Improved Centroid Peak Detection and Mass Accuracy using a Novel, Fast Data Reconstruction Method

James A. Ferguson [1], William G. Sawyers [1], Keith A. Waddell [1],
Anthony G. Ferrige [2], Robert Alecio [2], Stuart Ray [2]

[1] Applied Biosystems, Framingham, MA, [2] Positive Probability Ltd, Isleham, U.K.

Overview

An advanced data reconstruction method is described to determine peak centroids. Comparisons are made with proprietary peak detection methods. The advantages of this new method are:

1. More accurate mass centroids on small noisy peaks.
2. Peak information independent of noise level.
3. Results not dependent on threshold values.

Introduction

All data are corrupted by noise and the error in centroids using algebraic algorithms is related to abnormalities in the noise over the peaks. Algebraic calculations are very fast and they always work directly on the data. However, they do not take into account a varying noise level (leading to both over and under estimation of noise levels within the data set) and one cannot assess the error on reported centroids as a measure of assessing the validity of a peak. An alternative is to use a data reconstruction method. An advantage being that it uses a peak model to iteratively reconstruct the data until the reconstruction fits the data within the noise level. The data are separated into signal and noise channels, the efficiency of the separation being primarily determined by the peak model quality. Because the model is wider than high frequency noise, individual noise spikes do not fit the model and their influence is small and spikes that can distort the position of an algebraic centroid are less likely to distort the data reconstruction. This is best illustrated in Figure 1. Unfortunately, probability-based methods are slow because they are iterative. We describe here a reconstruction method for centroiding with only one input that is very fast, accounts for any variation in noise level and has the benefits of improved mass accuracy and quantified errors.

Method

Probability-based data reconstruction methods are only applicable to baseline corrected data since any positive baseline intensity will be interpreted as signal. They are also designed to reach a predetermined point as with the ***ReSpect***™ and maximum entropy software. In these solutions the reconstruction fits the data within the noise level and far exceed that required to efficiently separate signals from noise. An enhanced method was used to perform this separation in a single, fast computation without broadening peaks. The reconstruction has enhanced S/N and peak positions/intensities are obtained along with their quantified errors. The errors calculation is the most computationally intensive step and the speed is comparable to conventional centroiding if this is omitted. A key, novel feature of the method is that it takes into account any variation in the noise level and peaks are assigned a significance level that is independent of the noise. Therefore the setting of a conventional threshold is not required. In addition, the method can accommodate any positive error in the baseline corrected result since this gives weak peaks and noise features an enhanced significance. This program function examines the results and determines whether there is a bias in the centroid intensities so that noise may be rejected according to its significance, again, taking into account any variation in the noise level.

Various MALDI MS and Electrospray data were processed via either Data Explorer® software (AB1) or Analyst® software (AB2) and compared with the new peak detection (PPL method). The Analyst software and Data Explorer software methods used centroiding of the top 50% of the peak, other parameters were optimized for the analysis concerned.

Data Explorer software methods employed thresholds based on % maximum peak area (% MPA).

