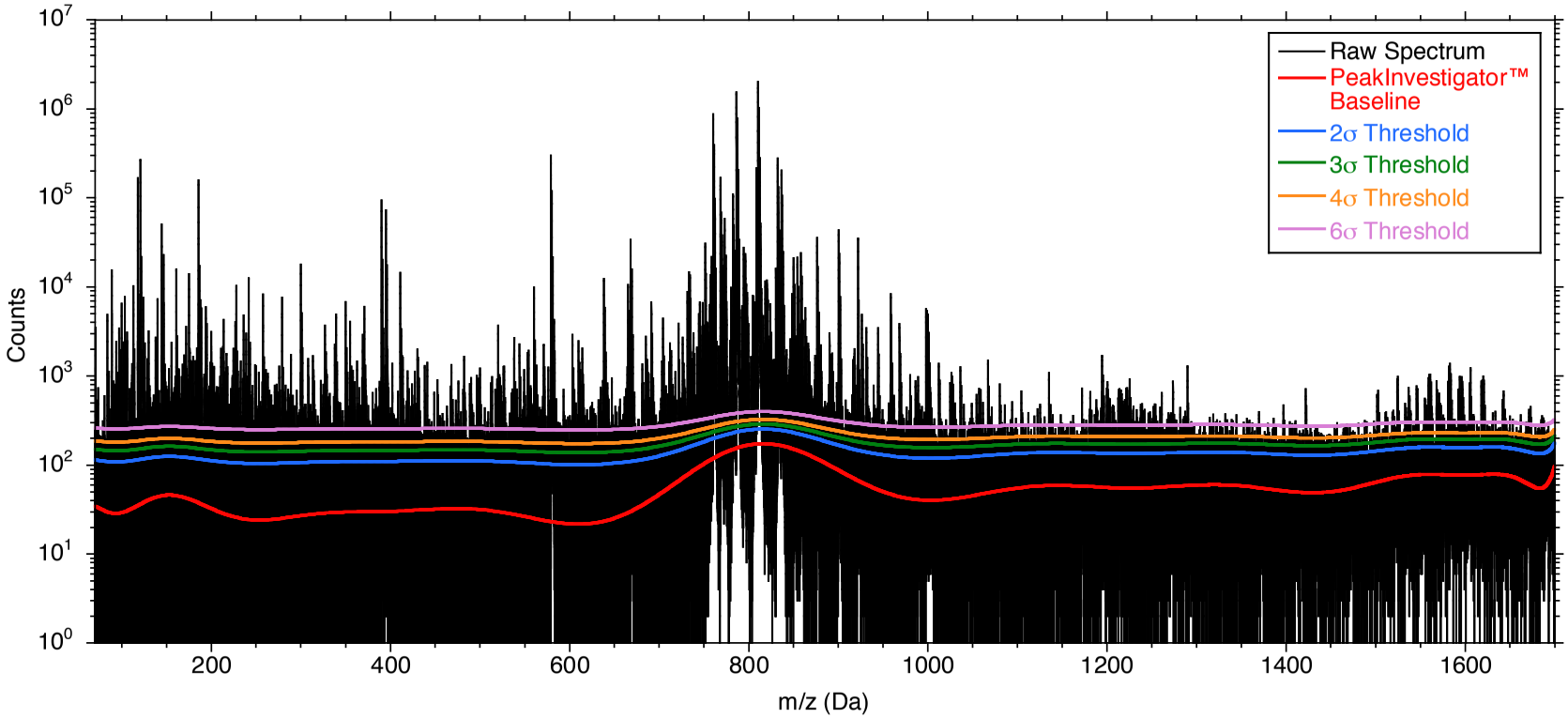
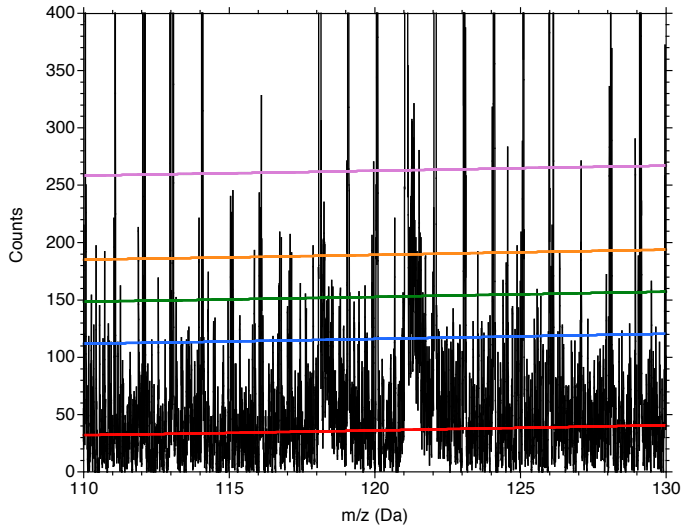
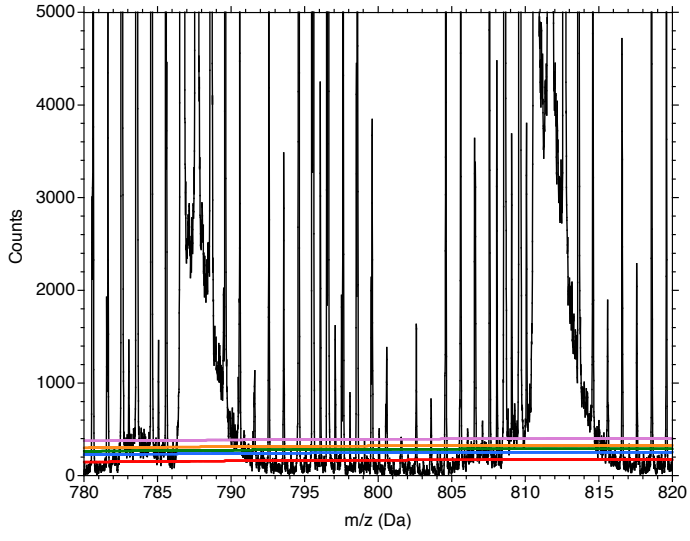


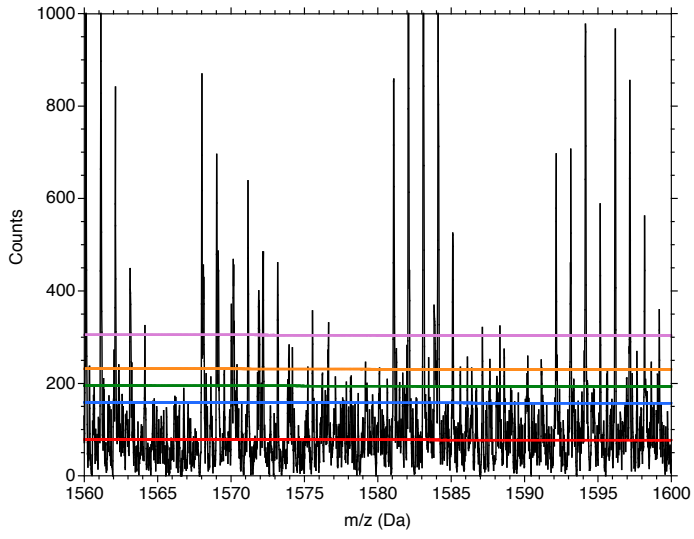
Figure 8.3. Signal-to-noise thresholding for a random LC/MS scan taken from a plasma lipidomics experiment. The PeakInvestigator™ approximation of the median baseline is shown, to which is added the reflected spectral noise at several different multiples of the standard deviation in its variance. The data are plotted on a log(abundance) scale for better visualization.



(a)



(b)



(c)

Figure 8.4. Close-up regions at the low, middle, and high mass ranges of the LC/MS scan shown in its entirety in Figure 8.3. These figures suggest that a signal-to-noise threshold between 2 and 4 times the overall noise variance would provide a dynamic threshold above the baseline noise over much of the spectrum. However in the area of the largest peaks (700 to 900 Da), the local baseline appears elevated by detector saturation in these regions. This localized baseline variation is of too small a wavelength to be adequately modeled by the PeakInvestigator™ baseline and so is not reflected in the corresponding S/N threshold. The same situation may be seen to a lesser extent near 122 Da in Figure 8.4a.

Peak Identities

Since the identities and abundances of the species present in these scans are unknown, it is necessary to define which centroids will be accepted as real (positive) peaks and which will be classified as noise (i.e., negative) peaks. Within a mass tolerance of ± 3 times the intrinsic mass spacing (± 0.0005 $Da^{0.5}$), all the centroided peaks found within the three consecutive scans were identified as positive or negative detection events based on the following criteria:

- A positive peak results from:
  - a) any peak seen in the central scan of the series that is also seen in at least one of the adjacent scans at greater than 50% relative abundance (a true positive), and
  - b) any peak that appears in both adjacent scans but fails to appear in the central scan at greater than 50% relative abundance seen in the adjacent scans (a false negative).
- A negative peak results from:
  - a) any detected peak in the central scan that can not be confirmed in either adjacent scan at greater than 50% relative abundance (a false positive), and
  - b) any peak seen in either adjacent scan that is not confirmed in the central scan or the other adjacent scan at greater than 50% relative abundance (a true negative).

The resulting distributions of positive and negative peaks plotted by their centroided abundances is shown in Figure 8.5. As we have seen previously,[36] the abundance distributions of positive and negative peaks overlap considerably. The goal of any thresholding algorithm is to discriminate between these positive and negative detection events, retaining as many of the positive peaks, while discriminating against as many of the negative peaks as possible.

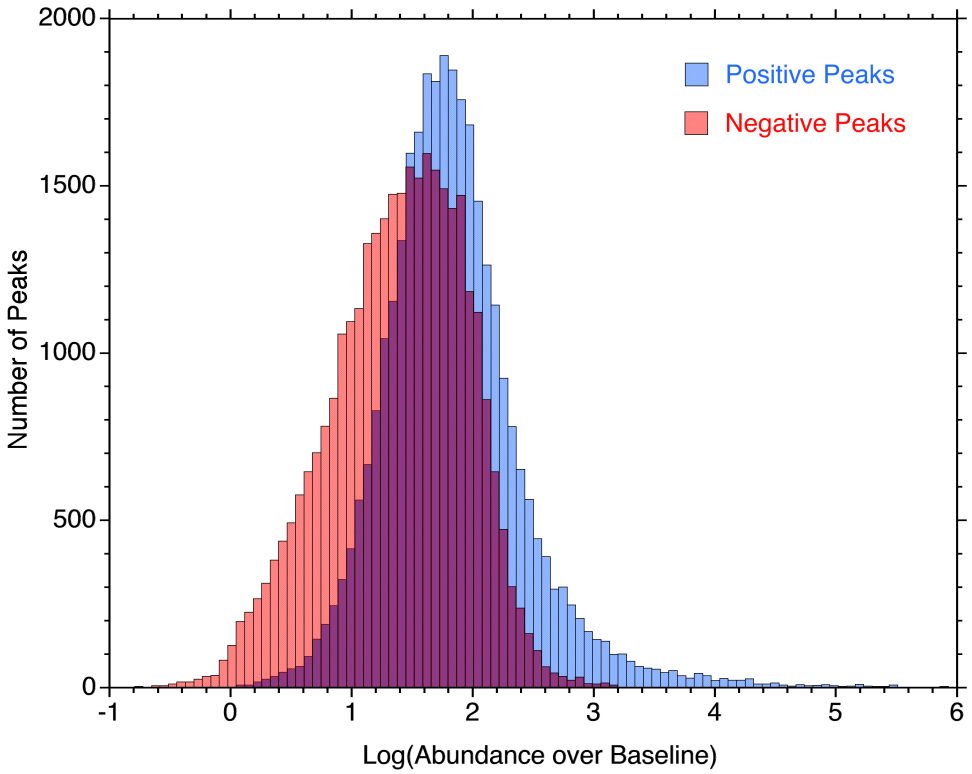


Figure 8.5. The abundance distribution of positive and negative peaks detected by mMass centroiding of the PeakInvestigator™ baseline subtracted scan by the criteria identified in the text.

