An Operator-Independent Approach to Mass Spectral Peak Identification and Integration

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2004

Impact Factor: 5.64 · DOI: 10.1021/ac0354701 · Source: PubMed

3 AUTHORS, INCLUDING:

A. J. Kearsley
National Institute of Standards and Technolo…
54 PUBLICATIONS **497** CITATIONS

Charles M Guttman
National Institute of Standards and Technolo…
103 PUBLICATIONS **2,196** CITATIONS

Anal. Chem. **2004,** *76,* 2446−2452

An Operator-Independent Approach to Mass Spectral Peak Identification and Integration

William E. Wallace,*[,†] Anthony J. Kearsley,[‡] and Charles M. Guttman[†]

Polymers Division, and Mathematical and Computational Sciences Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899-8541

A mathematical algorithm is presented that locates and calculates the area beneath peaks from real data using only reproducible mathematical operations and no user-selected parameters. It makes no assumptions about peak shape and requires no smoothing or preprocessing of the data. In fact, it is shown that for matrix-assisted laser desorption time-of-flight mass spectra noise exists at all frequency ranges making the smoothing of data without distortion of peak areas impossible. The algorithm is based on a time-series segmentation routine that reduces the data set to groups of three *strategic points* where each group defines the beginning, center, and ending of each peak located. The peak areas are found from the strategic points using a commonplace polygonal area calculation routine. Peaks with statistically insignificant height or area are then discarded. The performance of the algorithm is demonstrated on a polystyrene mass spectrum with varying degrees of noise added either mathematically or experimentally. An on-line implementation of the method, termed MassSpectator, for public use can be found at www.nist.gov/maldi.

The new generation of mass spectrometers produces an astounding amount of high-quality data very rapidly.[1] Robots for sample preparation combined with instrumentation having automated data collection routines are becoming common features in many analytical laboratories. Such high-throughput experimentation leads to inevitable data analysis bottlenecks. Algorithms that do not require human intervention are needed for rapid and repeatable quantitative processing of spectra that often contain hundreds of discrete peaks. New algorithms that work without user input will not only save operator time but also have the potential to eliminate operator bias.

This second criterion is crucial for NIST's goal of creating a synthetic polymer absolute molecular mass distribution Standard Reference Material by mass spectrometry. One of the key elements in creating such a standard is finding a robust, stable, and reproducible data analysis procedure that does not introduce operator bias. In previous work,[2] NIST conducted a matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) interlaboratory comparison on a low-mass, narrow-dispersity polystyrene homopolymer. Working from the raw data files provided by some of the participants, we compared our reduced values for number-average and mass-average molecular mass with those of the participants and found considerable disagreement in some cases. These disagreements were traced to decisions made by the investigator on the details of how to reduce the data. Surprisingly there was more concurrence in the raw data than in the reduced results. From this it was concluded, "A uniform method of polymer molecular mass distribution integration is needed to eliminate these problems"[2] and later reaffirmed at an industry−government workshop entitled Quantitative Synthetic Polymer Mass Spectrometry held at NIST in November 2002.[3] The work presented here is an attempt to provide such a uniform method. It follows our work on autocorrelation methods for systematic peak identification in complex mass spectra containing hundreds of peaks.[4] Autocorrelation methods allow for peak identification in noisy spectra without operator bias. In this work, we take a different approach not only to identify peaks in real data but to provide their relative areas as well.

Toward this end, a unified collection of algorithms is presented that locates peaks and calculates their associated areas using only reproducible mathematical operations and no user-selected parameters. As shown in Figure 1, the method consists of three steps: (1) statistical characterization of the data set and a corresponding analyte-free data set; (2) data set segmentation to determine the strategic points; and (3) deflation of the number of strategic points guided by the statistical properties of the original spectrum and its congruent analyte-free spectrum. The analyte-free spectrum (sometimes termed a "blank") is a spectrum taken under the same instrument conditions as the spectrum of interest. We refer to this case as being "congruent". The aim of this spectrum is to isolate only the noise without the signal of interest. It is desirable that the noise be of exactly the same character in the spectrum of interest and in its congruent analyte-free spectrum. We have found that this is not always the case for MALDI-TOF MS of synthetic polymers but is often close enough as to have no consequence for the method outlined here. The strategic points are deflated using the characteristics of both the original and the analyte-free spectra. The final deflated set of

* Corresponding author: (e-mail) william.wallace@nist.gov.

[†] Polymers Division.

[‡] Mathematical and Computational Sciences Division.

[1] Kassel, D. B. *Chem. Rev.* **2001**, *101*, 255.

[2] Guttman, C. M.; Wetzel, S. J.; Blair, W. J.; Fanconi, B. M.; Girard, J. E.; Goldschmidt, R. J.; Wallace, W. E.; Vanderhart, D. L. *Anal. Chem.* **2001**, *73*, 1252.

[3] Wallace, W. E.; Guttman, C. M.; Hanton, S. D. *J. Res. Natl. Inst. Stand. Technol.* **2003**, *108*, 79; www.nist.gov/jres.

[4] Wallace, W. E.; Guttman, C. M. *J. Res. Natl. Inst. Stand. Technol.* **2002**, *107*, 1; www.nist.gov/jres.