Results

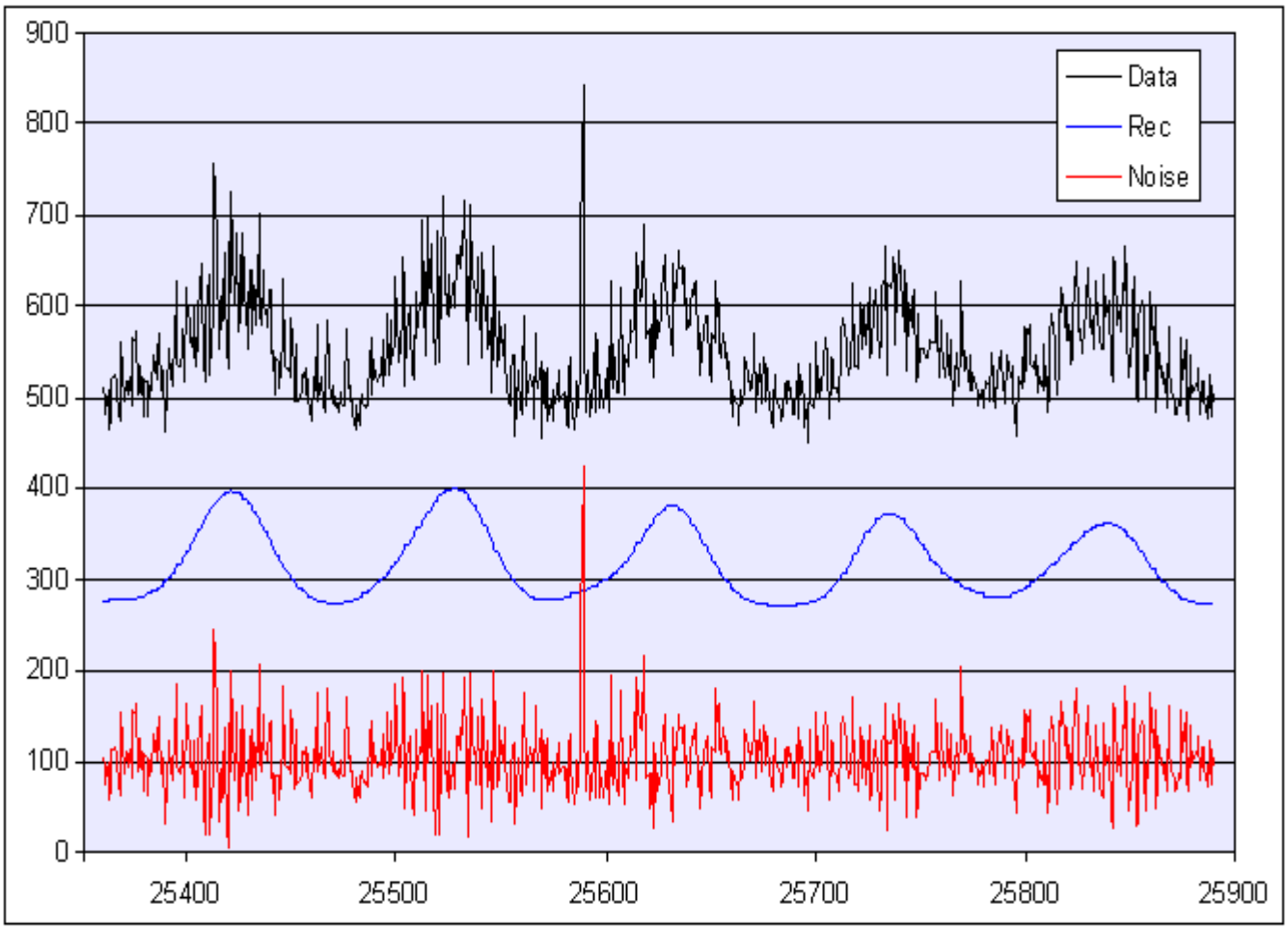


Figure 1: Principle of Reconstruction Process

Top: Data: **Middle**: Reconstructed data: **Bottom**: Data minus reconstruction, representing the noise.

The points to note are:

1. The intense spike does not fit the model. It is treated as noise and has little effect on the reconstruction.
2. The reconstruction is free of all high frequency noise and the peaks are not broadened.
3. The spike is detected as noise (lower trace) and that the noise over the peaks is evenly distributed about the mean, ensuring little effect on the reconstruction

Example 1

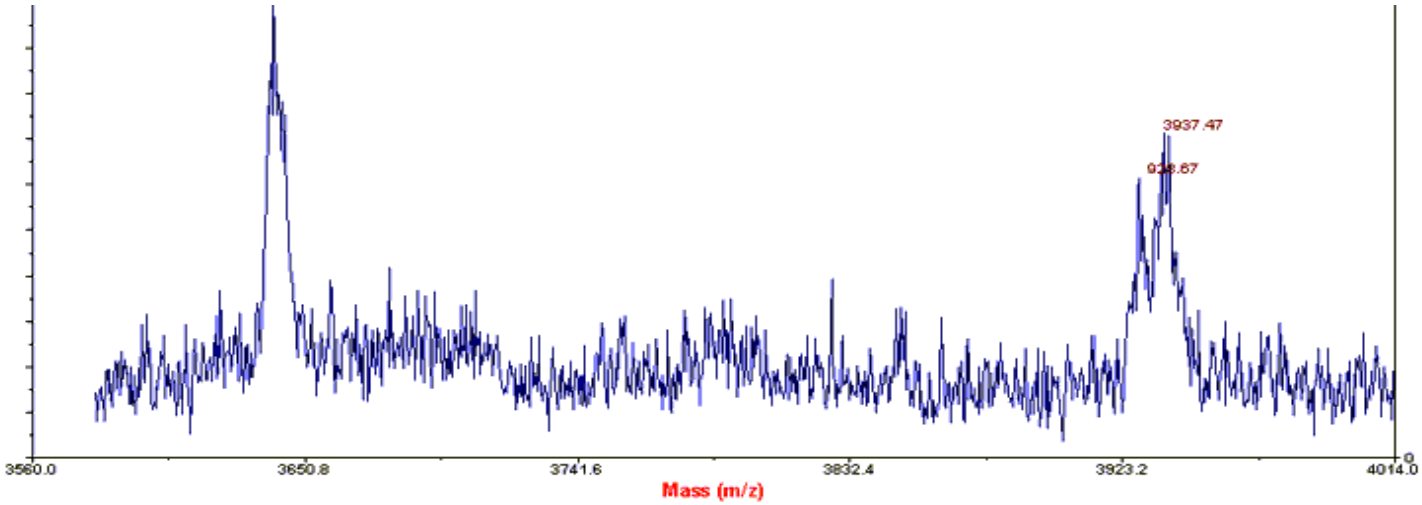


Figure 2 MALDI TOF mass spectrum of an oligonucleotide with A or T base extension. Theoretical mass differences are 288.2 Da for T and 297.21 Da for A for the doublet from the parent oligo.

Table 1. Oligo and AT base extension mass differences using different centroiding algorithms.

