Poster Reprint

ASMS 2018
TP-292

Improved Peak Detection for Mass Spectrometry via Augmented Dominant Peak Removal

Daniel Abramovitch

Agilent Technologies, Santa Clara, CA

Introduction

Peak assignment/picking/finding is a fundamental post measurement data compression step.

Basic peak finding algorithms are based on locating abundance profile apices (or zero crossings of abundance slope), characterizing the peak width, and returning a centroid for each peak [1] (Fig. 1, left). For isolated peaks with low noise, this works quite well.

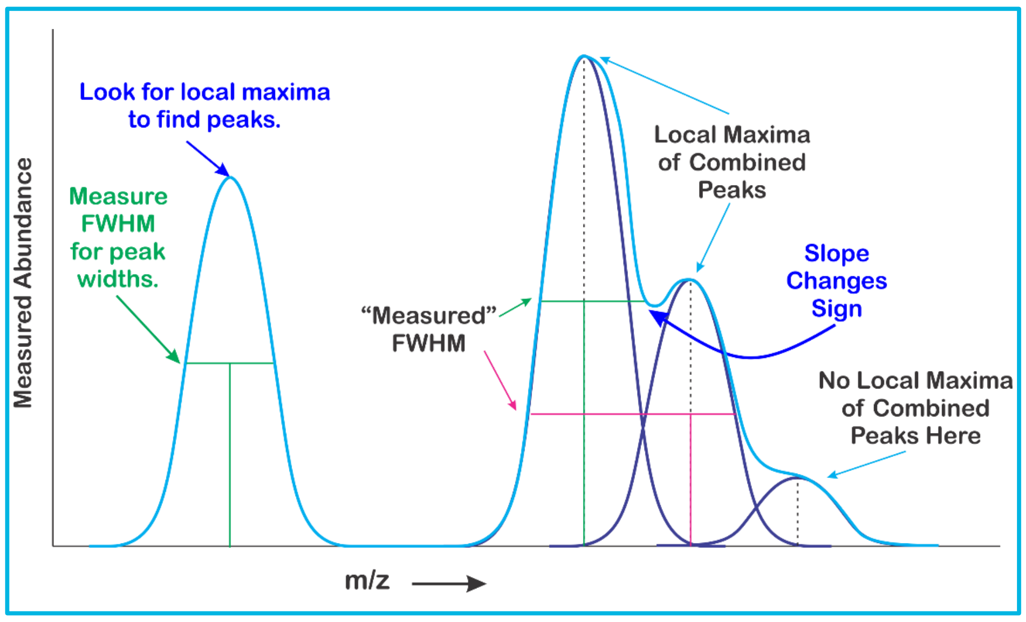
Look for local maxima to find peaks.
Measure FWHM for peak widths.
Measured Abundance
"Measured" FWHM
Local Maxima of Combined Peaks
Slope Changes Sign
No Local Maxima of Combined Peaks Here
m/z

Figure 1: Basic peak picking on sample mass axis (left), and issues caused by interfering peaks (right).

Overlap with side peaks interferes with peak detection and centroid calculations.

Interfering peaks can increase errors in peak width and centroid calculations and make some peak apices invisible under the curve (Fig. 1, right). There is much discussion of baseline correction and noise filtering, but not much is said about interfering peaks. Least squares fits (e.g. Levenberg-Marquardt) have computational complexity that goes up as the cube of the number of peaks. For densely populated mass axes, this makes the problem intractable except in post processing