10.1021/ac0354701
Published on Web 04/07/2004

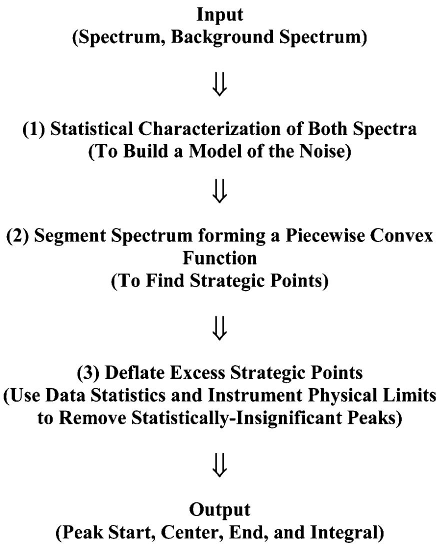
Input
(Spectrum, Background Spectrum)
⇓
(1) Statistical Characterization of Both Spectra
(To Build a Model of the Noise)
⇓
(2) Segment Spectrum forming a Piecewise Convex Function
(To Find Strategic Points)
⇓
(3) Deflate Excess Strategic Points
(Use Data Statistics and Instrument Physical Limits to Remove Statistically-Insignificant Peaks)
⇓
Output
(Peak Start, Center, End, and Integral)

Figure 1. Flow diagram for the suite of algorithms defining the MassSpectator data analysis method.

strategic points consists of groups of three points that define the beginning, center, and end of each peak in the data. Finally, an elementary polygonal fitting routine is used to calculate relative peak area.

After a brief description of the algorithm, examples will be shown using MALDT-TOF mass spectra of the standard polystyrene sample used in our previous interlaboratory comparison.[2] It should be noted that a mathematical description of our segmentation algorithm has been published previously.[5] An on-line implementation of the method for public use can be found at www.nist.gov/maldi and has been given the name MassSpectator for convenience.

DESCRIPTION OF METHOD

Spectrum Segmentation. A nonlinear programming algorithm using an L_2 (least squares) approximation to an L_1 (least absolute value) fit was employed.[5–8] L_1 fits are superior to L_2 fits due to their increased tolerance for outliers; that is, outlying points do not exert as much control over the final fit. Given a data set of *N* points, we find a collection of strategic points and find the unique optimal piecewise linear function passing through the *x* coordinate of each strategic point. This defines a set of function maximums and minimums corresponding to the peak maximums and the peak limits. The peaks' original data are then integrated by finding the area of the polygon determined by the strategic points.

Our segmentation method is a two-step algorithm. The first portion requires the selection of strategic points and is the same as the earlier work of Douglas and Peucker.[9] These points are selected based on an iterative procedure that identifies points whose orthogonal distance from the end-point connecting line segment is greatest. Once a point with greatest orthogonal distance from the mean has been identified, it joins the collection of strategic points and, in turn, becomes an end point for two new line segments from a point with greatest orthogonal distance. This numerical scheme is performed until the greatest orthogonal distance to any end-point connecting line segment drops beneath a prescribed threshold value. This threshold value is the only algorithmic parameter and is based on a statistical analysis of the data and its congruent analyte-free spectrum. Clearly the selection of these points does not require equally spaced data; therefore, the method is equally well suited for TOF data expressed in either time or mass space. Here we chose to work in time space with the data in its most basic state and to eliminate the need to do a point-by-point correction of intensity using partial integrals.[4,10] The second phase of the algorithm, developed specifically for this work, requires the solution of an optimization problem, specifically, locating strategic point heights (that is adjusting strategic point *y*-axis values at their associated strategic *x*-axis value) that minimize the sum of orthogonal distance from raw data. This problem is a nonlinear (and nonquadratic) optimization problem that can be accomplished quickly using a recently developed nonlinear programming algorithm.[11]

Figure 2 gives a graphical representation of the segmentation step of the method using a simple three-peak model with added high-frequency white (that is, uncorrelated random) noise. The first two strategic points chosen are always the first and last points of the data set. A line is drawn connecting these two points, and the data point the greatest orthogonal distance from this line is selected as a new strategic point. This process is iterated over all line segments until the orthogonal distance falls below a threshold parameter calculated from the statistical analysis of the data set and its congruent analyte-free data set as described in the next section. Finally, the strategic point heights are adjusted to minimize the distance from the original (full) data set. Clearly this method requires no knowledge of peak shape and no preprocessing of the data (e.g., smoothing) nor does it require equal spacing of data points.

Strategic Point Deflation. Once the data set is fully segmented, strategic points are discarded in accordance with the statistical analysis of the original data set and its congruent analyte-free data set. This "deflation" of strategic points using statistically derived thresholds is performed by first analyzing the analyte-free spectrum for peaks and peak areas. Once a collection of peaks and peak areas has been accumulated, the spectrum with sample is then analyzed. Each peak identified from the spectrum with analyte is compared to peaks found in relative proximity from the analyte-free spectrum algorithm output (i.e., peaks that appear with similar time or mass coordinates). If any peak in the spectrum with analyte has a smaller peak height or smaller peak area than most (~95%) of the background-spectrum peaks in proximity, then that peak is ignored. Thus, no peak is identified from the sample spectrum that could have been identified by height or area from the background spectrum. This discarding of strategic points also serves to prevent the inadvertent subdivision of larger peaks into a set of smaller peaks. This can sometimes occur if the noise in the analyte spectrum is much greater than the noise in the congruent background spectrum.