Receiver Operating Characteristic (ROC) Analysis

Receiver Operating Characteristic (ROC) analysis can be applied to the problem of how well a threshold discriminates between two distributions.[37] Depending on the threshold value, the abundance of any positive peak may be above the threshold and is counted as a true positive (TP), or may lie below the threshold value and is counted as a false negative (FN). Similarly, the abundance of any negative peak that is above the threshold value is counted as a false positive (FP) or is counted as a true negative (TN) when below the threshold value. Sensitivity and specificity values are determined for each threshold value from the peak counts in these four categories using equations 8.4 and 8.5. Note, as the threshold is raised, more positive peaks will move from TP to FN and negative peaks from FP to TN, suggesting that sensitivity will drop and specificity will increase as the threshold rises.

$$Sensitivity = \frac{TP}{TP + FN} \tag{8.4}$$

[36] Spectral Baselining.docx

[37] Reciever operating characteristic, https://en.wikipedia.org/wiki/Receiver_operating_characteristic (accessed 21Jul2016).

$$Specificity = \frac{TN}{FP + TN} \tag{8.5}$$

By plotting the sensitivity against 1-specificity determined for each threshold value, an ROC curve is constructed (Figure 8.6). If no discrimination is seen between the two peak distributions the ROC curve would lie on the diagonal between 0,0 and 1,1. The more discrimination that the threshold provides between members of the two distributions, the closer the ROC curve will approach the upper left corner (1,0) of the graph. The area between the actual ROC curve and the diagonal is called the area under the curve (AUC) and is a measure of the total difference between the two distributions. The AUC has a maximum value of 0.5 and a minimum of zero. The point of maximum ROC curve deviation from the diagonal is called the Youden index and corresponds to that threshold providing the greatest discrimination between the two distributions.

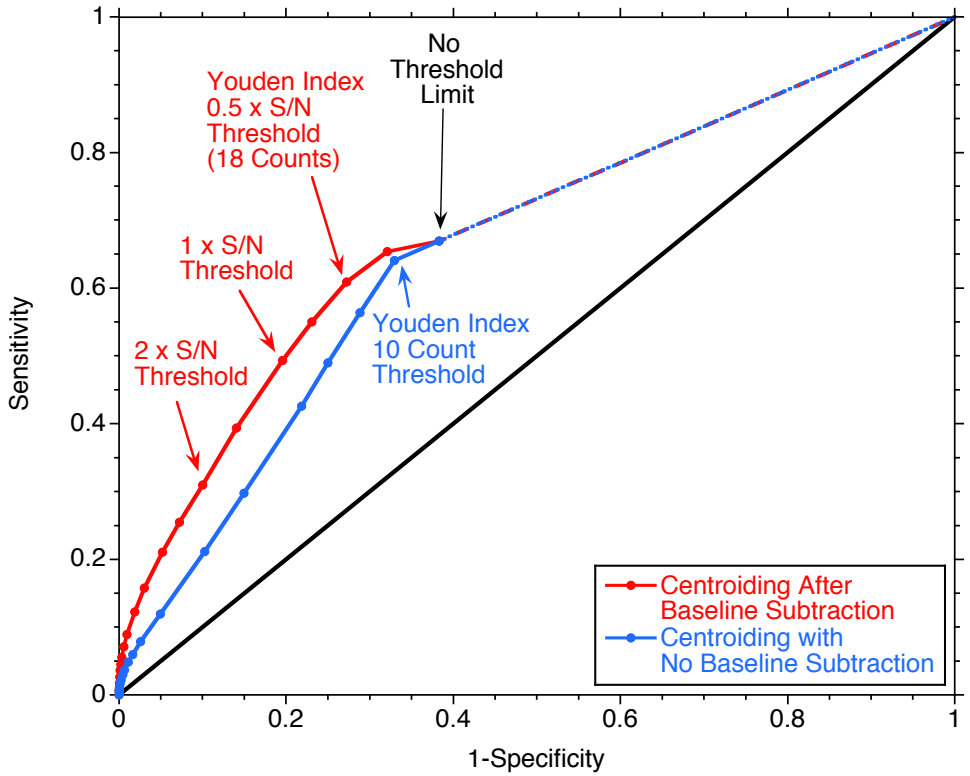


Figure 8.6. The Receiver Operator Characteristic (ROC) curves determined for the LC/MS scan of Figure 8.3 both with and without baseline subtraction. In both of these curves a flat threshold was applied over the entire spectrum during centroiding to generate the ROC curve. The area under the curve (AUC) shows that only 37.4% of the positive and negative peaks can be discriminated based on abundance alone, after baseline subtraction, and 31.7% without prior baseline subtraction. Youdon Indicies (optimum thresholds) before and after baseline subtraction were 10 and 18 counts, respectively.

The AUC (0.187) from figure 8.6 suggests that only 37.4% of the positive and negative peaks (Figure 8.5) can be discriminated using the signal-to-noise threshold after baseline subtraction. A similar ROC analysis of the raw spectrum (centroiding without baseline correction) produces an AUC of 0.159, suggesting only 31.7% of the positive and negative peaks could be discriminated without any baseline adjustment. Therefore, baseline correction

prior to centroiding yields nearly 18% better signal to noise discrimination. The position of the Youden index is at 0.5σ, which corresponds to the optimum s/n threshold level for this spectrum. The obvious remaining question is why does a flat threshold (either s/n or absolute counts) produce such poor peak discrimination results even after baseline correction?

Remaining Detection Issues

Mid-Frequency Baseline Distortions

By subtracting the PeakInvestigator™ dynamic baseline the longest wavelength (lowest frequency) variation in the baseline is effectively removed from the spectrum providing more consistent centroiding results for a flat threshold. The s/n analysis effectively accounts for the shortest wavelength noise variation in the spectrum. What is left after applying each of these corrections is the medium wavelength noise (i.e., that which approaches the inherent peak width in wavelength).

Some of this medium wavelength noise is seen in the baseline distortions around the larger peaks in the spectrum (particularly Figure 8.4b). It is readily seen that the mMass centroiding method over-estimates the abundance of those peaks riding on top of the medium-wavelength baseline distortion in the vicinity of larger peaks (Figure 8.7). This overestimation of peak abundance artificially moves these centroids above the threshold s/n value, causing the retention of FP detection events. When the threshold is raised, TP peaks in regions of the spectrum unaffected by detector saturation are then lost as FN events. Note that the abundance distortion in peak intensity in this example exceeds two orders-of-magnitude.

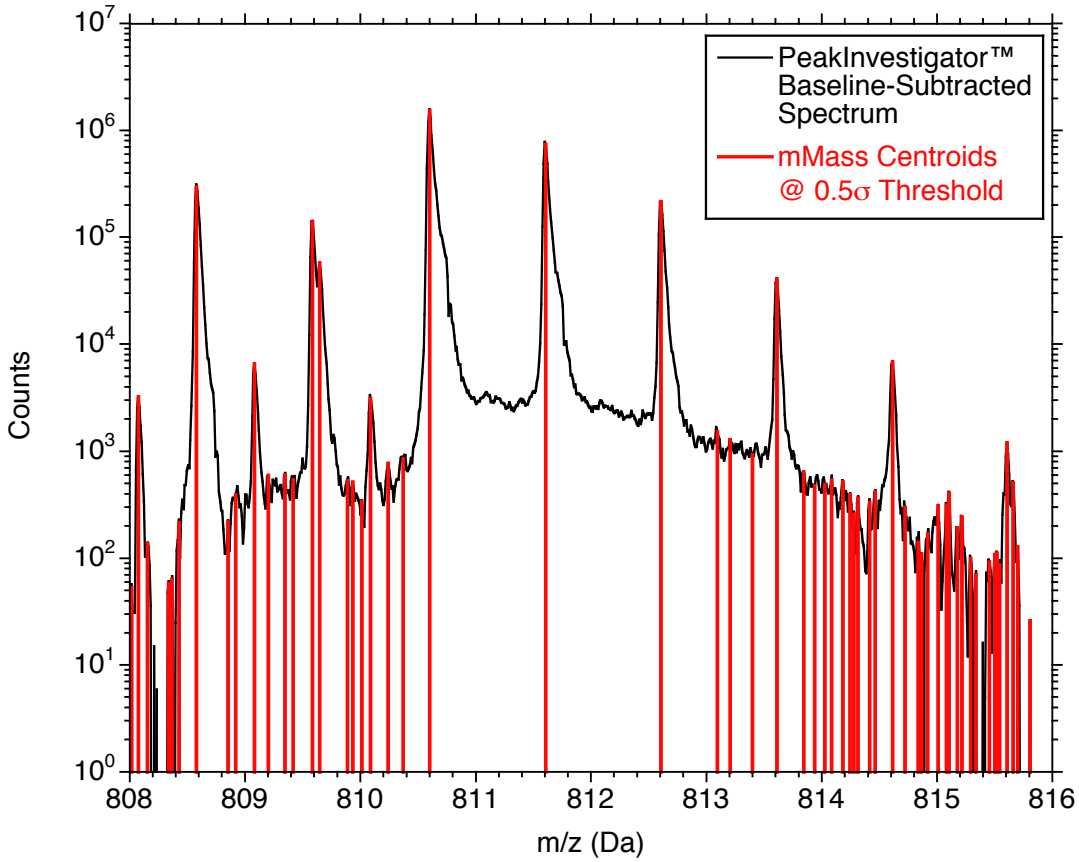


Figure 8.7. Centroids (for the optimum 0.5σ s/n threshold) determined for the baseline-subtracted spectrum of Figure 8.3 for peaks near the most abundant peaks (and the regions of baseline distortion these create). The mMass centroiding method is drawing its centroid abundance from the baseline (zero) to the peak apex. Because of the localized baseline distortion around the abundant isotopic series between 810.5 and 815 Da, what would normally be considered interpeak noise is now larger than the threshold and kept in the peak list. The spectrum is shown on a log(abundance) scale to better illustrate the localized baseline distortion of over 1000 counts from the baseline.

Superimposed Noise on Peaks

Another source of peak detection errors is caused by the superimposition of noise on top of the peaks themselves. The mMass Peak Picking is a standard local-maxima based centroiding algorithm based on the first and second derivatives of abundance with m/z. Any putative peak is determined from the local derivative and its abundance (which must exceed the flat threshold abundance criterion provided by the user). The centroid height (peak abundance) is drawn from the baseline (zero) through the center of mass of the peak to the interpolated line of the spectrum above it. So any noise on the side of a peak that causes a detectable derivative, carries with it the abundance of the underlying peak raising it above the threshold. An example is shown in Figure 8.8 (at 136.85 Da) where the noise peak sits on the side of main peak (at 136.79 Da). The Centroid (base to apex is greater than the 1σ threshold applied, but the peak height from the extrapolated side of the main peak would be less than that 1σ threshold.