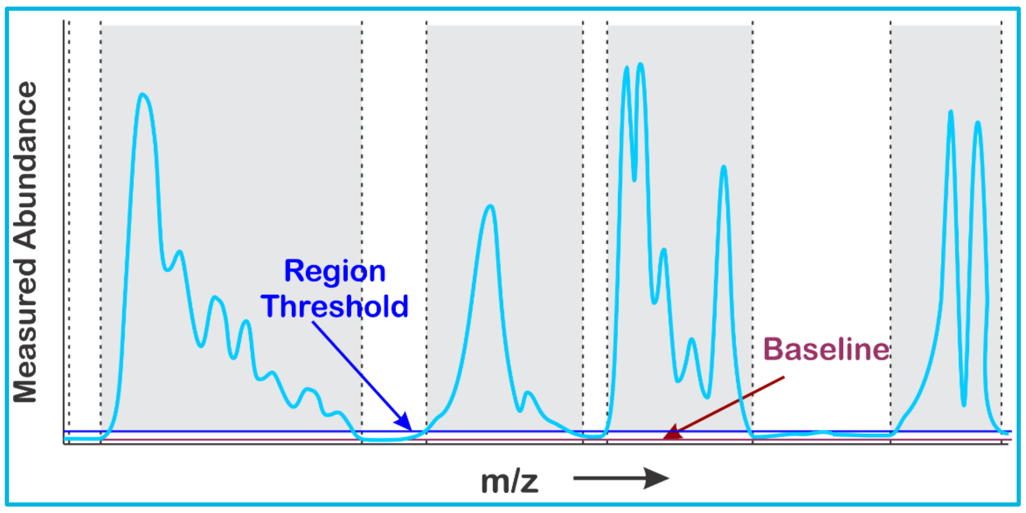

Figure 2: Segmenting interesting regions for iteration.

Segment the Mass Axis

Fix #1: Segment the mass axis into smaller regions.

In many measurements, there are regions of interesting mass abundance and those that are not. By dividing interesting regions (Fig. 2), and doing separate peak detection in each interesting region, the search complexity can be dramatically reduced.

Successive Dominant Peak Removal

Fix #2: Fit largest remaining peak in segment to model. Remove model abundance from abundance curve. Repeat on residual abundance curve.

Iterating on each segment reduces the search space. Removing largest remaining peak (Fig. 3) can reveal smaller, previously hidden peaks. Some similarity to [3]. The largest remaining peak is most likely to allow a clean measurement of its width from measurement. Many peak models possible, but Gaussian is easy to generate from measured data parameters.

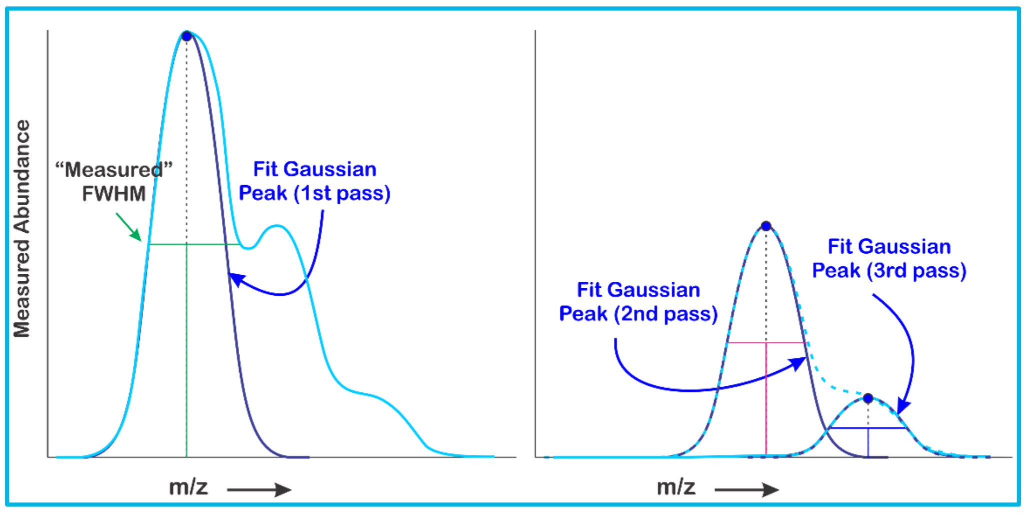
Measured Abundance
"Measured" FWHM
Fit Gaussian Peak (1st pass)
Fit Gaussian Peak (2nd pass)
Fit Gaussian Peak (3rd pass)
m/z
m/z

Figure 3: Successive Dominant Peak Removal

When side peaks are too close in distance and height to "dominant" peak, SDPR breaks down.

Unable to cleanly establish peak width for largest remaining peak due to interference (left of Fig. 4).