	Theory	PPL	AB1 - Raw	AB1 - NF 0.7	AB1 - NR1	AB1 SM5
T	288.20	287.57	287.42	288.04	287.86	287.98
A	297.21	296.68	296.22	296.15	295.87	295.99
Difference	9.01	9.11	8.80	8.11	8.01	8.01

The Data Explorer software (AB1) centroiding algorithm was used on both raw and preprocessed data. NF0.7 was noise filtered, NR1 was noise reduced and SM5 had a 5 pt smooth applied. "PPL" was the new algorithm. The results show as one smoothes or noise-processes the data, the 2 peaks in the doublet merge together, giving a larger mass difference error. The PPL algorithm compares very favorably with the established method and gives half the mass error.

Example 2

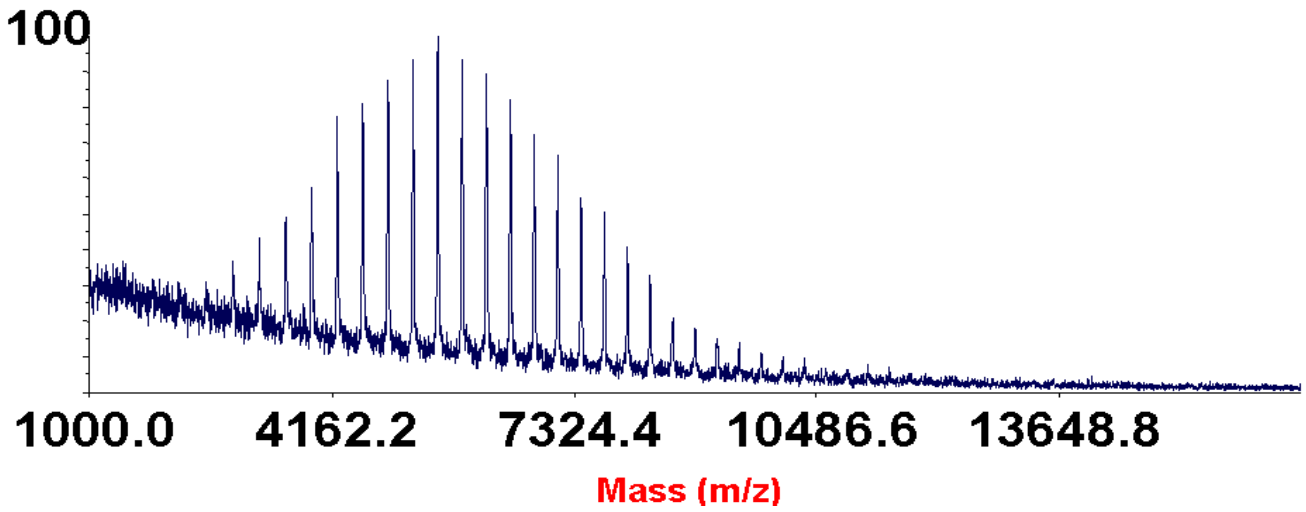


Figure 3a: MALDI MS of a polymer.
Note that the noise level is considerably higher at low mass than at high mass.