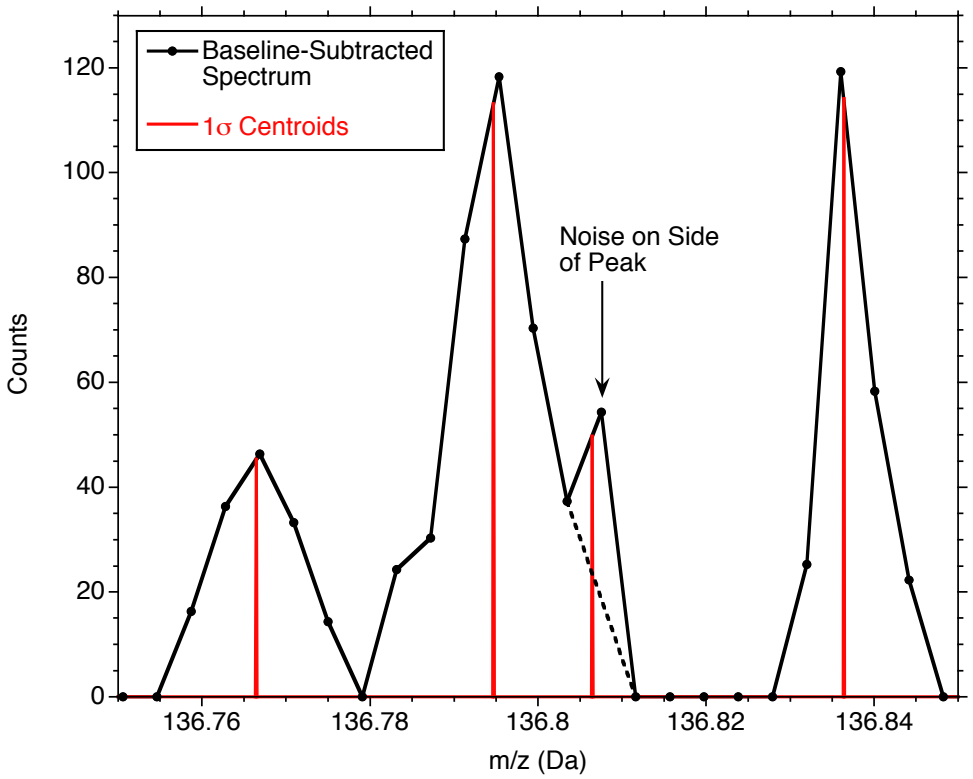


Figure 8.8. Centroids determined for the baseline-subtracted spectrum of Figure 8.3 at a 1σ S/N flat threshold. The peak detected at 136.85 Da is a False Positive created by noise on the side of the peak that is higher than the 1σ S/N

Threshold. The centroid abundance of this peak includes part of the neighboring peak upon whose shoulder (dashed line) it rides.

The two remaining detection issues, then, are: 1) adapting the centroiding algorithm to deal with local variations in the spectral baseline of wavelengths approaching that of the peak width and 2) creating a centroiding algorithm able to detect and quantitatively deconvolve peaks overlapping the sides of other peaks to generate a more appropriate centroid abundance value for each deconvolved peak.

9. SPECTRAL PEAK CENTROIDING

At the heart of every mass spectral analysis is the detection, location and quantification of the real analyte peaks in the mass spectrum. These activities are all accomplished as part of "centroiding," which converts the mass spectrum as precisely as possible into a list of peak masses and abundances (a "mass list" or "peak list"). By the classic definition, centroiding is the process of determining the center of mass of the detected analyte molecules that are dispersed into separate detector bins surrounding the true m/z of the analyte molecule (i.e., the "centroid" of the peak). However, peak centroiding actually consists of three distinct mathematical steps: 1) detecting the presence of a peak, 2) determining the center of mass of that peak, and 3) quantifying the abundance of that peak.

Peak Detection

Finite Difference Calculus

Basic calculus tells us that a peak in a continuous distribution is characterized by the first derivative of that function passing through zero and the second derivative being negative at that point. A mass spectrum can be thought of as a periodically sampled continuous distribution. Therefore, finite difference calculus can be used to determine the local first and second derivative around any mass point in that spectrum.

Assuming the peaks are roughly uniform in shape, the central difference equations provide the greatest accuracy. In the simplest case, the first derivative of abundance (Equation 9.1) and second derivative of abundance (Equation 9.2) can be determined at any mass position (m/z_i) from the following equations:

$$\left(\frac{d\,Abundance}{d\,(m/z)}\right)_i = \frac{Abundance\,(m/z_{i+1}) - Abundance\,(m/z_{i-1})}{m/z_{i+1} - m/z_{i-1}} \tag{9.1}$$

$$\left(\frac{d^2\,Abundance}{d\,(m/z)^2}\right)_i = \frac{Abundance\,(m/z_{i+1}) - 2\,Abundance\,(m/z_i) + Abundance\,(m/z_{i-1})}{\left(\frac{m/z_{i+1} - m/z_{i-1}}{2}\right)^2} \tag{9.2}$$

The quality of the finite difference approximation to the actual first and second derivatives depends on two assumptions. The first is that the m/z data is evenly spaced; hence, the need for accurate spectral decompression.[38] While m/z data is only truly evenly spaced in ion trap spectra, the variation in the m/z spacing around any given point is generally small enough to be ignored in the above calculation for all other analyzer types. The second assumption is that noise is small relative to the change in signal from point to point in the spectrum. Where the peaks are large, this second assumption is reasonably valid, but as the peak abundance starts to approach that of the spectral noise, single points of random noise can be easily mistaken for real peaks. Conversely, nearly isobaric partially overlapped peaks can be ignored if their abundance is low relative to that of their nearly isobaric neighbor.

[38] Spectral Data Compression and Decompression.docx

We can see the effect of noise on finite difference peak detection in Figure 9.1. Here we have a synthetic TOF spectrum of an isolated 500-count peak with superimposed random noise of between 0 and 30 counts. The first and second derivatives for each of the members of this series are shown as are the resulting centroids. Noise superimposed on the tailing edge of the main peak is detected as additional peaks above the threshold value.

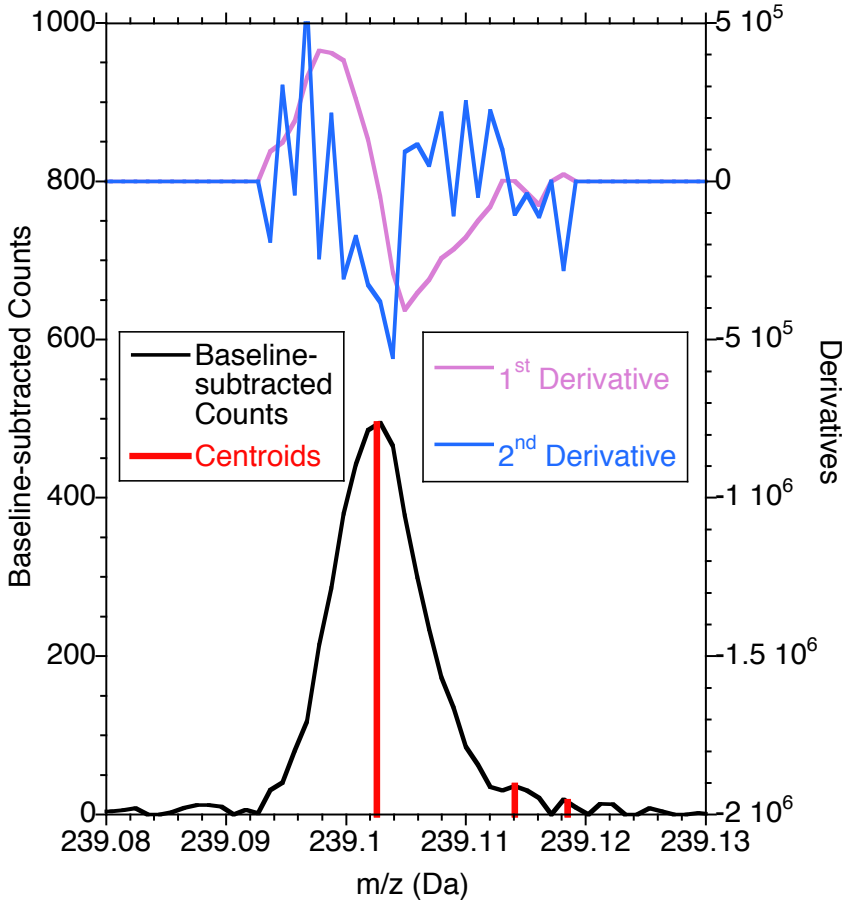


Figure 9.1. Finite difference peak detection applied to a synthetic baseline-subtracted TOF spectrum of a singly-charged 500-count monoisotopic peak with superimposed random noise of between 0 and 30 counts. The centroiding threshold was set at 1σ (15 counts over baseline). The resulting centroids are shown.

The obvious and traditional solution to this inherent problem of false positive detection using finite difference centroiding is to smooth the spectrum. We have independently presented arguments against spectral smoothing as applied to baselining in that all such techniques are: 1) intrinsically data destructive and 2) require the specification of at least one user-adjustable parameter.[39] However, they present the only solution to the false positive peak detection problem inherent in the finite difference method for peak detection. Smoothing techniques and their limitations are discussed separately.[40]

Apex Finding

[39] Spectral Baselining.docx
[40] Spectral Smoothing.docx

A second approach to peak finding is to start at the most abundant apex in the spectrum and work downward until the threshold is reached. However, a mass spectrometric peak is wider than a single mass point. Hence, the mass points leading up to any apex may also be higher than the apexes of other peaks in the spectrum. Therefore, this peak detection method necessarily must be accompanied by some concept of the general width or shape of a peak, to preclude multiple imperfect detections of the same peak in the mass list by blocking the selection of another apex from some mass range around each apex already called.

Alternatively, the highest abundance actual mass point next to a peak determined by finite difference methods can be chosen as the mass and abundance apex. Since the true peak may fall between two mass points, the best mass resolution for apex picking is the ± distance between the neighboring m/z points in the spectrum. Where nearly isobaric peaks overlap, a peak model or width guide used to block multiple detections of the same peak may also block the detection of a partially overlapped neighboring peak.

Accurate Peak Mass Determination

Center of Mass (Traditional Centroiding)

The center of any object can be defined mathematically by least squares fit of the distances from each sampled position on the surface of that object measured to a common center point (Equation 9.3). This is illustrated in Figure 9.2. One user-specified variable in this calculation is which sampled points should be used in the objective function. It is common to use only those points within some abundance of the apex since the larger the abundance associated with any mass point, the less the effect spectral noise should have on this calculation. In the absence of spectral noise with a perfectly symmetrical peak (e.g., a Gaussian) the center of mass would be the same no matter which points are used for its calculation. The more asymmetrical the peak shape, however, the more the centroiding abundance cutoff will affect the precision of the center of mass calculation (Figure 9.2).

$$min\left[\sum_i r\right] = \sum_i \left[(m/z_i - m/z_{center})^2 + (Abundance_i - Abundance_{center})^2\right] \quad (9.3)$$

Many variations on this basic approach are possible. For example if only 4 points are used, the resulting system of simultaneous equations and be solved explicitly.[41] Alternatively, abundance weighted moments of the points can be used to give more weight to the higher abundance values.[42]

The challenge in this approach is defining what is part of a peak and what is not. The presence of any overlapped nearly isobaric peaks will skew the center of mass calculation with this method unless there is a way to determine where the second peak starts (e.g., finding a trough between two peaks).

Even with a known isolated peak, as shown in Figure 9.2, even small peak asymmetries can cause variations in the center mass, depending on which points are used in the centroiding objective function. Therefore, it is typically important to re-calibrate the masses of the centroids after centroiding and to use a consistent cutoff (% of apex height) to centroid all peaks in a spectrum.

---

[41] MZmine Development Team, MZmine 2.3 User Manual, Exact Mass Calculation, pg. 19 (2011), http://mzmine.sourceforge.net/manual.pdf (accessed Oct, 10, 2016).

[42] Agilent Technologies, Mass Accuracy and Mass Resolution in TOF MS, pg. 13 (Oct, 2011), http://www.agilent.com/cs/library/eseminars/public/Mass%20Accuracy%20and%20Mass%20Resolution%20-%20October%202011.pdf (accessed Oct, 10, 2016).