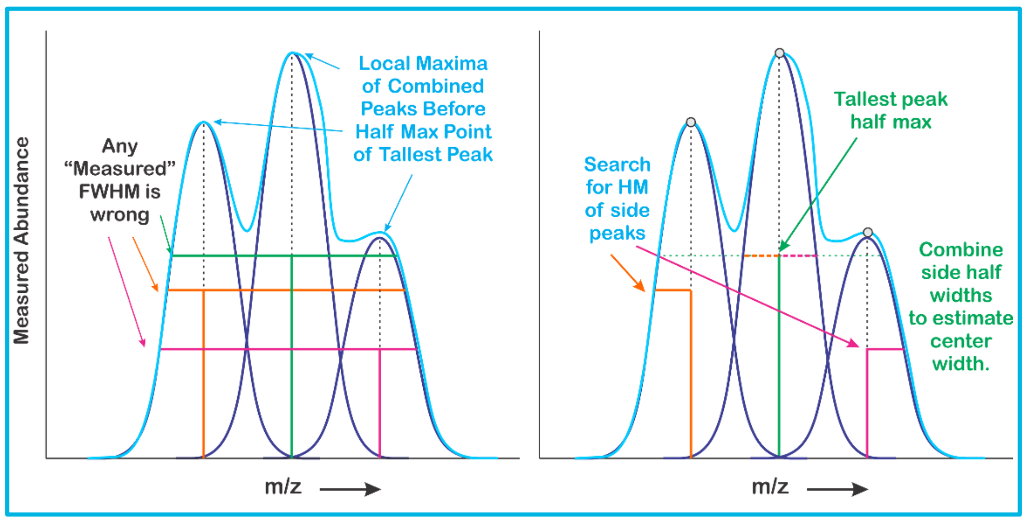
Measured Abundance
Any "Measured" FWHM is wrong
Local Maxima of Combined Peaks Before Half Max Point of Tallest Peak
Search for HM of side peaks
Tallest peak half max
Combine side half widths to estimate center width.
m/z
m/z

Figure 4: Issues with closely interfering side peaks..

Successive Multi-Peak Removal

Fix #3: Detecting and modeling side peaks, taking superposition effects into account, eliminates many of these issues. The steps include:

- Infer width of largest remaining peak from measurements of half widths of side peaks (right side of Fig. 4).
- Remove peak clusters as a group from residual abundance curve.
- Scale height of combined peak model to not exceed measured residual curve.
- If side peaks have their own side peaks, model these to calibrate the heights of inner side peak.
- Sanity check residual curve to not introduce artifacts.
- Iterate until no more peaks to be found in residual curve.
- Do this for each segment of significant abundance.