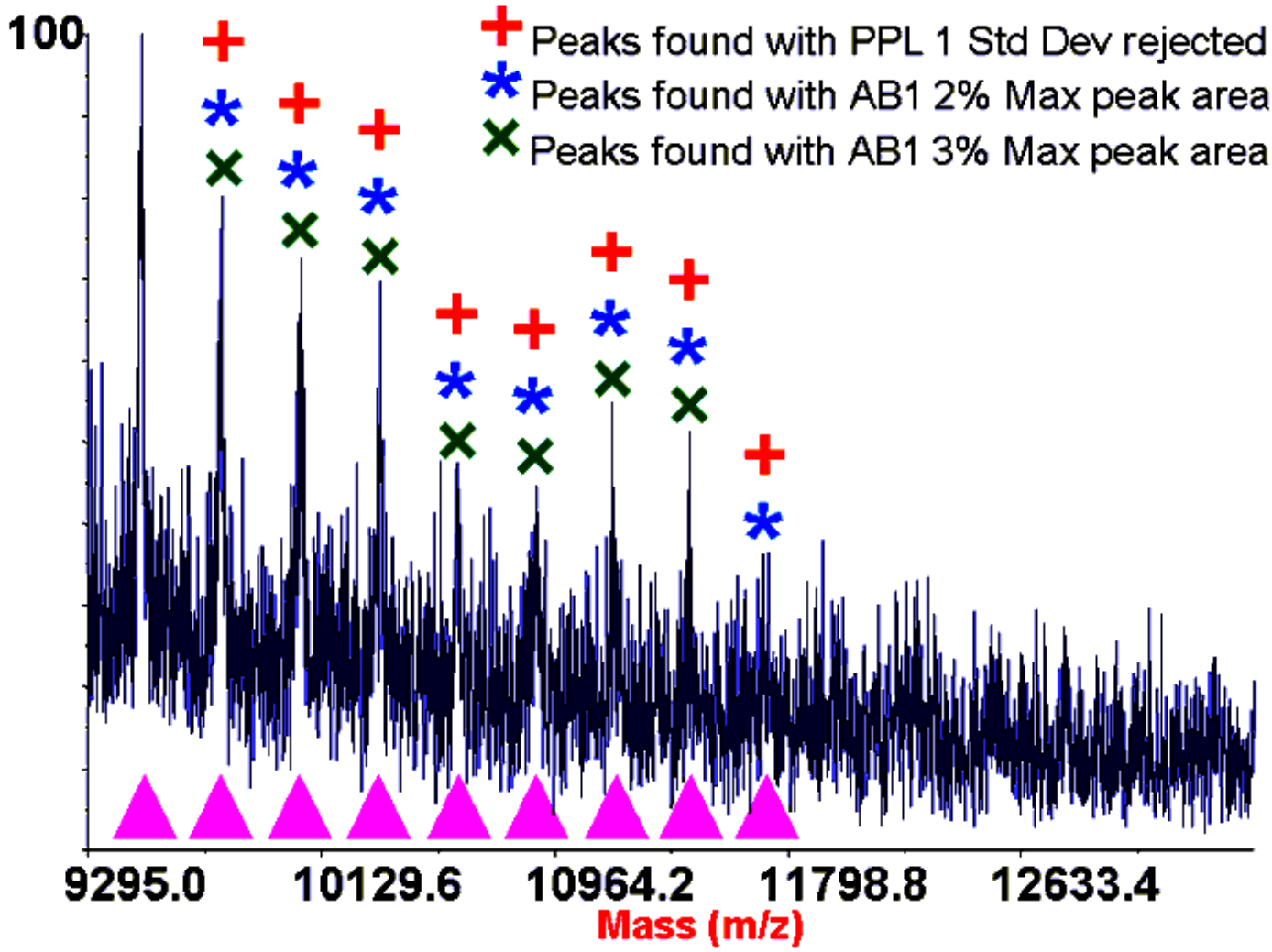


Figure 3b: Horizontal expansion showing found peaks. The polymer spectrum has a sequence of peaks which decrease in intensity at higher mass.

Table 2. Peaks between masses 1,000 and 16,811.

Software	Criterion	Peaks
AB 1	0% MPA	496
	1% MPA	452
	2% MPA	310
	3% MPA	170
PPL	All	474
	Minimum	292
	Normal	159

Although both the AB1 peak detection (2% max peak area) and the PPL program (normal) found as many of the polymer series as possible, the setting required on AB1 reported nearly twice as many peaks in the whole spectrum.