(5) Kearsley, A. J.; Wallace, W. E.; Guttman, C. M. *Appl. Math. Lett.* In press.
(6) Barrondale, I. *Appl. Stat.* **1968**, *17*, 51.
(7) Barrondale, I.; Roberts, F. D. K. *SIAM J. Numer. Anal.* **1973**, *10*, 839.
(8) Duda, R. O.; Hart, P. E. *Pattern classification and scene analysis*; John Wiley and Sons: New York, 1973; p338.
(9) Douglas, D. H.; Peucker, T. K. *Can. Cartographer* **1973**, *10*, 112.

(10) Guttman, C. M. *ACS Polym. Prepr.* **1996**, *37*, 837.
(11) Boggs, P. T.; Kearsley, A. J.; Tolle, J. W. *SIAM J. Opt.* **1999**, *9*, 755.

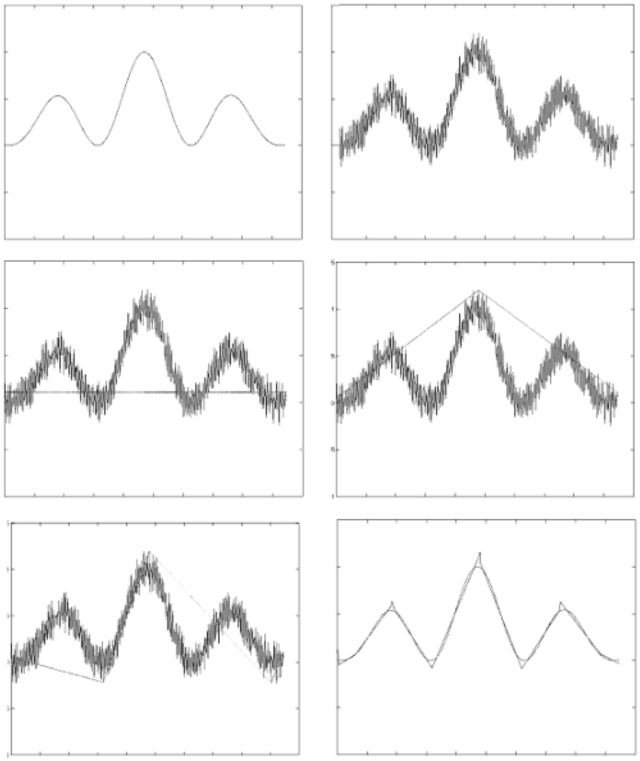

Figure 2. Graphical representation of time series segmentation algorithm used in step 2 of the method. The top left panel shows the ideal measurement response, and the top right panel shows the ideal measurement adulterated with added random noise. The left middle panel shows the first step of the algorithm: the simple selection of the first and last points in the data set, which are always defined as strategic points. The right middle panel shows the location of the point the greatest distance from the line segment connection of the first two strategic points. The lower left panel shows the second iteration where the two line segments found in the previous step are further segmented. The lower right panel shows the final result superimposed on the ideal measurement response. Recall that the y-axis values of the strategic points are adjusted to best fit the data points between them. In this simplified model, the threshold for segmentation was chosen as twice the height of the noise added to the ideal measurement spectrum.

Polygonal Area Calculation. Once the final set of strategic points has been found, the area of the polygon defined by these points is calculated. (The polygon is often, but not always, a triangle. The algorithm will work on polygons of any number of vertexes connected by line segments.) The line connecting the first and last strategic points for a given peak determines a "local baseline". The mathematical basis for the polygonal area calculation algorithm is Green's theorem in the plane and can be interpreted as repeated application of the trapezoidal rule for integration.[12] The method returns the exact area of the polygon.

EXPERIMENTAL SECTION

Figure 3 shows a MALDI-TOF mass spectrum measured using a Bruker Reflex II[13] of a 7190 u mass-average molecular mass polystyrene (available from NIST as Standard Reference Material 2888[14]) obtained using protocol 1 as defined in our previous interlaboratory comparison on this material.[2] Protocol 1 calls for the *all-trans*-retinoic acid MALDI matrix, the polystyrene analyte, and the silver trifluoroacetate salt to be dissolved in tetrahydrofuran in a ratio of 15:1:1 by mass. For the data shown here, this mixture was electrosprayed in ambient at an applied capillary voltage of 5 kV onto the stainless steel target of the MALDI mass spectrometer. This sample preparation was used to ensure sample homogeneity and reproducibility.[15] In particular, Figure 3 shows an optimal spectrum (and its congruent background spectrum) for our instrument in terms of signal-to-noise ratio. Optimization was performed by systematically varying the voltages on the ion optics, the detector voltage, and the nitrogen laser intensity until the best signal-to-noise ratio was achieved. Estimated standard uncertainty (type A) of the peak position from calibration and repeatability studies is 0.2 u, and the estimated standard uncertainty in overall signal intensity from repeatability studies is 15%.