Sanity Check Against QTOF

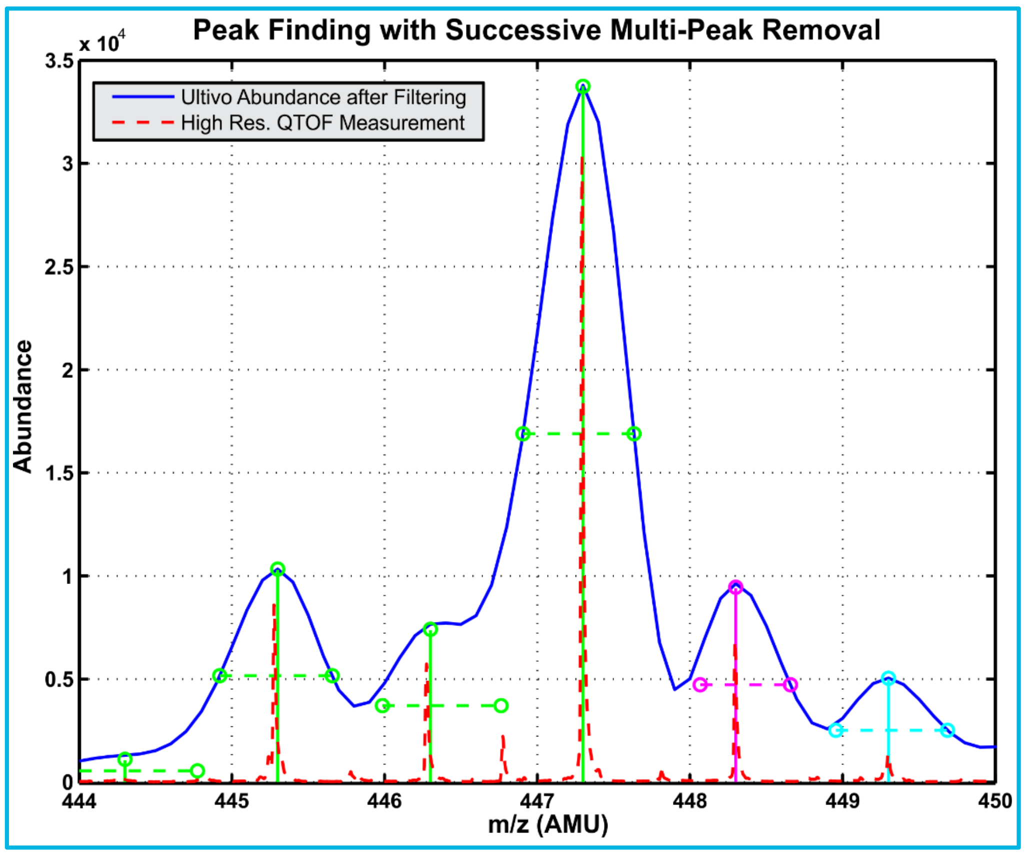

Figure 5: SMPR Checked Against QTOF

Results and Discussion

Experimental data taken from an Agilent Ultivo Tandem Quad Mass Spectrometer [4]. Measured sample was Polypropylene Glycol (PPG), with an average molecular weight of 1000 Da.

Data has content from 10 to 1400 AMU.

Data is saved unfiltered to MATLAB for processing.

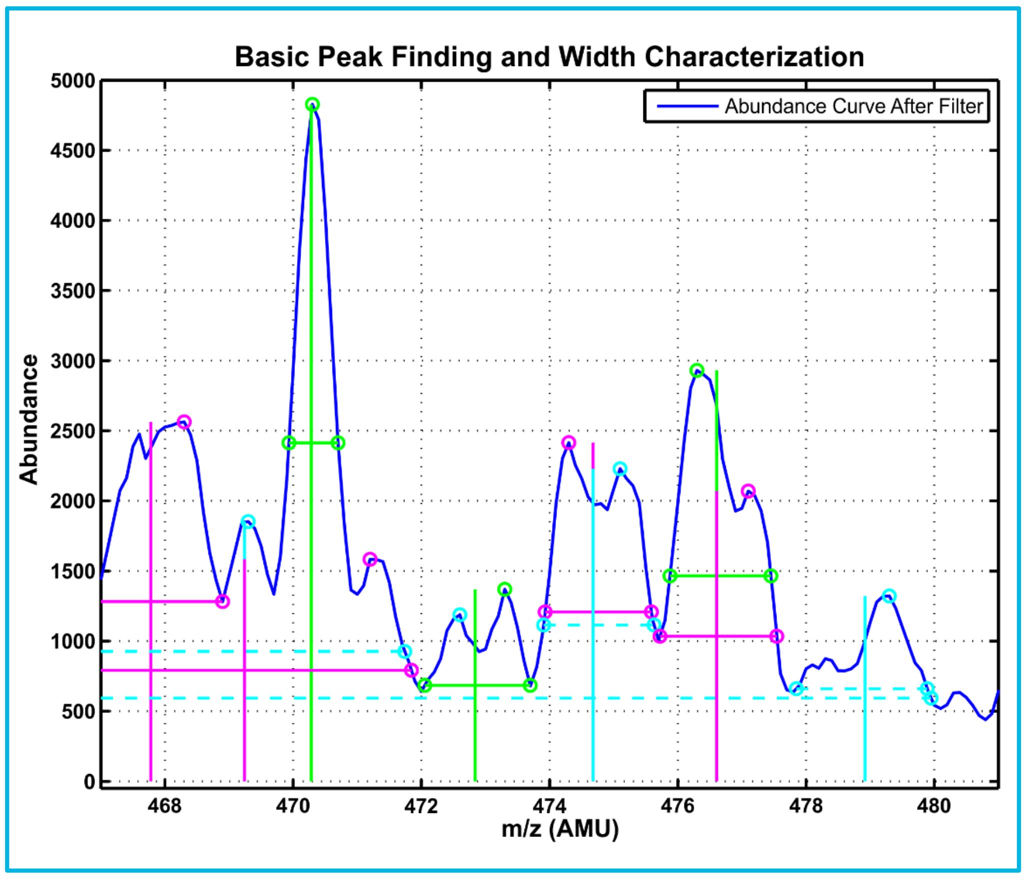

Figure 6: Basic peak finding has issues with overlap. Apex marked with circle, centroid with vertical line.