Example 3

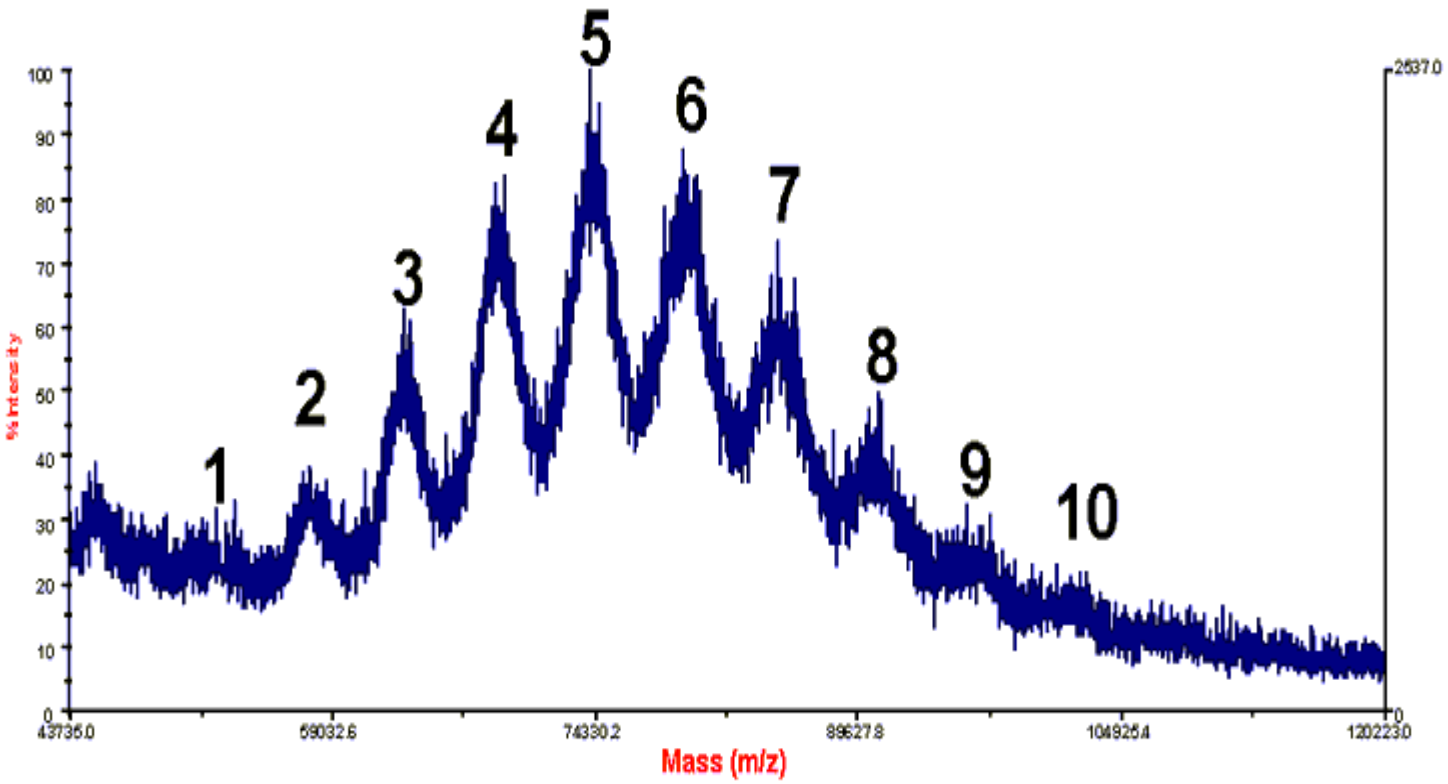


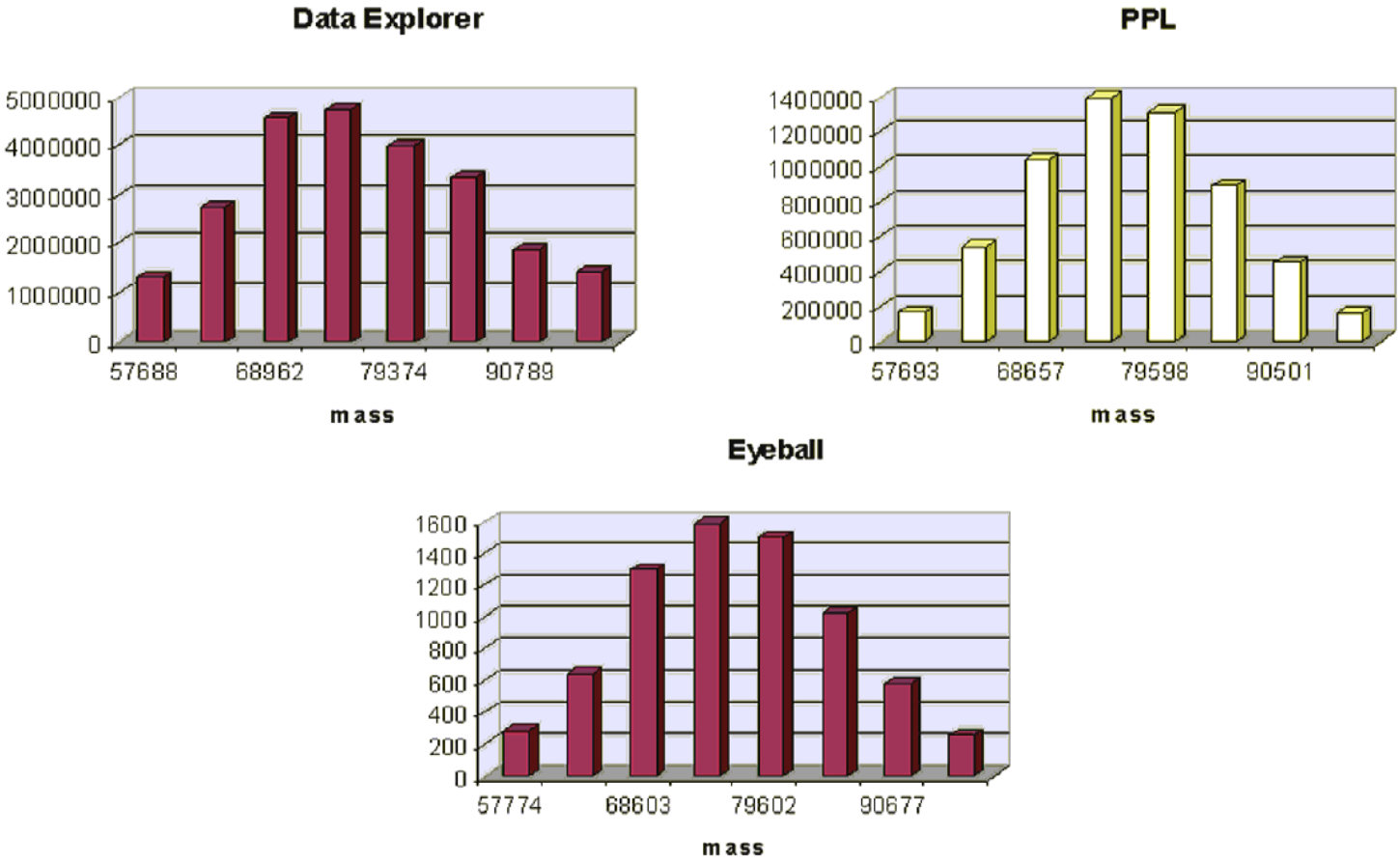
Figure 4. High mass polymer sample analyzed by MALDI TOF MS. Each broad peak contains about 900 data points. The interval between each peak should theoretically be the same.

Table 4. Comparison of the three methods including a pencil and ruler ("eyeball") method for this high mass polymer.