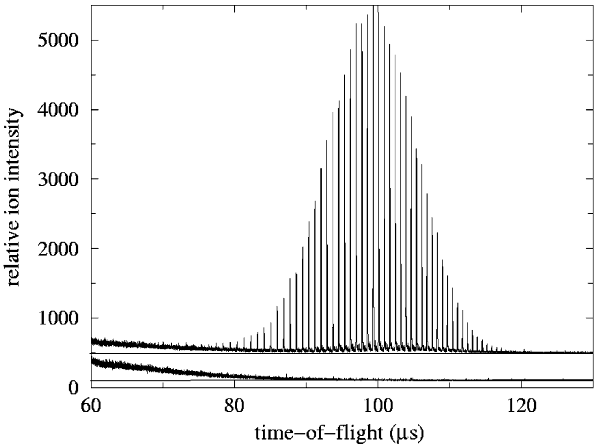

Figure 3. MALDI-TOF mass spectrum of the low-mass, narrow-polydispersity polystyrene (SRM 2888) and its congruent analyte-free spectrum used to demonstrate the method's capabilities. Spectra are offset for clarity. Expanded views of the analyte spectrum are shown in Figures 8 and 9.

STATISTICAL OVERVIEW OF THE DATA

The normal probability plot for the synthetic polymer MALDI-TOF mass spectrum in Figure 3 is shown in Figure 4. If the data (in this case the relative ion intensities) have a standard normal (Gaussian) distribution when plotted against their normal score, the characteristic shape of the plot will be linear.[16,17] (For a tutorial introduction on normal probability plots see section 1.3.3.21 of the *NIST/SEMATECH e-Handbook of Statistical Methods*.[18]) It is observed from Figure 4 that the normal probability plot is far from linear and, in fact, has several obvious sharp changes in slope separating quasi-linear regions indicated by Roman numerals I, II, and III. This signifies that noise is present at a number of widely different frequency ranges. Stated another way, the noise is not purely random (or "white"). In particular, the noise spans the high-

[12] Beyer, W. H. *CRC Standard Mathematical Tables*; CRC Press: Boca Raton, FL, 1981.

[13] The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

[14] http://ts.nist.gov/ts/htdocs/230/232/232.htm.

[15] Hanton, S. D.; Hyder, I. Z.; Stets, J. R.; Owens, K. G.; Blair, W. R.; Guttman, C. M.; Giuseppetti, A. A. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 168.

[16] Filliben, J. J. *Technometrics* **1975**, *17*, 111.

[17] Chambers, J.; Cleveland, W.; Kleiner, B.; Tukey, P. *Graphical Methods for Data Analysis*; Wadsworth International Group: Belmont, CA and Duxbury Press: Boston, MA, 1983.

[18] http://www.itl.nist.gov/div898/handbook/index.htm.

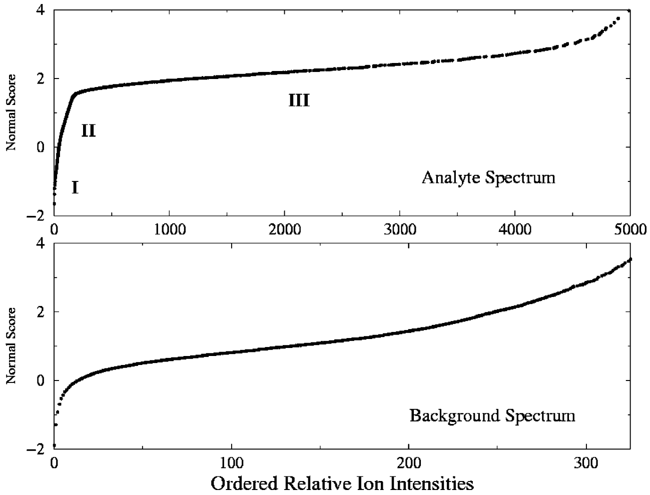

Figure 4. Normal probability plot for the MALDI-TOF mass spectrum shown in Figure 3 revealing the complex shape indicative of several sources of noise in the analyte spectrum and in the congruent analyte-free background spectrum. Roman numerals in the analyte spectrum refer to three distinct linear regions likely stemming from three sources of noise.

frequency range that defines the peak shape for each oligomer as well as the lower frequency range dictated by the time spacing between oligomer peaks. The characteristic shape of the normal probability plot suggests that the noise appears to arise from multiple sources and that different sources of noise appear to interfere with each other. Furthermore, our experience shows that the normal probability plots have subtle variations in shape as changes to instrument parameters, matrixes, and analytes are made. Last, the normal probability plot of the background spectrum has a quite different shape from the normal probability plot of the analyte spectrum. This makes the a priori prediction of the noise spectrum from instrumental parameters, or from a background spectrum, very difficult.

Without adequate models of the sources of noise, smoothing of raw data by *any* algorithm could alter peak structure. Thus, smoothing could remove peaks that are present or create spurious peaks from noise. Most importantly for quantitation, smoothing will change relative peak area. All of these limit the accuracy of any quantitative reduction of the data. Our experience shows that the power spectrum of the noise cannot be predicted solely from the experimental conditions; therefore, blind application of smoothing or filtering algorithms may unintentionally remove information necessary for quantitation of instrument response from the data. This caution against blind application of smoothing routines has recently been discussed by Eilers in this journal.[19] An innovative method to identify peaks while avoiding smoothing of the data has just been published.[20] This method entails using local histograms over variable window widths to isolate regions of the spectrum that deviate from the baseline.