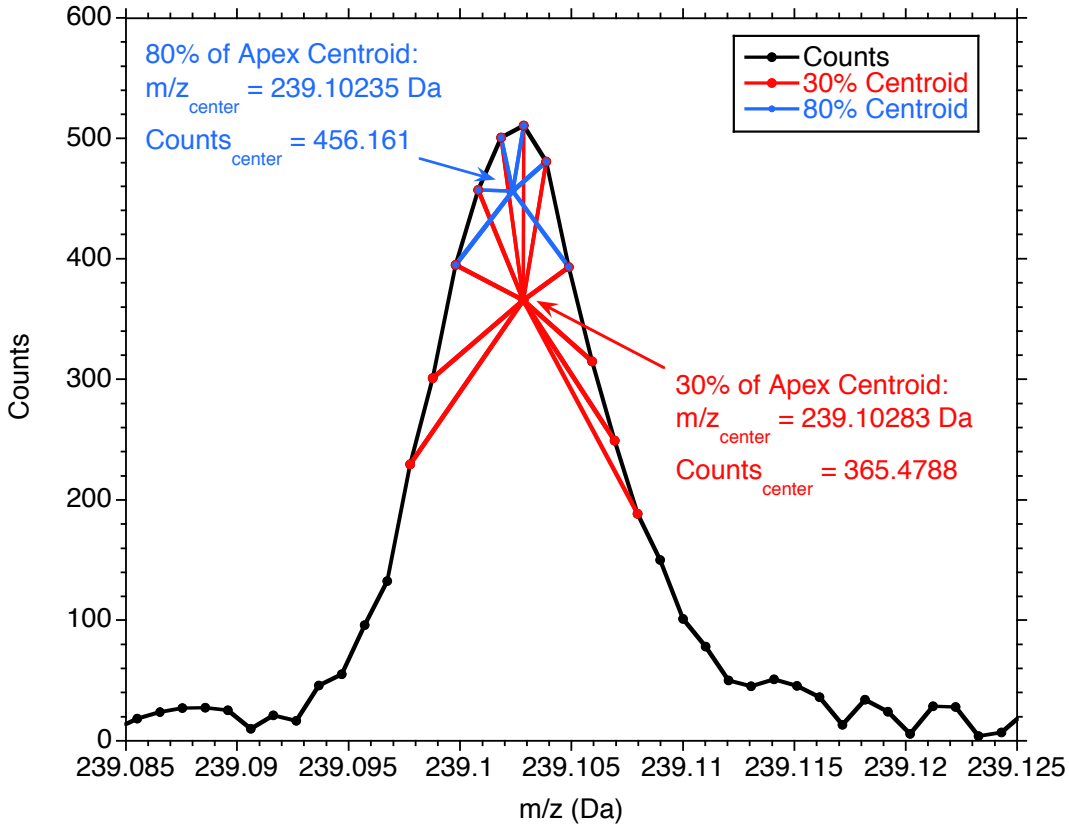


Figure 9.2. Determination of the center of mass of a mass spectral peak using all points within 30% of the peak apex abundance and all within 80% of the apex abundance. Note, the calculated center of mass shifts by 2 ppm between these two point selections because the peak shape is asymmetrical.

Fitting of Peak Models

An alternative to classical centroiding is to define a model peak that is fit to the spectrum at each spectral apex. Usually a Gaussian[43, 44] peak shape model is used, but other model distributions can also be applied.[45] Consistent instrument-specific deviations from the model distribution can also be calculated and added to the core peak shape model to improve the quality of the fit where necessary.[46] A major benefit of this approach is that spectral noise is effectively averaged out by the model fitting process. Both mass and abundance are determined simultaneously when the model is successfully fit to the experimental peak.

---

[43] Wang, Y., Methods for operating mass spectrometry (MS) instrument systems," US6,983,213 (3 Jan 2006).

[44] Hall, M. P. et al., "Mass defect tags for biomolecular mass spectrometry," *J. Mass Spectrom.,* **38**:809-816 (2003).

[45] Leopold, P. et al., "Peak shape self-modeling for low abundance analytes in complex mixtures," http://www.positiveprobability.com/POSTERS/2006Modelling.pdf

[46] Wang, Y., "Methods for operating mass spectrometry (MS) instrument systems," US6,983,213 (3 Jan 2006).

The primary disadvantage of this approach is that an appropriate peak model (or characteristic deviation from a known model)[47] must be determined for the spectrum of interest. This can be exceptionally difficult where the peak shape varies with m/z as in all mass analyzer types except ion trap. Even in ion trap analyzers the peak shape depends on the time spent in the trap (i.e., becomes narrower at longer trap retention times). Any change in instrument tuning parameters or analyzer mass range also forces changes to be made in the peak model. Therefore, if the model is drawn from the spectrum itself, it must be based on well-isolated training peaks within a narrow m/z range of the target peak to which it is applied for centroiding.

This still leaves the issue of where to apply peak models to the spectral data (i.e., peak detection discussed above). One approach is to start at the most abundant apex and work downwards until the threshold of detection is reached. If the peak model adequately explains the peak it can be subtracted from the spectrum with minimal residual (i.e., the residual error is less than the threshold specified). The next highest apex in the residual spectrum is then fit with the model, the model subtracted and this process repeated, until there are no longer any residuals above the spectral threshold. While computationally laborious, the basic principle appears sound on the surface.

Problems arise, however, when there is more than one nearly isobaric peak near the apex forcing the experimental peak to deviate from the peak model, or the single peak shape model does not adequately explain the observed peak shape and leaves a residue higher than the threshold, which becomes a false positive detection on a subsequent pass. Overlapping peak abundances are additive, so the presence of a partially-overlapping, nearly-isobaric, side peak increases the apparent abundance at every mass point of overlap, and alters the center of mass of any peak with which it overlaps. In the presence of such a peak overlap, the first peak model fit to the highest apex will carry both a mass error and abundance overshoot. After subtraction of the first fitted peak model, the resulting residual peak(s) will also exhibit shifted center(s) of mass, and its (or their) abundance(s) will then be under-estimated by fit of the second model. This is illustrated in Figure 9.3.

[47] Wang, Y., Methods for operating mass spectrometry (MS) instrument systems," US6,983,213 (3 Jan 2006).

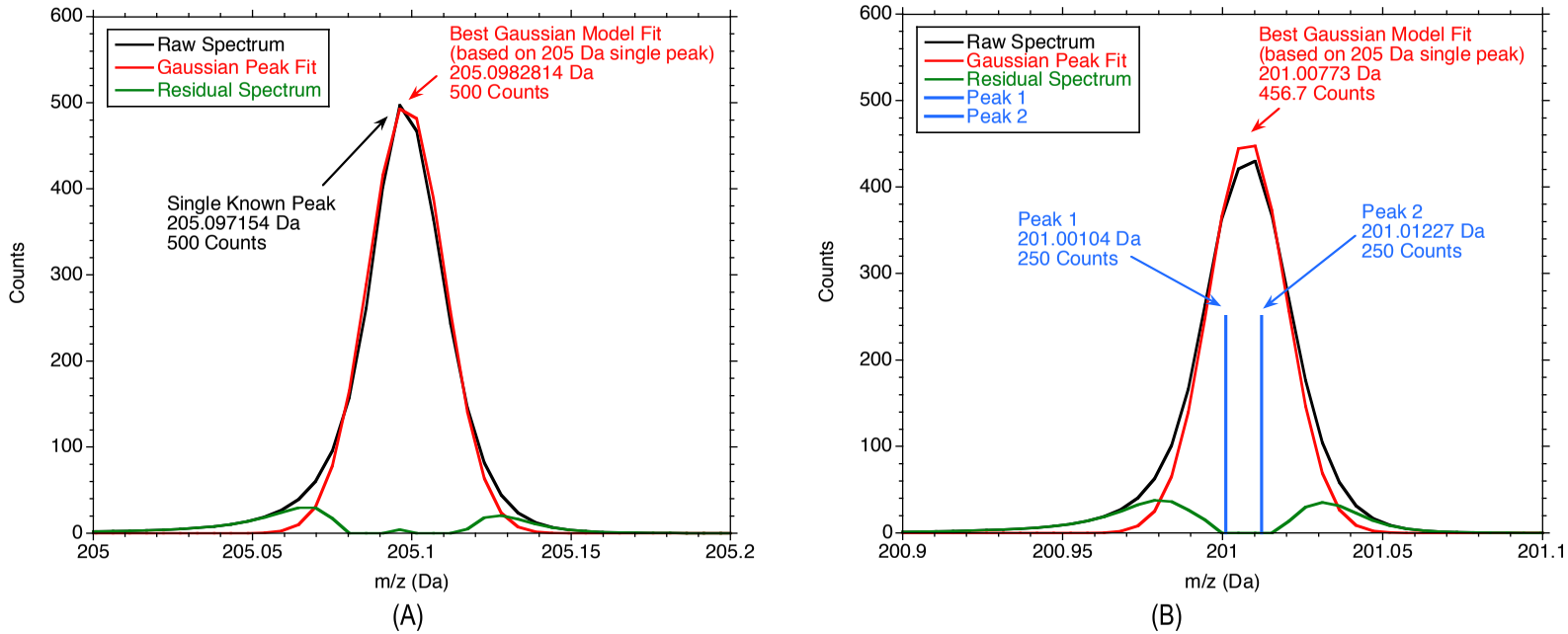


Figure 9.3. A Gaussian peak model is developed on a known isolated 500-count training peak at 205 Da in a synthetic TOF spectrum (Panel A). The deviation of the real peak from a true Gaussian peak shape is seen in the residual spectrum. This model is then fit to neighboring peak in the same spectrum at 201 Da, which appears to be a single peak, but is in fact an overlapped pair of nearly isobaric 250-count Gaussian peaks (Panel B). The mass error of the first single peak model centroid versus the actual underlying peaks is 33 ppm to Peak 1 and 23 ppm to Peak 2. The abundance is over-estimated by 83% because the Gaussian peak model determined from the 205 Da training peak is narrower than the fattened 201 Da peak composed of nearly isobaric overlapped peaks. After subtraction of the single peak model, the residual spectrum contains two widely separated residual peaks with much lower abundances than either of the underlying peaks.

The method of fitting the model to the spectrum is also an important decision. The observed experimental mass points may straddle the actual center of mass position, so the observed peak apex may not be co-located with the centroid. Each peak in the spectrum also has a unique abundance associated with its centroid, which is not simply the highest observed point in the experimental peak, but must be determined during model fitting. So, optimization of the model fit must allow both mass and abundance shifts of the model, typically through minimization of the least squared residual error of all the abundances that make up the rising and falling edges of the peak. The model abundances must be recalculated at each measured mass point every time the mass position of the model is shifted. Then there is the problem of which experimental and model points to use in the optimization. Should the data used in the least squared error calculation be limited to the width of the model peak or to those above the S/N threshold? For the same average noise level, there is a higher percentage abundance error in experimental mass points that are close to the threshold than for mass points that are higher in abundance; therefore, should the former be discounted or eliminated from the least squares objective function of the optimization? Since noise error is always additive to the peak, should just the positive residuals be included in the objective function and the negatives ignored? There is no clear and unambiguous guidance in the literature to answer these questions.

More importantly, how should the model fitting algorithm decide when more than one model should be fit to a given peak (as in the example of Figure 9.3B)? Where two or more nearly isobaric peaks are known or suspected, the number of models to apply could be specified by the user, as in MassWorks™ (Cerno Bioscience, Norwalk, CT); however, this cannot be automatically applied to unknown peaks without some goodness of fit criteria being used to justify the addition of another peak model.