Peak	AB1	PPL	"Eyeball"
1	52235	52066	51987
2	57688	57693	57774
3	63075	63179	63312
4	68962	68657	68603
5	74044	74130	74102
6	79374	79598	79602
7	84831	85041	85101
8	90789	90501	90677
9	95932	96167	96247
10	Split peak	101871	101960
Mean interval	5462	5534	5533
Std deviation	372	101	118

The AB1 software gave 2 peaks at 99,000 and 102,000, so this peak was deleted from the mean calculation. The PPL std. dev. reduces to 84 using the same number of peaks. The AB1 peak detection method also gave 3 other false peaks within this mass range. The AB1 algorithm performed better when the raw data was smoothed (std. dev. of 186), however the data integrity is then lost.

Figure 5 (below). Comparison of data intensity (area) across major peaks.



Results show that the relative intensities of the minor peaks are over estimated with AB1 method. This may produce incorrect centroids. Compare with eyeball method (Intensity in this case was height.)

Example 4.

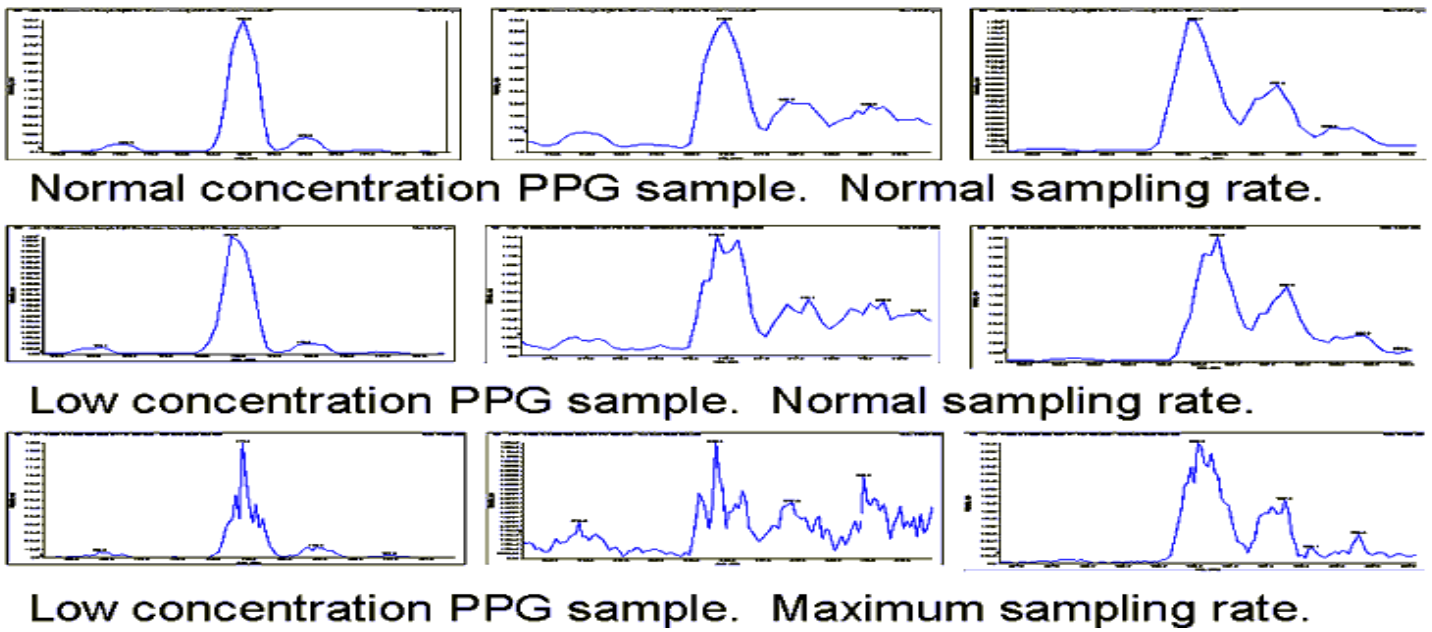


Figure 6. PPGs data from triple quad using ESI.

Peaks selected were PPG calibrant peaks plus their isotope peaks and others in the discrete ranges chosen. The masses were m/z: 57, 59, 60, 173, 175, 176, 616, 617, 618, 906, 907, and 908. Table 5, in the next column summarizes the data.

Table 5. Comparison of the two centroiding methods

Difference	AB2 processed		PPL processed	
	#1	#2	#1	#2
Mean	-0.0002	0.0334	0.0075	0.0277
Std dev.	0.0774	0.0951	0.0665	0.0341

Difference #1 is the mass difference between the normal sampling rate high and low concentration samples. Difference #2 is between the high conc. normal sampling rate and low conc. high sampling rate samples.

Example 5.