RESULTS

Noise Added Mathematically. As a first demonstration, let us return to the example shown in Figure 2. In the upper left panel, the exact areas of the three peaks are in the ratio 1:1.75:1. The method finds a ratio of 1.02:1.79:1.03, or stated another way, a value no more than 3% off the exact value. This example is rather easy in that the noise is of a much higher frequency than the signal so separating them by any method would not be too difficult. More astute models of noise are required for a more challenging test of the method.

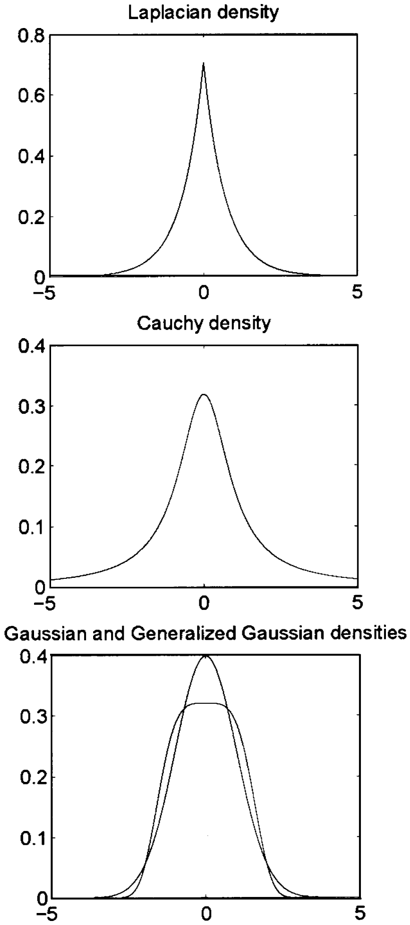

Figure 5. Graphical representation of the model noise PDF used. The Laplacian PDF was used to model noise on the time axis, and the Cauchy and generalized Gaussian PDFs were used to model noise on the intensity axis.

To further test the robustness of the method, we applied the algorithm to data sets corrupted by several types of simulated noise (error). The first seeks to simulate errors made in the time measurement (e.g., digitizer jitter). While it is our opinion that this error will be small when compared to other instrument errors, we include it here for completeness. This is well modeled by a Bernoulli process using a Laplacian noise model probability density function (PDF) as shown in Figure 5. If t_i^r denotes the ith component of the true time series, then the observed t_i^0 can be modeled as

$$t_i^0 = t_i^r\left(1 \pm \epsilon_1 \frac{\exp(-|t|/(\sigma/\sqrt{2}))}{\sqrt{2}\sigma}\right) \qquad (1)$$

where ϵ_1 defines the magnitude of the error and σ is a constant known as the Bernoulli probability parameter.[21]

Second, we sought to model the combined type A (“random”) and type B (“systematic”) measurement uncertainties encountered

[19] Eilers, P. H. C. *Anal. Chem.* **2003**, *75*, 3299.

[20] Jarman, K. H.; Daly, D. S.; Anderson, K. K.; Wahl, K. L. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 61.

Table 1. Performance of the Method with Variable Amounts of Numerically Simulated Noise Using the Cauchy Probability Distribution Function (Eq 2) To Define the Error in Intensity

p	ϵ_1	ϵ_2	no. of peaks	M_n (μs)	M_w (μs)
0.01	0.1	0.1	176	99.24	99.69
0.05	0.2	1	180	99.17	99.83
0.1	0.3	10	278	93.08	141.4
0.25	0.5	100	1912	298.6	359.2

Table 2. Performance of the Method with Variable Amounts of Numerically Simulated Noise Using the Generalized Gaussian Probability Distribution Function (Eq 2) To Define the Error in Intensity

p	ϵ_1	ϵ_2	no. of peaks	M_n (μs)	M_w (μs)
0.01	0.1	0.1	176	99.24	99.73
0.05	0.2	1	177	99.28	99.77
0.1	0.3	10	225	99.97	103.2
0.25	0.5	100	1078	238.1	245.3

in the intensity coordinate of the data. Type A arises from such things as poor counting statistics for small peaks while type B arises from such things as mass-biased ionization probabilities or mass-biased detector sensitivity. Truncation and round-off errors introduced by the data system are also included in this model; however, they are generally not as significant as the other errors. The errors in the intensity coordinate are expected to play the major role in the final determination of molecular mass distribution. To model this more detrimental noise, we added varying amounts of noise selected from one of the following noise distributions to the intensity measurement, I_i^r, at randomly selected points in the spectrum shown in Figure 3 such that the selected data points suffer some nonzero error. We used two distributions the Cauchy

$$I_i^0 = I_i^r\left(1 \pm \epsilon_2 \frac{1}{\pi\sigma(1 + (t/\sigma)^2)}\right) \tag{2}$$

where ϵ_2 defines the magnitude of the error and σ is the Cauchy probability distribution constant, and the (generalized) Gaussian

$$I_i^0 = I_i^r\left(1 \pm \epsilon_2 \frac{1}{2\Gamma(5/4)A} \exp\left(-\frac{t^4}{A^4}\right)\right) \tag{3}$$

where ϵ_2 defines the magnitude of the error and Γ and A are the constants for the generalized Gaussian.[21] See Figure 5 for a graphical representation of the shape of these probability density functions. Note that the Cauchy PDF has a greater relative width than the generalized Gaussian. Given these three models of uncertainty, the robustness of the algorithm can be tested. We choose only to operate on only some points of the spectrum as a means to control the level of added mathematical noise; other means could of course be chosen.