This brings us to the problem of when to stop a serial model peak-fitting process. As discussed previously,[48] noise is always additive in mass spectrometry. Standard (parametric) statistical methods (e.g., goodness-of-fit ANOVA or χ^2) cannot be applied to this problem, as they produce random variations in the p-score as subsequent peaks are added to the fit. In the example of Figure 9.3b, the two model fit to the overlapped peak pair at 201 Da generates a lower residual error than the single model fit and is better aligned with the theoretical masses and abundances, but the goodness-of-fit ANOVA p-score of the two model fit is higher than that of the single model fit. Yet, adding a third model to the fit improves the p-score over that of either the single or two model fits, even though there is little improvement in the residual error over the two model fit. The underlying problem is that experimental peaks are either over-fit or under-fit by model peaks because of the superimposed one-sided spectral noise. The standard goodness-of-fit ANOVA implicitly assumes that the noise is evenly distributed across both sides of the model curve, when the reality is that it is distributed only to one side. Therefore, standard parametric statistical methods fail to reliably optimize the number of partially-overlapped peak model fits. Furthermore, unless a limit is placed on the number of simultaneous models applied to a particular apex, smaller and smaller model peaks will continue to be added to the same peak until all the spectral noise above the threshold of detection is completely modeled.

Another approach is to insert a single peak model at every mass point above the threshold in the spectrum, and not allow these models to move in the mass dimension. The resulting over-specified sparse matrix of simultaneous equations can then be solved for the global least squares height of all the models, and all those models with optimized heights less than the threshold eliminated as noise. The remaining model heights are then re-optimized in both the mass and abundance domains. This approach creates a massive n x m matrix optimization process, where n is the total number of mass points in the spectrum file and m is

[48] Spectral Characteristics.docx

the number of non-zero abundance points in the spectrum file (i.e., m models). Furthermore, this process results in loss of mass precision since the closest that the centroid of any model can get to the true mass is ± the intrinsic mass spacing of the spectrum at that mass position. In a typical mass spectrum, peaks are only 5-7 mass bins wide, so the mass accuracy of this method is effectively limited to 1/5 to 1/7 of the spectral peak width, which would constitute an unacceptable reduction in mass resolution. Furthermore, where the true measured peak apex falls between two experimental mass points, this method has a tendency to fit two models of half height into the spectrum at that peak.

The challenges posed by peak model fitting have greatly inhibited its adoption, and finite difference peak detection methods continue to prevail as the method of choice in almost all spectral centroiding software packages. MassWorks™ (Cerno Bioscience, Norwalk, CT), which continues to apply a modified-Gaussian peak model fitting strategy, is the only commercial exception noted at this writing.

Abundance Quantification

Continuum Spectrum Intersection

Once a center of mass has been determined for a peak, there are two basic options for determining its abundance. The simplest option is to draw a vertical line from the baseline (or threshold) up to where it intersects the spectrum (or smoothed spectrum). If another partially-overlapped nearly isobaric peak is present, the abundance of both peaks will be over-estimated by this method. This can be easily visualized in Figure 9.3B, where the true heights of the overlapped nearly isobaric peaks are shown, but if these were correctly detected by the centroiding algorithm, then a line drawn to intersect the spectrum at each mass would overestimate each of the peak abundances by about 50%.

Peak Area

An alternative approach is to estimate the area contained under the peak and use this as the abundance. This has an advantage in detecting the presence of partially overlapped ions in an isotopic pattern since the members of the isotopic series will not display the expected ratios if overlapped by another nearly isobaric species. However, the challenge here is how to define the ends of a peak. Is it where the peak intersects the threshold? If so, what happens when the trough between to neighboring peaks is elevated above the threshold due to the overlap? Such troughs can be found by finite difference calculus as positioned where the first derivative goes to zero and the second derivative is positive. However, this simple boundary test fails to function when the peak asymptotically approaches a constant baseline. This problem is common to that experienced by every chromatographer in trying to determine the proper limits of analyte peaks, and for which there has never been a fully satisfactory automated solution.

Use of Model Peaks

Where multiple peak models are fit to an experimental peak (e.g., Cerno MassWorks™), the combined overlap may more effectively deconvolve the correct abundances of each of the overlapped peaks versus the abundance results from standard centroiding. As discussed previously, however, this assumes: 1) that the peak model is a good approximation of the peak shape; and 2) that the number of overlapping peaks is known with certainty. The additive composite of the multiple peak models is fit to the spectrum, so regions of spectral overlap are not counted twice. For the example of Figure 9.3, fitting a pair of the Gaussian peak models from 205 Da (Figure 9.3a) to the known peak pair at 201 Da (Figure 9.3b) results in one peak disappearing to zero counts and the other peak centering to become the Figure 9.3b result. Constraining both peaks to a minimum counts of 240 yields the results of Figure 9.4, where the mass error for the first peak is 33 ppm and the abundance of that peak hits the 240 count

constraint. Fitting multiple peaks is very difficult because the non-linearity of the model creates lots of local minima.

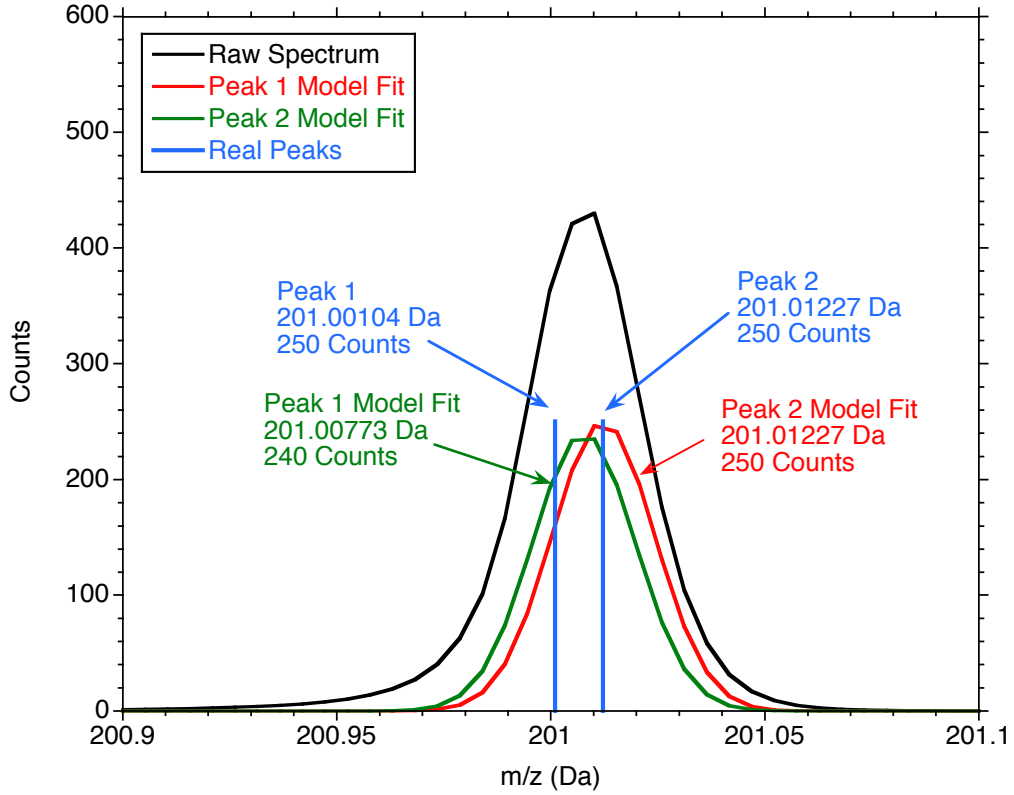


Figure 9.4. The synthetic spectrum of Figure 9.3b with two Gaussian models fit simultaneously to the known nearly isobaric overlapped peak pair at 201 Da.

PeakInvestigator™

The rapidly increasing volume and importance of mass spectrometry in analytical chemistry and the life sciences (including medical research and clinical diagnostics) have created a need for a more automated and robust method to convert raw profile mass spectral data into accurate mass lists for further processing. Such a new approach has been developed by Veritomyx, Inc. and is accessed via their PeakInvestigator™ software services. PeakInvestigator treats the raw mass spectrum as a periodically-sampled continuous signal and uses a proprietary self-trained algorithm to identify and quantify the features of that signal with statistically-measurable precision. The advantages of this approach are:

- no user-adjustable parameters
- fully-automated and locally-adaptive baselining and statistically-determined S/N thresholding for optimized sensitivity above background noise
- reproducible and more precise masslists
- fully-automated spectral detection and deconvolution of nearly isobaric overlapped peaks, effectively providing up to four times the spectral resolution of the native mass analyzer
- statistical confidence (error bars) defined for the mass and abundance result on every peak reported.

While the algorithms are proprietary, it is possible to compare the results to the alternatives described above.

Lipidomics Study

In an LC/MS/MS lipidomics experiment[49], human plasma samples from a diabetes study were analyzed at 10K resolution on an Agilent 6530 and with their standard centroiding software. An overlapped set of peaks were detected in the MS^1 spectrum only after tandem MS analysis of the peak and confirmation by re-running the sample on a 100K resolution LECO Citius instrument. The peaks were found to correspond to phosphatidylethanolamine (36:2) and plasmenol-phosphatidylcholine (P34:1). These two nearly isobaric species were unresolved by standard centroiding methods in any of the 10K resolution MS scans in the entire study.[1] Nor could they be resolved chromatographically since they were nearly co-eluting. The relevant LC/MS scans for over 100 patients were provided to Veritomyx for PeakInvestigator analysis. The PeakInvestigator software was able to blindly detect and deconvolve both peaks (Figure 9.5) in scans for 98% of the patients, allowing them to finally be resolved chromatographically (Figure 9.6). Not only did PeakInvestigator correctly identify and deconvolve the peak pair in question, but it also found an additional 40 previously undetected pairs of nearly-isobaric peak overlaps within the same samples, yielding a remarkable and unanticipated new discovery rate opportunity.

[49] PeakInvestigator™ Deconvolution & Centroding Software: UC Davis Beta Collaboration-Phase 1, https://veritomyx.box.com/s/v15u85f4b47nyge71nq7bnxe4zy7dwul (Accessed 9/15/16).

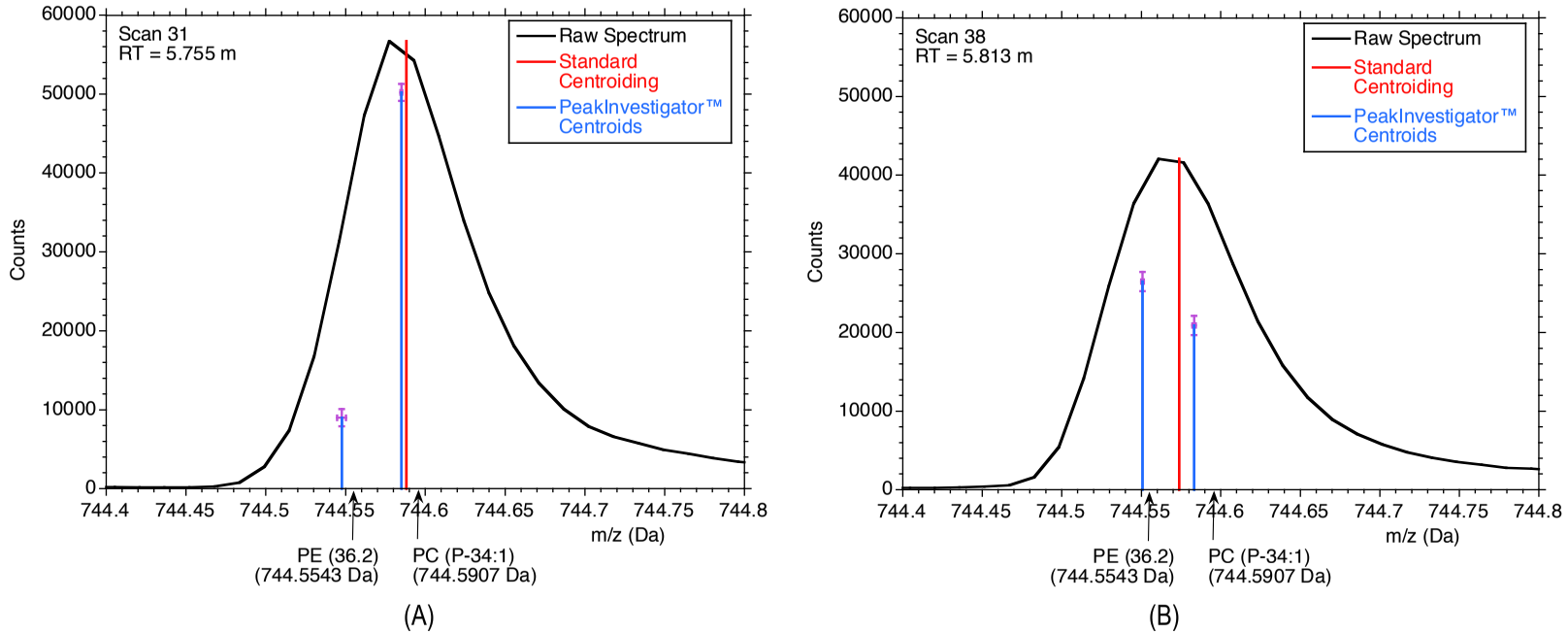


Figure 9.5. Single Agilent 6530 scans at 10K resolution, for (A) 5.755 min and (B) 5.813 min retention times, from an LC/MS/MS lipidomics study on human plasma. In the region of 744.5 Da there were two nearly-isobaric overlapped peaks that were only detected by tandem MS and confirmed in a higher resolution TOF instrument (LECO Citius). These peaks could not be resolved in any scan by standard centroiding software, even though the overlap was known. PeakInvestigator™ was able to blindly identify and quantitatively deconvolve these overlapped peaks in the 10K resolution scans, allowing separate chromatograms to be assembled for each peak (Figure 9.6). The mass and abundance error bars, a feature uniquely available in the statistically-driven PeakInvestigator algorithms are shown for each peak found.