We compare 3 algorithms in regions that emphasize their differences:

- Basic Peak Finding and Width Characterization (Fig. 6)
- Successive Dominant Peak Removal (Fig. 7)
- Successive Multi-Peak Removal (Fig. 8)

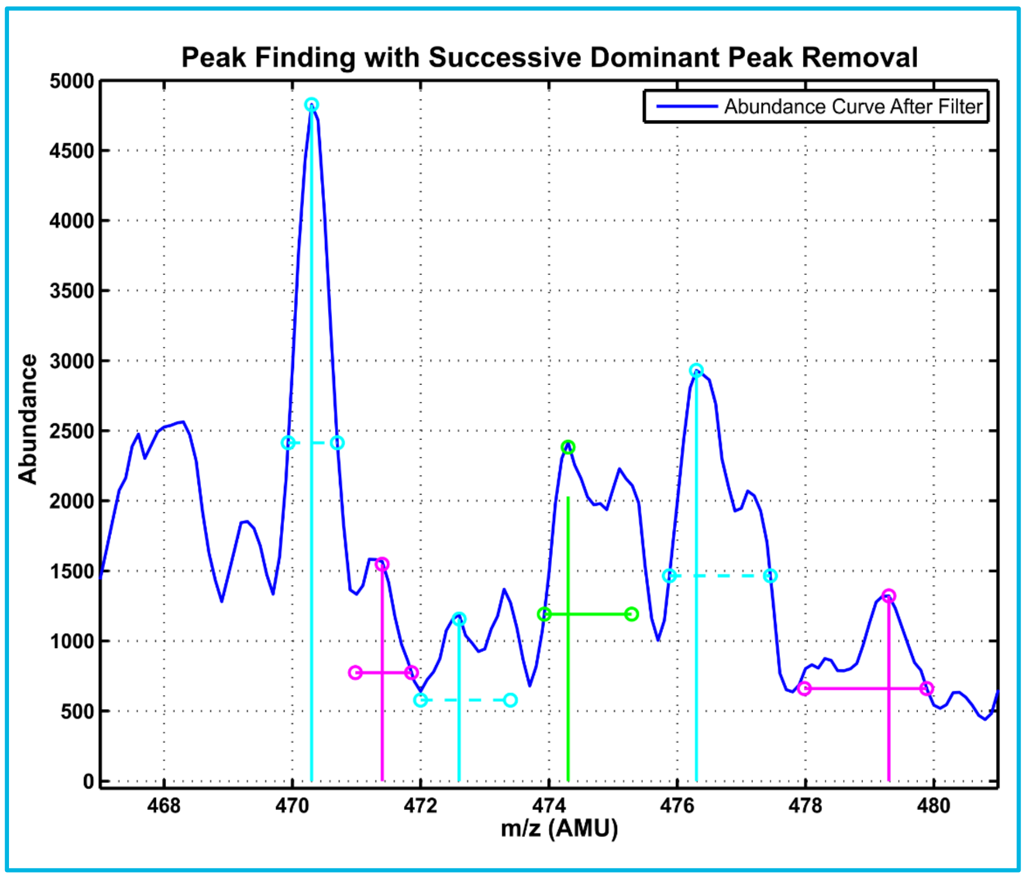

Figure 7: SDPR misses close side peaks. Horizontal segments on all plots are FWHM estimates.

Results and Discussion

Sanity checks between QTOF and Ultivo show that SMPR finds peaks close to QTOF abundance spikes.

- Differences generally accounted for by instrument/timing variations (Fig. 5).

Between 467 and 481 AMU, differences are obvious.

- Basic algorithm finds most apices, but miscalculates width, leading to poor centroid calculations (Fig. 6).
- Successive Dominant Peak Removal (SDPR) does a better job on many side peaks, but misses some that are assumed to be part of dominant peak (Fig. 7).
- Successive Multi-Peak Removal (SMPR) finds more peaks and calibrates for overlaps. Peak widths are all reasonable, meaning the peak centroids will be far more accurate than the previous two methods (Fig. 8).