In Tables 1 and 2 we report on the behavior of the algorithm on data sets with varying degrees of simulated noise corruption for the Cauchy and the generalized Gaussian PDFs. In the tables, the parameter p gives the fraction of points in the spectrum corrupted by noise; that is, $p = 0.01$ indicates 1% of the data points have been effected. Of those points, each is subjected to an error in t defined by eq 1 and an error in I defined by either the Cauchy (Table 1) or the generalized Gaussian (Table 2) PDF. The algorithm performs similarly in the face of large and small amounts of noise, up to a critical point. As noise was increased, the number of strategic points identified increased representing the creation of "false" peaks through the random correlation of noise fluctuations. At the critical point, the signal-to-noise ratio is such that it is impossible to extract meaningful quantitative measures. At reasonably small changes to the signal-to-noise ratio, the algorithm finds only a small change to the moments of the molecular mass distribution (specifically M_n the number-average molecular mass and M_w the mass-average molecular mass as defined in eqs 4 and 5).

$$M_n = \frac{\sum N_i M_i}{\sum N_i} \tag{4}$$

$$M_w = \frac{\sum N_i M_i^2}{\sum N_i M_i} \tag{5}$$

where N_i is the number of molecules of mass M_i.

Noise Added Experimentally. To degrade the polystyrene MALDI-TOF mass spectrum signal-to-noise ratio experimentally, the optimal laser intensity of 0.2 μJ/pulse at the sample was first decreased to 0.1 μJ/pulse and then increased to 0.33 μJ/pulse and to 0.5 μJ/pulse. (All values are ±0.01 μJ/pulse, which is the standard deviation of 100 pulses and is taken as an estimate of the standard uncertainty.) Increasing the laser energy creates more "chemical" noise in the spectrum. For the 0.33 μJ/pulse experiment, this noise is concentrated near the main peaks at the center of the spectrum and is attributed to ions arriving at the improper time for their corresponding mass. This can result from excessive initial velocity imparted by the ablation process or from metastable ion fragmentation in the flight tube or in the ion mirror of the mass spectrometer. For the 0.5 μJ/pulse experiment, this "noise" appears at low mass due to silver clusters. Changing the voltage on the channel plate detector also degraded the spectrum. The voltage to produce the best signal-to-noise ratio was 1.5 kV. The voltage was decreased to 1.4 kV, making the detector quiet but not very sensitive, and increased to 1.6 kV, making a sensitive but not very quiet detector. (All voltages are ±0.01 kV, which is taken from the instrument's voltage monitoring hardware.) Raising the detector voltage produced increased "electronic" noise, that is, high-frequency noise across the entire spectrum.

Figure 6 shows the original spectrum optimized for signal-to-noise ratio (from Figure 3) as well as one laser power setting below and two above the optimum value. The signal-to-noise ratio has been lowered especially for the small peaks found at the wings of the molecular mass distribution. In addition, silver cluster peaks (Ag_n^+) with the characteristic alternating peak height structure

[21] Hogg, R. V.; Craig, A. T. *Introduction to mathematical statistics*; MacMillan: New York, 1978.

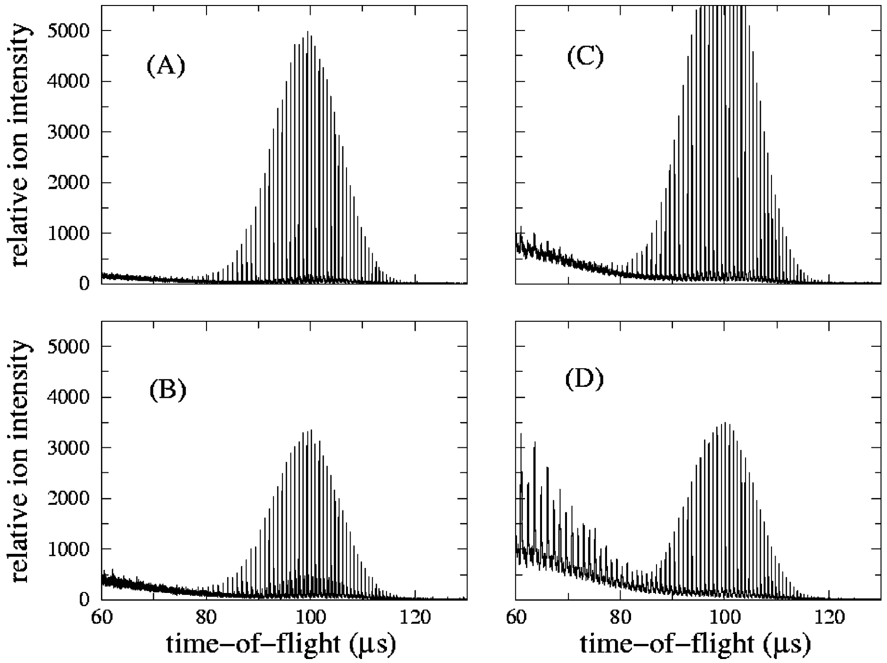

Figure 6. MALDI-TOF mass spectrum of the low-mass, narrow-polydispersity polystyrene taken at laser powers for (A) optimal signal-to-noise ratio (same data as in Figure 3), as well as (B) one laser power setting below and (C, D) two above the optimum value. (All figures are plotted on the same scale; tick labels carry over to unlabeled tick marks in adjacent plots.)