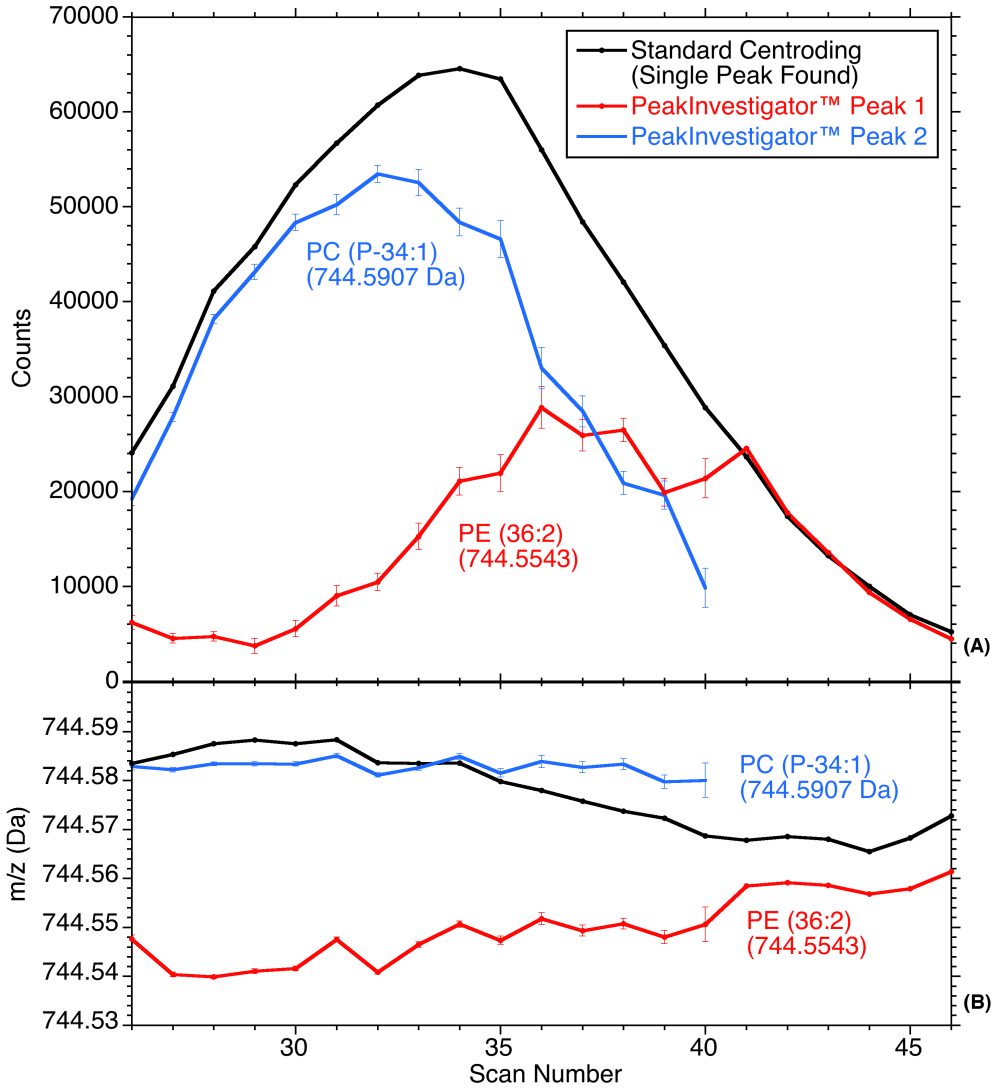


Figure 9.6. Results of standard centroiding (Exact Mass in MZmine) versus PeakInvestigator centroiding of the two nearly-isobaric peaks from each of the 20 LC/MS scans in which either component eluted. The reported mass of the standard centroid (Panel 6B) is seen to decline with increasing scan number as first one component and then the other dominates the peak found. Versus the two masses deconvolved and reported by PeakInvestigator, the mass offset of the standard centroiding result is caused by the asymmetry of the peak shape (Figure 9.5) as discussed in the text. Each component is correctly isolated by PeakInvestigator with little overall mass drift. The corresponding abundance chromatograms (Panel 6A) show the quantitative precision of PeakInvestigator across the changing dynamic range of these two peaks. PeakInvestigator is uniquely able to provide both mass and abundance error bars (indicated). It should be noted that error bars less than the sampling spacing of the profile spectrum are effectively zero since the minimum mass tolerance of the PeakInvestigator method is the intrinsic mass spacing of the spectrum (discussed further in the text).

Nearly Isobaric Admixtures

The following two admixtures of nearly-isobaric chemicals were prepared and analyzed by ESI Q-TOF (Agilent 6550) at 20K resolution and by Q Exactive Plus Orbitrap at 100K resolution.[50] The spectra were acquired both in raw format and in centroid mode on the Agilent instrument. The Agilent raw spectra were re-analyzed with Exact Mass standard centroiding (MZmine v2.0) and by PeakInvestigator.

The first admixture contained two components differing in mass by 90 ppm:

- 4-Imidazole acetic acid ($C_5H_7N_2O_2^+$ - $[M+H^+]$ = 127.050752 Da)
- 5-Aminoimidazole-4-carboxamide ($C_4H_7N_4O^+$ - $[M+H^+]$ = 127.061986 Da)

These results are summarized in Figure 9.7. Panel A shows the acquisition in centroid mode on the Agilent 6550, which produced a single centroid and an extraneous noise peak. Reanalysis using exact mass (Panel B) produced two standard centroids differing by just 93 ppm (a 3 ppm precision variance). However, the relative abundances of the two peaks were within 16% of each other because they were drawn from zero until they intersected the spectrum curve. This varies greatly from the 70% relative abundance difference seen in the fully-resolved Orbitrap spectrum (Panel D). The PeakInvestigator result (Panel C) also deconvolved the two peaks with 92 ppm difference (a 2 ppm precision) and deconvolved the relative abundances of the two species to a relative abundance difference of 67%, very close to the 70% difference found in the Orbitrap result (Panel D).

The second admixture contained three components differing in mass by 65 and144 ppm, respectively:

- N-Acetyl-L-ornithine ($C_7H_{15}N_2O_3^+$ - $[M+H^+]$ = 175.108267 Da)
- L-Arginine ($C_6H_{15}N_4O_2^+$ - $[M+H^+]$ =175.119501 Da)
- Ne,Ne+dimethyllysine ($C_8H_{19}N_2O_2^+$ - $[M+H^+]$ = 175.144653 Da)

These results are summarized in Figure 9.8. Panel A again shows the acquisition in centroid mode on the Agilent 6550, which detected the second two of the three peaks at the correct 144 ppm mass tolerance. However, the relative abundances of these two peaks were reversed from that of the Orbitrap result because of the added contribution of the first component to the relative abundance of the second peak from which it was not resolved. Panel B again shows the Exact Mass result from the raw spectrum, with only the second two peaks being resolved by standard centroiding. In this case the mass difference between the two peaks was 147 ppm (a 7 ppm precision error) and the relative abundances of the two detected peaks were again reversed because of the added contribution of the first component counts that were unresolved from the second peak. PeakInvestigator resolved all three peaks (Panel C) with mass differences of 64 ppm for the first two peaks (a 1 ppm precision error) and a 142 ppm mass difference for the second two peaks (a 2 ppm mass precision error). The abundances of the three peaks were also much closer to that seen in the fully-resolved Orbitrap result (Panel D).

[50] PeakInvestigator™ Deconvolution & Centroiding Software, Stanford Beta Collaboration, https://veritomyx.app.box.com/s/u14tsrsrg36xrie1okotpi9sdqv03zzg (Accessed 9/15/16).

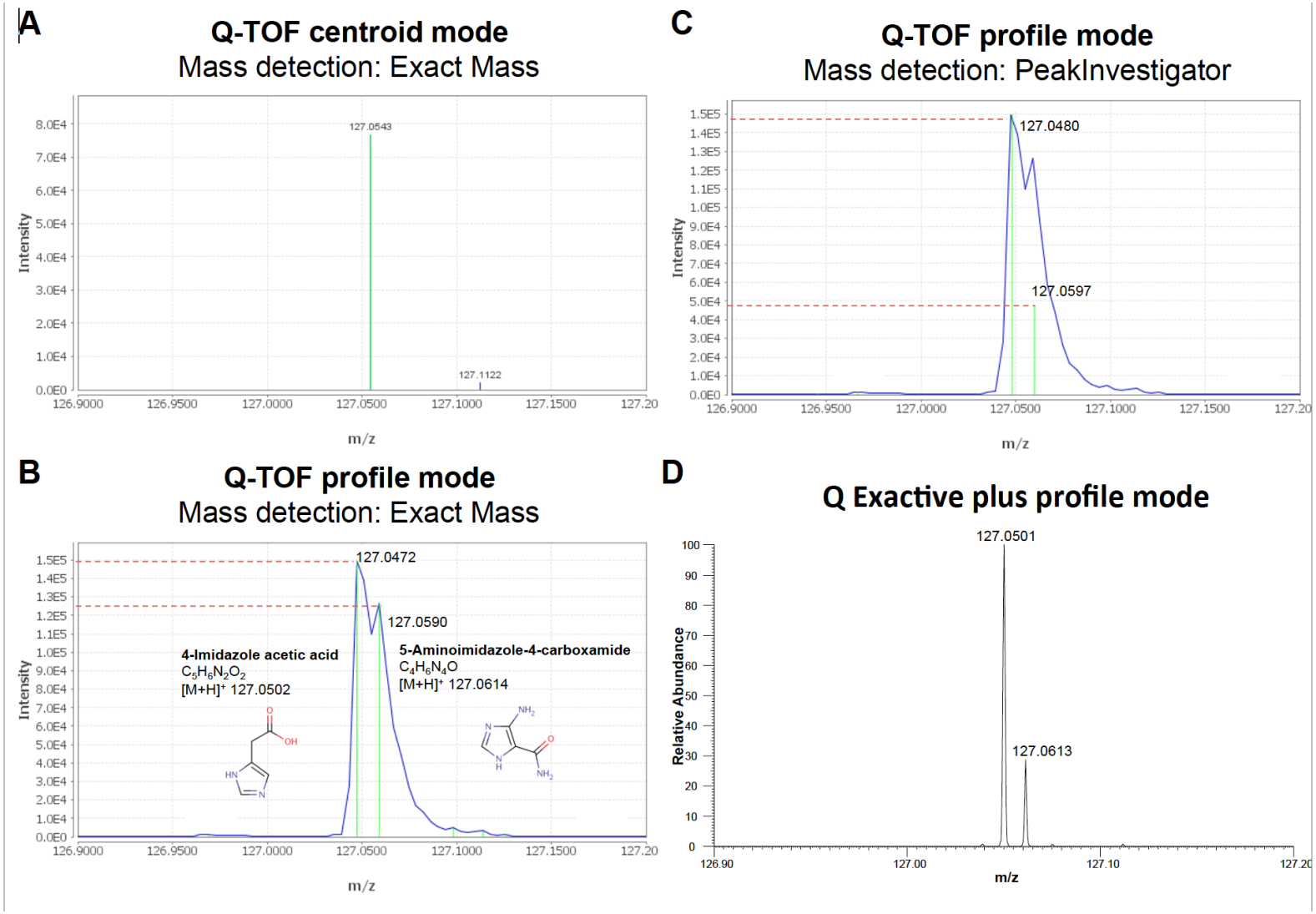


Figure 9.7. Resolving admixtures of nearly-isobaric species by standard centroiding methods (Panels A and B) and PeakInvestigator (Panel C) from spectra acquired in postive ESI mode on an Agilent 6500 at 20K resolution. The same admixture was fully resolved on an Orbitrap Q Exactive Plus at 100K resolution (Panel D).