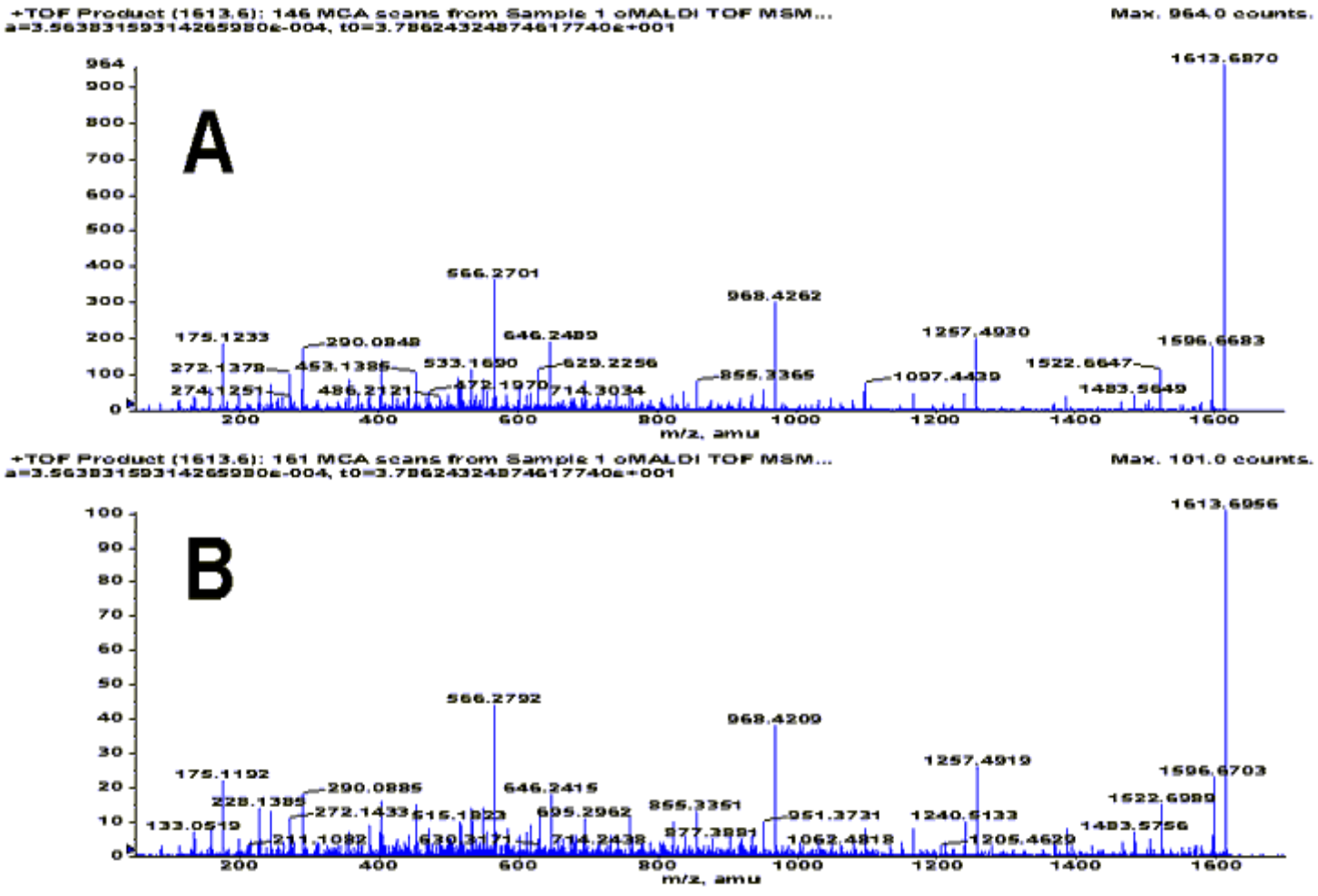


Figure 7. QqTOF MS/MS spectra from a peptide A) concentrated sample, B) diluted sample. Identical conditions, and calibration.