Table 3. Performance of the Method with Variable Amounts of Experimentally Added Noise

noise source	no. of peaks	M_n (μs)	M_w (μs)
optimal settings (0.2 μJ/pulse) (1.5 kV detector voltage)	176	99.2547	99.7114
low laser power (0.1 μJ/pulse)	228	98.4099	99.2461
high laser power (0.33 μJ/pulse)	236	97.2339	98.3408
high laser power (0.5 μJ/pulse)	94	88.0035	90.9924
low detector voltage (1.4 kV)	92	99.2170	99.5802
high detector voltage (1.6 kV)	128	99.1670	99.7521

(due to the even-*n* clusters being more stable than the odd-*n* clusters) are observed in Figure 6D for the highest laser power. Last, the baseline is much higher at lower mass in the two spectra created by the excess laser power. This is due to the abundance of clusters (matrix, silver, and matrix + silver) and, to a lesser extent, metastable polystyrene fragments, each of which is unfocused in time. These effects will alter the measured molecular mass distribution; however, what we wished to show is that the method presented here will find the statistically significant peaks in a variety of situations. Table 3 gives the number of peaks and the mass distribution moments M_n and M_w for the spectra studied. Notice that for the highest laser power data the mass distribution moments are very low due to the inclusion of the silver cluster peaks in the calculation. At this point, the operator must intervene to sort out the polymer peaks from the spurious silver cluster peaks. The number of peaks is also lower at the highest laser power because the small interstitial peaks between the main series peaks have been lost in the noise. Nevertheless, the method performed its function by identifying and calculating the area of the peaks in the spectrum and returning results that would be expected from the experimental conditions.

Changing the detector voltage (and thus the detector sensitivity) played a much smaller role in altering the molecular mass moments, as seen in Table 3. The number of peaks found drops in both cases. At the lower voltage, the smaller peaks are not registered, while at the higher voltage, they may be registered but are lost in the noise. In either case, the method either fails to find a peak or deems it to be not statistically significant.

Last, the original data set and its congruent analyte-free spectrum were smoothed using a 7-point moving average. While the original spectrum had 175 peaks identified by the algorithm, the smoothed spectrum has only 160. Thus, 15 peaks were smoothed to below the statistical lower limit. This was true even though the background spectrum was smoothed in the same manner. Shown in Figure 7 are the areas of these two peaks sets in descending order. Note that the ion intensities are now plotted on a logarithmic scale. Also shown is the ratio of peak areas (original spectrum divided by smoothed spectrum) on a peak-by-peak basis across both sets. While the area of the large peaks was unaffected by the smoothing, the smaller peaks were significantly reduced in area, some to only one-third of their original value. Recognize that while the total integrated counts in the original spectrum are not changed by the running average (exclusive of small end-point effects), the total peak area found by the algorithm decreased by 1.7% due to smoothing both through the loss of the 15 smallest peaks and decreased peak area of the smaller peaks. Clearly smoothing can lead to the loss of information and can pose a significant bar to quantitation when dealing with oligomers found in small quantities in the analyte.

Closer observation of the details of the peaks identified in the two spectra (original and smoothed) is shown in Figures 8 and 9. Note again that the ion intensities are plotted on a logarithmic

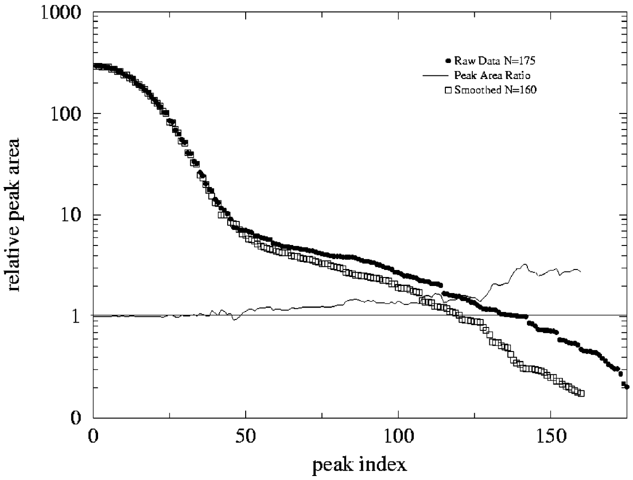

Figure 7. Relative peak areas sorted by size for the data of Figure 3 before (solid circles) and after (open squares) application of a 7-point moving average. Solid line is the ratio of relative peak areas before and after smoothing showing how ion intensity has been effectively smoothed away.