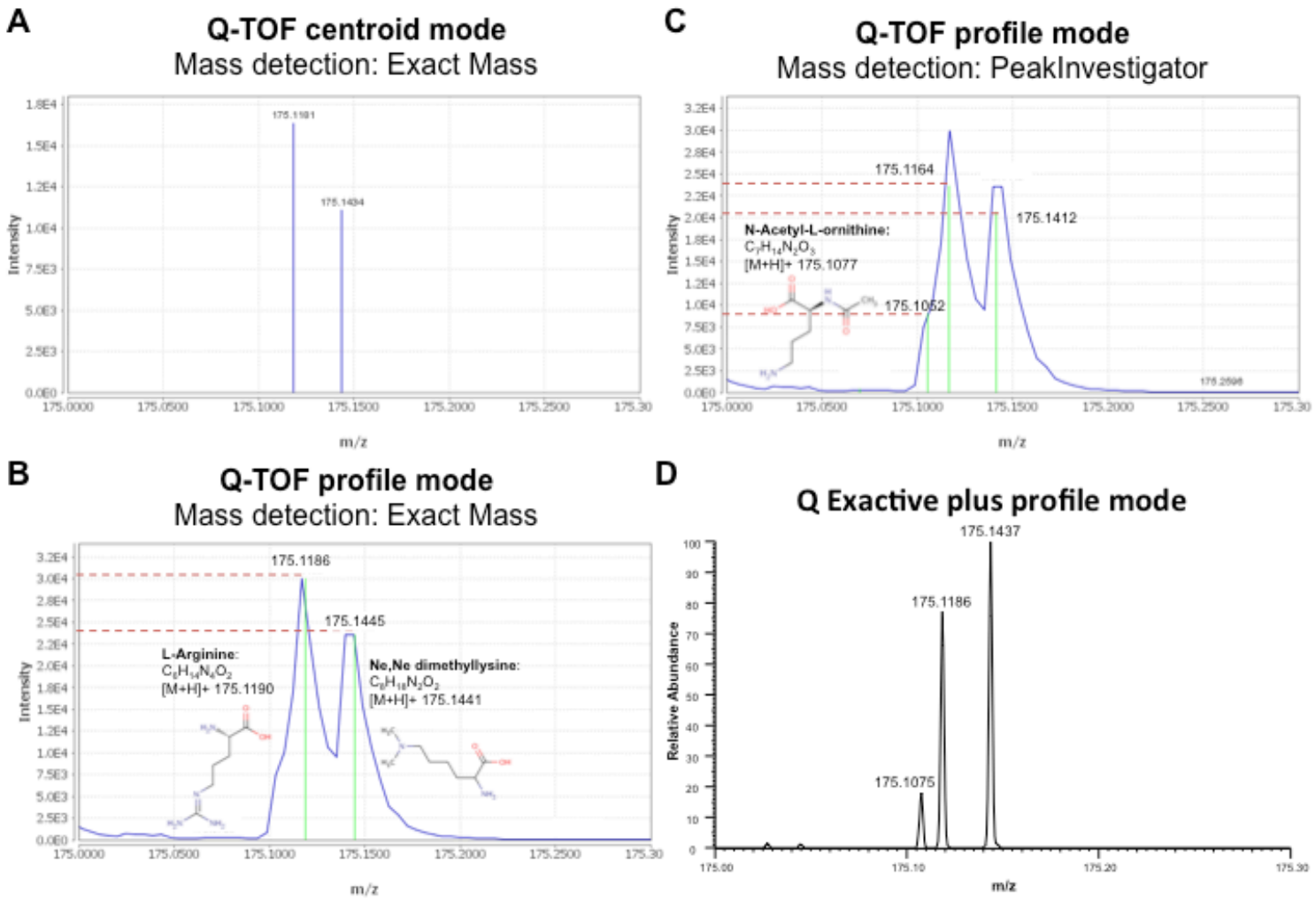


Figure 9.8. Resolving admixtures of three nearly-isobaric species by standard centroiding methods (Panels A and B) and PeakInvestigator (Panel C) that were acquired in positive ESI mode on an Agilent 6550 at 20K resolution. The same admixture was resolved on an Orbitrap Q-Exactive Plus at 100K resolution (Panel D).

Peptide Abundance in Orbitrap

By determining the rate of stable isotope incorporation into proteins, (or their constitutive peptides) the rate of protein synthesis in cellular systems can be estimated. In such studies, it is critically important to use the full isotopic abundance of the peptides, yet often multiply-charged states of these peptides can overlap and remain unresolved from one another making it difficult to get an accurate measurement of all the members of the isotopic pattern. In the following scan (Figure 9.9) we see the overlap of such tryptic peptides. $Peptide_1$ consists of a doubly charged species of a peptide of the nominal $[M+H^+]$ composition $C_{39}H_{62}N_{11}O_{12}$. $Peptide_2$ consists of the triply-charged species of a peptide of the nominal [M+H+] composition $C_{58}H_{93}N_{17}O_{17}S_1$. The $^{13}C_2$ peaks of both peptides are nearly isobaric in this 30K resolution Orbitrap spectrum. Standard centroiding (Exact Mass in MZmine) fails to independently resolve these two species, but PeakInvestigator does quantitatively resolve the isotopic pair.

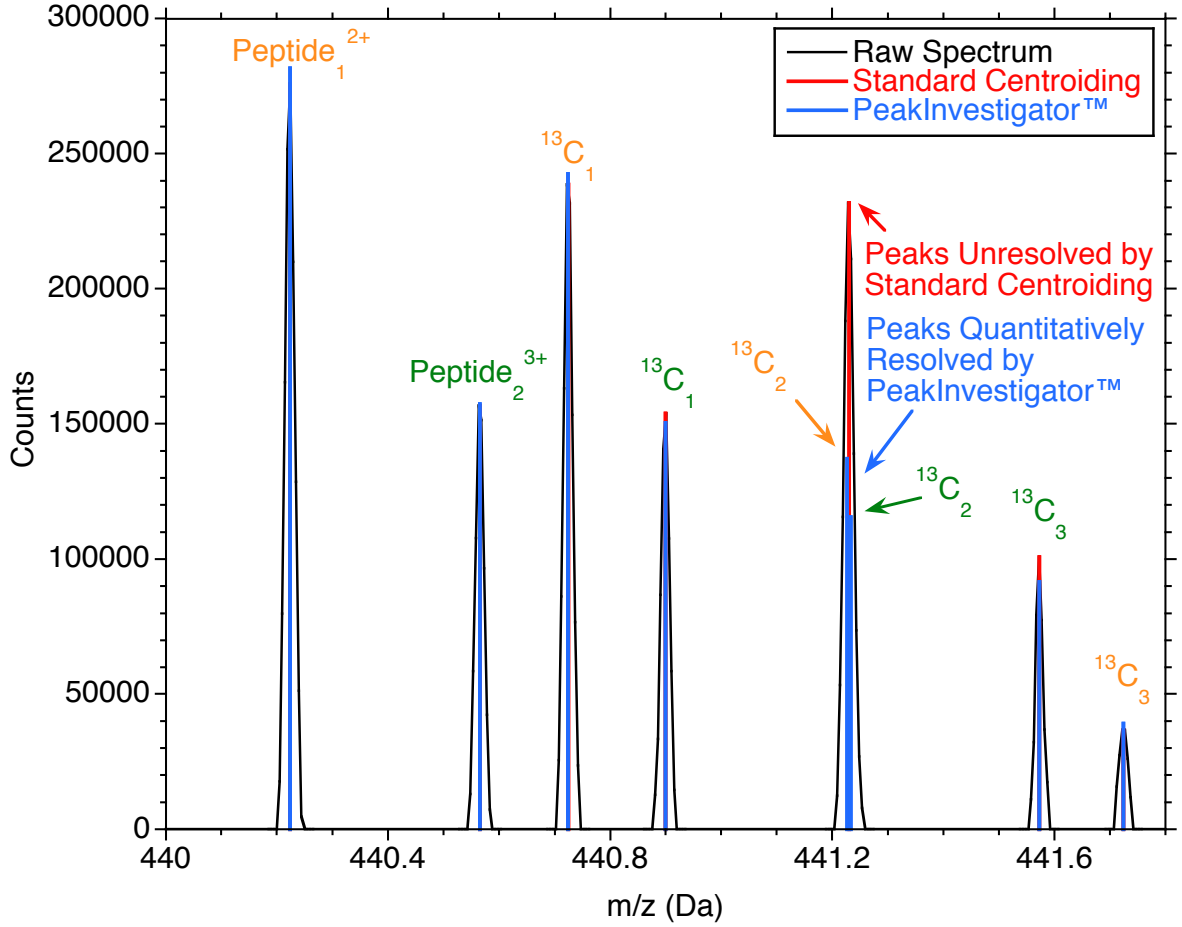


Figure 9.9. Overlapped second ^{13}C isotopes of two peptides in an 30K resolution Orbitrap mass spectrum are not detected by standard centroiding, but are blindly and quantitatively deconvolved by PeakInvestigator. $Peptide_1$ has the nominal composition $C_{39}H_{62}N_{11}O_{12}$. $Peptide_2$ has the nominal composition $C_{58}H_{93}N_{17}O_{17}S_1$. The isotopic vector angles for each peptide compared to their theoretical patterns obtained from standard centroiding (where the overlap is unresolved) show a 32.8% and 19.9% errors respectively. PeakInvestigator quantitatively resolves

these peaks improving the isotopic vector angle errors to 22.6% and 7.8%, respectively.

In stable isotope labeling experiments like these, Isotopic Vector Angle Analysis[51] is used to compare the centroided abundances to those expected from theory, effectively incoporating the relative abundances of all members of the isotopic series. In this case PeakInvestigator™ provides an average of 46% better isotopic match for these two peptides than could be obtained by standard centroiding methods.

Peak Error Bars

The statistical nature of the PeakInvestigator peak finding algorithms enables the estimation of both mass and abundance error bars for every peak found. These error bars are derived with statistical confidence. The reported values are $\pm 1\sigma$ (68% confidence) and can be multiplied by the appropriate 2-tailed Student's t-value for the statistical degrees of freedom reported for each peak to get estimates for the statistical confidence intervals.,[52].