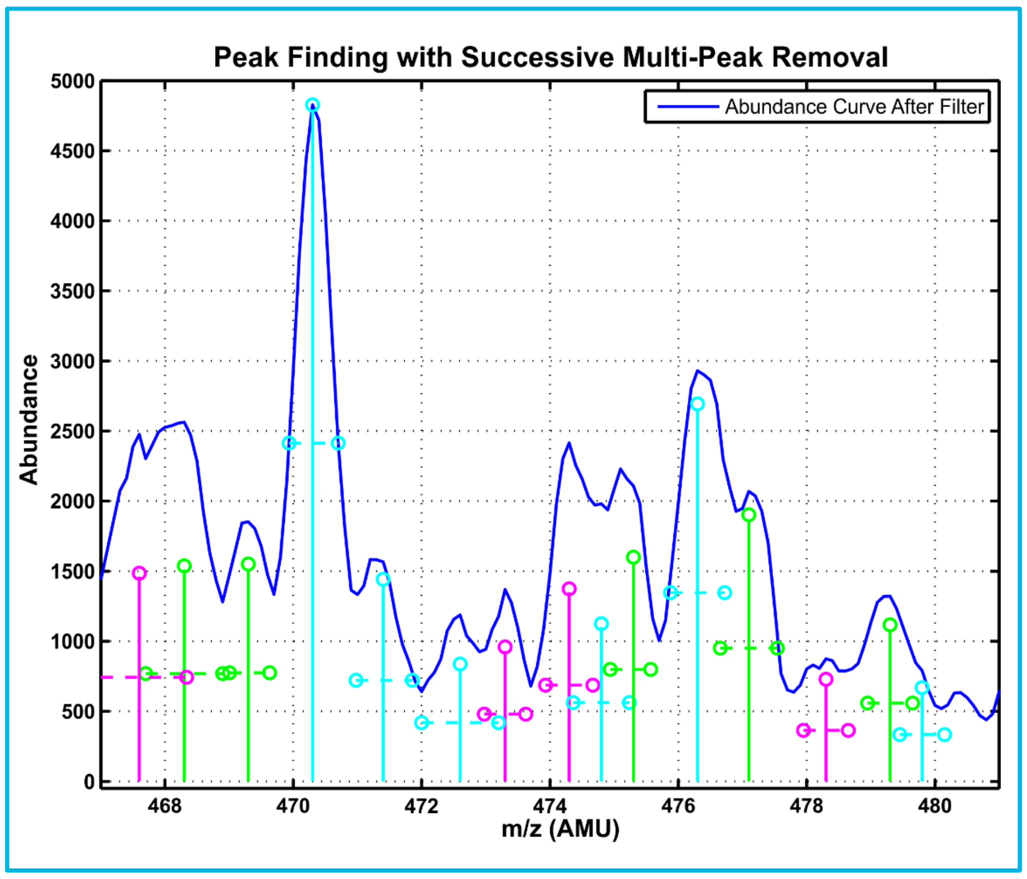

Figure 8: SMPR is more immune to interfering peaks.

Peak Type	Count	Percent (w.r.t. center peak)	Percent (w.r.t. left/right peaks)
Center	493	100	NA
Left	12	2.4341	NA
Left 2	1	0.2028	8.3316
Right	27	5.4767	NA
Right 2	5	1.0142	18.5185

Table 1: Cluster location statistics of found peaks

With the obvious improvement to abundance curve regions containing interfering peaks (Figure 8), it is reasonable to ask how significant a problem side peaks are. Table 1 shows statistics on center peaks (always detected), first, and second side peaks. We can see that even on this fairly complex matrix, the percentage of center peaks with significantly interfering side peaks (e.g. occurring before the half max point can be found) is about 8%. For those side peaks, about 15% had secondary side peaks.

For simple and isolated peaks, basic methods work well. In a small but significant percentage of overlapping peaks, SMPR has significant advantages.

Conclusions

Through improved segmentation of measurements and clever use of simple models, a dramatic improvement in peak finding is achieved, at least in regions where there is substantial overlap between peaks [5].

- “Divide and conquer” nature of the algorithm – dividing mass axis into smaller, separate regions of interest and then applying iterative methodology only on those smaller regions – keeps computational complexity of new algorithm low.
- This means, improved peak detection and compound identification can be made directly on instrument, rather than being relegated to specialized post-processing.
- Segmentation and improved peak characterization can be used as better starting point for more complex search and optimization schemes in post processing. Machine learning algorithms get smaller search spaces, and improved starting guesses in each of those spaces.

References

[1] Yang et. al., Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis, BMC BioInformatics, 2009.

[2] Press et. al., Numerical Recipes 3rd Edition: The Art of Scientific Computing, Cambridge U. Press, 2007.

[3] Wright, Methods of automated spectral peak detection and quantification without user input, PCT/US2010/036090, 2010.

[4] Agilent Technologies, Ultivo Triple Quadrupole Mass Spectrometer, Note # 5991-8146EN, 2017.

[5] Abramovitch, Method for finding species peaks in mass spectrometry, Patent Pending, 2018.

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

Agilent
Trusted Answers