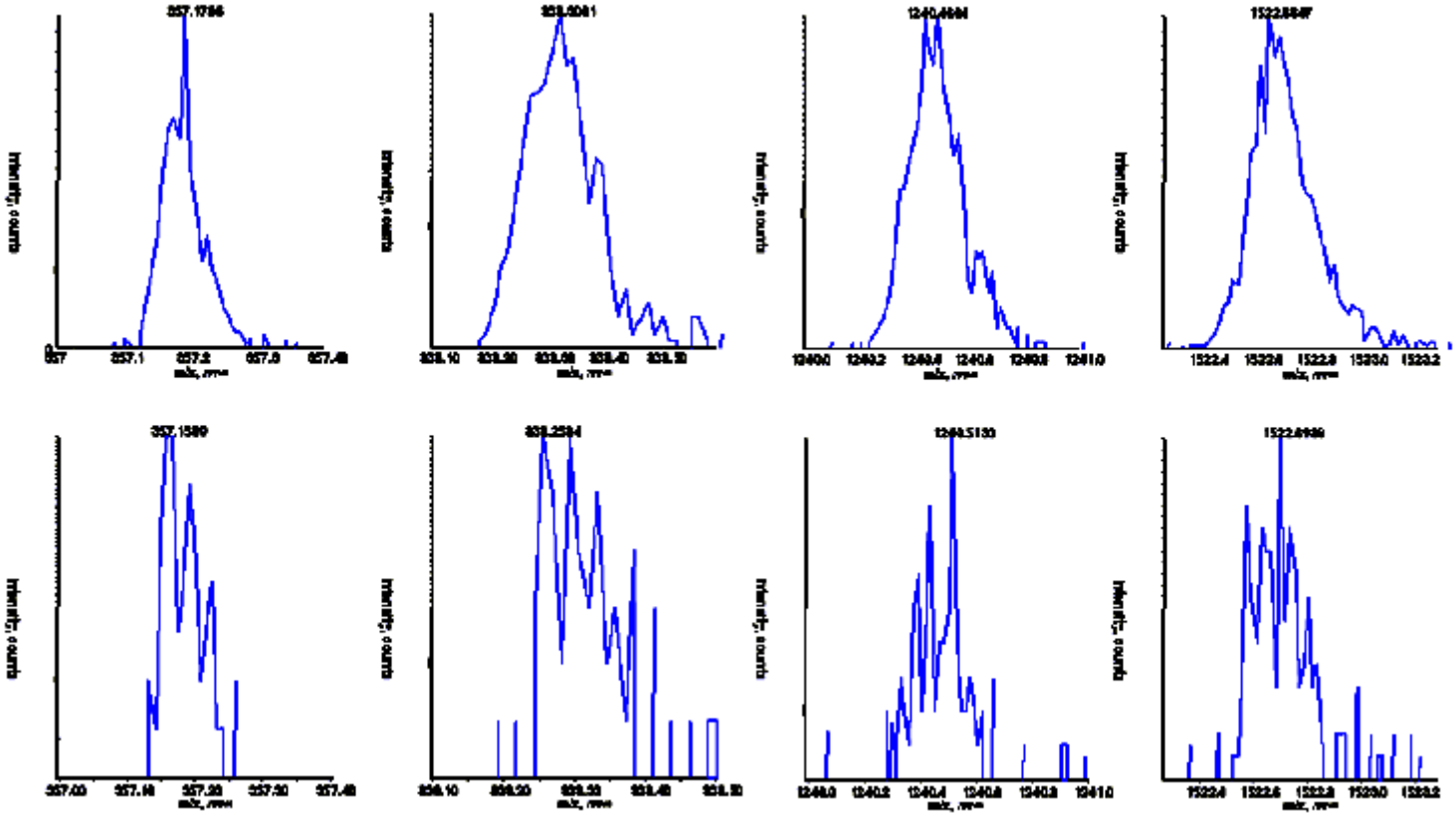


Figure 8. Partial MS/MS regions from high signal sample and low signal sample. The lower traces illustrate the challenge involved with centroiding "spiky" data.

Table 6. Mean and standard deviation data from the difference between strong and weak MS/MS spectral peaks.

		Mass Difference (Da)	Mass Difference (ppm)
PPL	Mean	0.001503	2.1
	Std dev.	0.009927	17.3
	Spread	0.0443	100.3
AB2	Mean	-0.000436	1.8
	Std dev.	0.030351	39.4
	Spread	0.2077	236.5

***Note** that the standard Analyst software (AB2) algorithm gave substantially (2×) better values after the data were smoothed. The most intense and common peaks were taken for the above results.*

Results Summary

Variable Noise Level - It has been demonstrated that a properly designed data reconstruction method will allow genuine signals to be identified regardless of the way the noise level changes in the data. The method also allows thresholds to be set according to the significance of peaks so that weak peaks are not "missed" in regions of low noise.

Centroid Errors - Table 7 shows the size of the standard deviations results from the different methods.

Table7. Comparison of the errors for the different methods (see individual results panels)

Spectrum	ABI	PPL	ABI / PPL
Peptide MS/MS	0.0242	0.0071	3.41
Oligonucleotide MALDI TOF	0.2	0.1	2.00
PPG analysis #1	0.0774	0.0665	1.16
PPG analysis #2	0.0951	0.0341	2.79
High mass polymer	372	84	4.43

ABI refers to the peaks determined by the Analyst software or Data Explorer software.

Discussion

Algebraic centroiding methods typically perform local baseline corrections in an attempt to increase the reliability of reported centroids. The user is required to inspect the data and select appropriate options if the centroiding is to be successful. Data are also normally smoothed before centroiding to reduce the effect of noise. This is only effective because, for example, the top 50% of a broadened signal contains more noise but the noise statistics are more uniform. Serious errors are introduced when filtering or smoothing as overlapped peaks begin to merge. In extreme situations, peaks resolved in the raw data are no longer resolved after filtering.

In order to reduce the peak table to a convenient size it is necessary to apply some form of threshold, usually a user-selected percentage of the area of the strongest peak. However, in this approach, the number of peaks reported will depend on both the intensity of the strongest peak and the magnitude of the noise. Unless noise level variations are taken into account, the final table may contain excessive noise centroids where the noise level is high and will miss genuine weak peaks in regions of low noise.

The data reconstruction method requires little user input - only an estimate of the peak width. First, noise variation is accounted for so that the computed baseline is at the noise center regardless of its amplitude. Second, the estimated peak width is used as a model for a fast data reconstruction incorporating the noise vector so that peaks may be identified by their S/N. Quantified errors are available if required. Finally, noise features may be automatically rejected according to their significance level and/or S/N.

Conclusions

In the work presented here, it has been demonstrated that:

1. Using a baseline correction method and correctly taking into account any varying noise level, the application of subjective conventional thresholds is redundant.
2. The fast data reconstruction method provides substantially improved mass accuracy over conventional algebraic centroiding, particularly for broad, noisy data.

Acknowledgements

Cheni Krishnan, Melanie Lin, Marge Minkoff, and Song Ye of Applied Biosystems for supplying some of the data presented here.