Mass Precision

Unfortunately, the mass spectrum is a sampled distribution in the mass domain. Therefore, there is a lower limit to the measured mass precision (i.e., effectively, the PeakInvestigator algorithm's intrinsic "Nyquist rate"[53]), which is one intrinsic mass spacing.[54] It should be remembered that the units of the intrinsic mass spacing of any mass spectrum are not expressed in units of mass (except for ion trap) but are related to mass by simple mathematical relations. Therefore, the minimum mass precision of any PeakInvestigator peak is the mass equivalent of the intrinsic mass spacing of the spectrum or the confidence limit, whichever is larger.

Abundance Precision

Abundance measurements are effectively continuous in a mass spectrum, but are often reported as integer equivalents. Therefore, the abundance error bar reported can only be as accurate as the least significant figure used to report counts.

Some recent AB/Sciex TOF analyzers use dynamic ion throttling to control ion flow to the detector to prevent detector saturation. The corresponding mass spectrum is automatically scaled by the ratio to which the ion stream has been throttled so that continuity in abundance across the series of autoscaled spectra is maintained. This process can make the minimum abundance error in any given spectrum a multiple of the normal integer count spacing. Therefore, some care must be taken in applying the PeakInvestigator error bars to spectra of this type, since the intrinsic precision limit may be something other than a single significant figure and will vary from scan to scan in an LC/MS run on automated ion-throttled mass analyzers.

---

51 Sokkalingam, N., Schneider, L., Tenderholt, A., Chu, F., Corillo, Y. E., Marshall, A.G., *Deconvolution and isotopic vector analysis for improved peak identification.* Poster presented at: 64th Annual Am. Soc. Mass Spec., 2016 June 5-9; San Antonio, TX, https://veritomyx.app.box.com/s/bm7xuw54mplgujdt4eojakybwnx9kodb (accessed 9/27/2016).

52 Confidence intervals, https://en.wikipedia.org/wiki/Confidence_interval.

53 Nyquist frequency, https://en.wikipedia.org/wiki/Nyquist_frequency.

54 Spectral Characteristics.docx.

Deconvolution Resolution

Mass spectral resolution is often defined by the mass of a peak divided by the width of that peak at half the peak height in Da (Equation 9.4). Therefore, resolution is a dimensionless number. It is similarly possible to define the ability of any centroiding process to correctly discriminate any two peaks in a mass spectrum, as the ratio of the the average mass of the two peaks divided by the mass difference between the two peaks (Equation 9.5). We can call this the Discrimination Resolution. As the distance between the two peaks becomes negligible approaching the mass of a single peak at the average mass position, this analysis can be taken one step further to define another dimensionless number, Deconvolution Resolution. The Deconvolution resolution is defined as the ratio of the discrimination and spectral resolution numbers (Equation 9.6). This, of course reduces to the actual mass difference between any two peaks divided by the width of an isolated peak in the spectrum at half its height, where the average mass is the same as the mass used in the spectral resolution calculation.

$$Spectral\ Resolution\ =\ \frac{peak\ mass\ (Da)}{peak\ width_{half-height}\ (Da)} \tag{9.4}$$

$$Discrimination\ Resolution\ =\ \frac{average\ peak\ mass\ (Da)}{|peak_1\ mass\ -\ peak_2\ mass|\ (Da)} \tag{9.5}$$

$$\begin{aligned} Deconvolution\ Resolution\ &=\ \frac{Discrimination\ Resolution}{Spectral\ Resolution} \\ &=\ \frac{peak\ width_{half-height}\ (Da)}{|peak_1\ mass\ -\ peak_2\ mass|\ (Da)} \end{aligned} \tag{9.6}$$

Deconvolution Resolution represents the relative spacing between peaks in any mass spectrometer. As the resolution of the mass spectrometer increases, its ability to resolve two peaks becomes greater, yet the relative overlap of those two peaks maintains a constant deconvolution resolution (Figure 9.10). Any mass spectrometer can just begin to resolve two peaks of the same height at a deconvolution resolution of 1, where there is the beginning of a trough between the two peaks. It can easily resolve peaks with lower deconvolution resolutions. The ability to deconvolve two peaks at deconvolution resolutions above one is entirely dependent on the peak picking or centroding software.

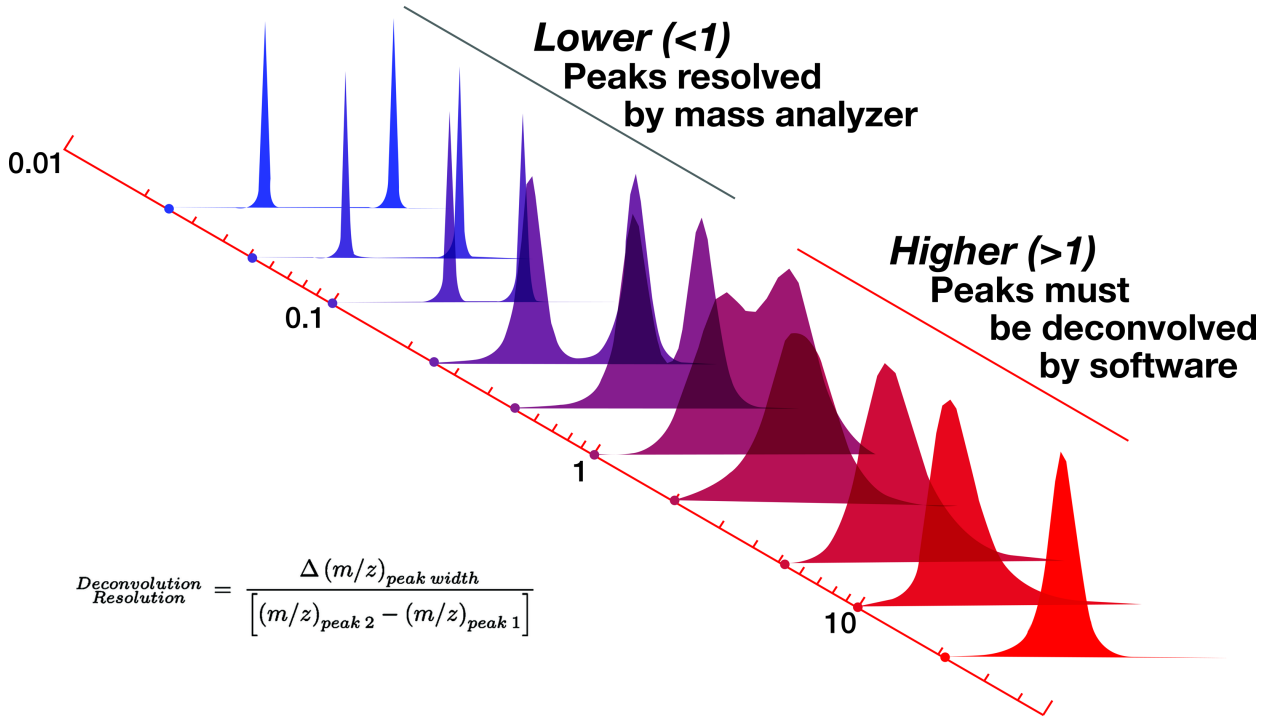


Figure 9.10. The progressive overlap of two peaks as a function of Deconvolution Resolution, a dimensionless variable that quantifies the degree of overlap for adjacent peaks in mass spectrometer outputs at any resolution.

If we evaluate the different centroiding methods presented above on their abilities to blindly detect and deconvolve any two overlapped peaks on the continuum of Deconvolution Resolution, we find that classic finite difference centroiding and peak model fitting (using finite difference methods to locate the peaks as implemented by Cerno Massworks™) have limiting deconvolution resolutions of approximately one (Figure 9.11).

When more than one peak is known (or suspected) at any given mass position, it is possible to guide Cerno Massworks™ to a higher limiting deconvolution resolution, but this can not be achieved on a blind basis as automatically delivered in PeakInvestigator. A limiting decovolution resolution of one corresponds to the point where the “saddle” or trough between overlapping peaks is approaching the point of disappearance (Figure 9.10).

PeakInvestigator multiplies and extends the limiting deconvolution resolution up to four-fold over the current centroiding techniques discussed above, for peaks with strong signal to noise ratios. PeakInvestigator outperforms the other centroiding methods for all peaks down to a signal to noise of 10 for any relative abundance of the overlapped peak heights.

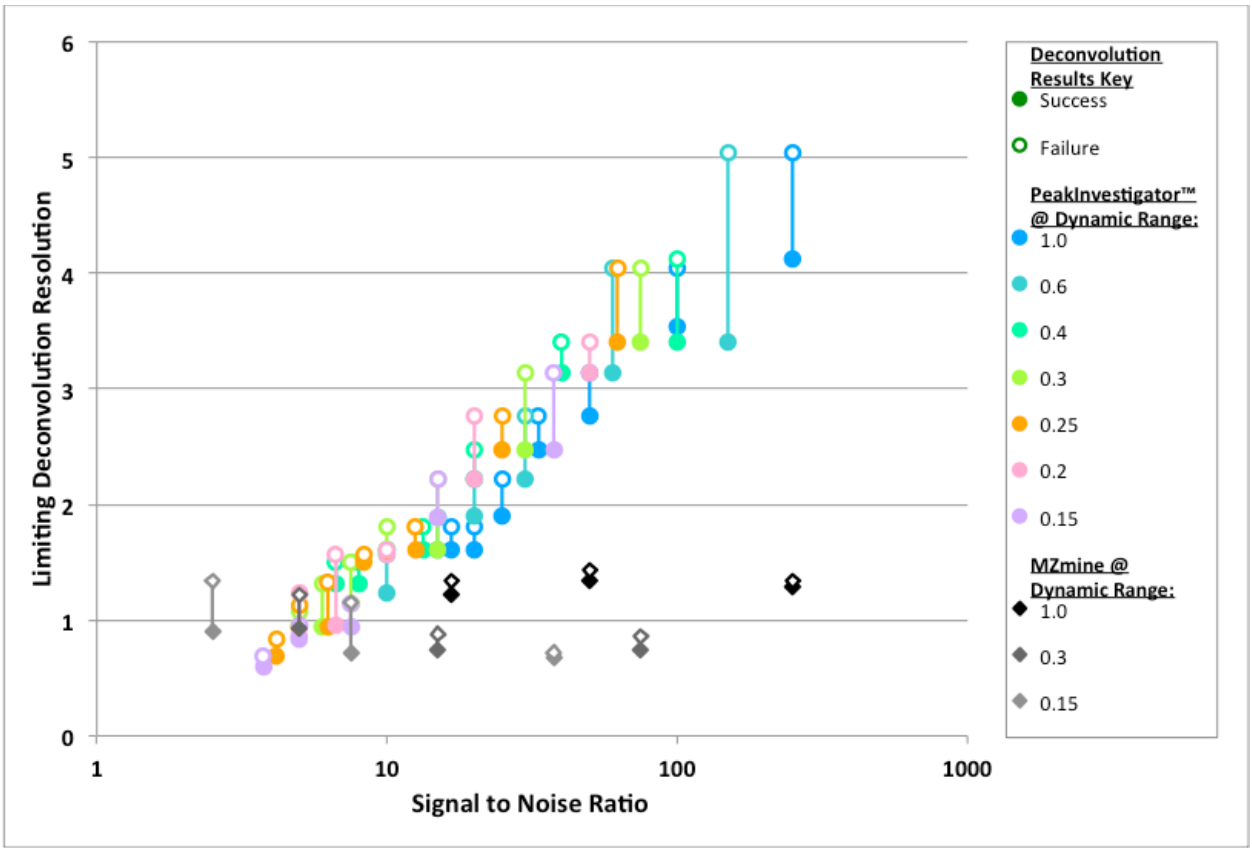


Figure 9.11. The Limiting Deconvolution Resolution of PeakInvestigator and standard centroiding methods, as a function of dynamic range of overlapped peaks and signal-to-noise levels. The true Limiting Deconvolution Resolution is between the solid and open symbols at each condition because it can not be measured any finer than the known spacing between peaks in alternative test spectra.