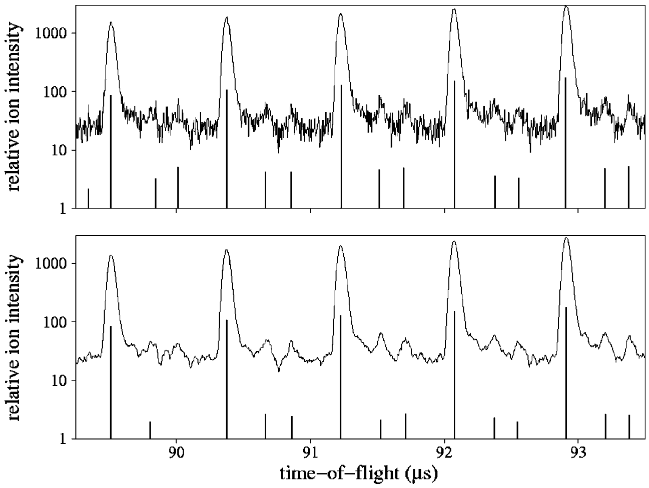

Figure 8. Direct comparison of the data of Figure 3 with the 7-point moving averaged data showing the peak positions and relative peak areas given by the vertical impulses originating on the time axis. In this part of the spectrum, peaks are lost from smoothing.

scale in order to observe the smallest peaks. Peak areas are clearly changed by smoothing. For example, in Figure 8, the small peaks in the smoothed spectrum have consistently less area (given by the magnitude of the vertical lines) than in the raw spectrum. Furthermore, notice that smoothing can lose peaks, for example, the peak at 90 μs. In Figure 9, which displays a low signal intensity region at the beginning of the spectrum, it can be observed that smoothing, specifically at 81.7 μs and at 83.8 μs, creates peaks. An analyst might say that those peaks were there all along and that the 7-point average only brought them out. This notion can be supported by the periodicity of the data, that is, in parts of the spectrum where the signal-to-noise ratio is better, peaks are found at these positions relative to the main series.[4] However, from a purely statistical point of view, these peaks did not have enough local convexity or area in the raw spectrum to be selected by the method. Lack of local convexity (“peakiness”) would prevent the method from finding any strategic points in the segmentation algorithm. A small relative peak area would cause them to be discarded in the deflation algorithm of the method.

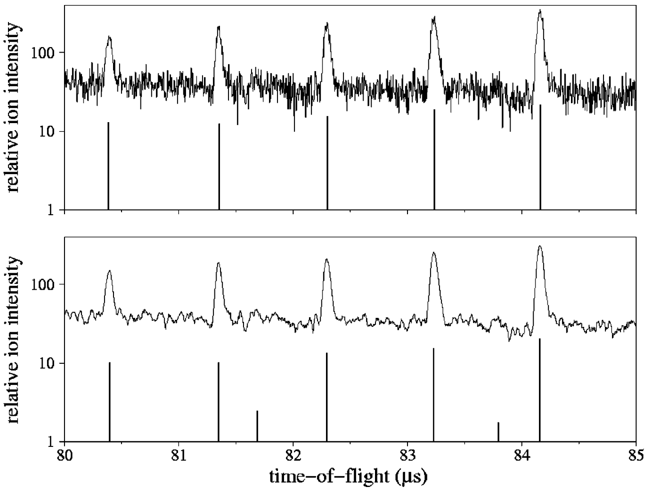

Figure 9. Direct comparison of the data of Figure 3 with the 7-point moving averaged data showing the peak positions and relative peak areas given by the vertical impulses originating on the time axis. In this part of the spectrum, peaks are gained from smoothing.

Last, peak positions are slightly modified by smoothing; however, the magnitude of this change is far less significant than the ion intensity axis distortion. In general, the main series peaks were shifted to longer flight times and the smaller intermediate peaks were shifted to shorter flight times. However, the time shifts were only on the order of nanoseconds, which will not affect the determination of molecular mass distribution, but will greatly affect instrument calibration if a smoothed spectrum is used as a calibration reference.

CONCLUSION

A unified collection of algorithms was presented that accurately locates peaks and calculates their area using only reproducible mathematical operations and no user-selected parameters. The method works best with an analyte spectrum and its congruent analyte-free spectrum to build a model of the measurement noise. Examples were given of mass spectra corrupted by purely mathematical means and by instrumental means. The addition of mathematically derived noise demonstrated the stability of the algorithm in that even with the addition of large amounts of noise on either the time or the ion intensity axis did not prevent the method from finding a reasonable selection of peaks. The addition of experimentally derived noise demonstrated the marginal nature of small peaks that can easily be lost. It also showed how unintended peaks, in this case of silver clusters, can have a profound effect of molecular mass moments when unbiased analytical methods are used. Last, smoothing by a 7-point moving average was shown to change peak area as well as remove and create peaks in an unpredictable manner.

ACKNOWLEDGMENT

The authors thank Javier Bernal of the NIST Mathematical and Computational Sciences Division for assistance in testing the computer code, Kathleen M. Flynn of the NIST Polymers Division for taking the MALDI-TOF mass spectra of SRM 2888, and Andrew R. Roosen of the NIST Center for Theoretical and Computational Materials Science for creating the on-line public version of the method.

Received for review December 12, 2003. Accepted March 2, 2004.

AC0354701