Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal, Quebec, Canada, June 8 - 12, 2003

Automated Peak Picking and Integration Algorithm for Mass Spectral Data

Anthony J. Kearsley[+], William E. Wallace*, Javier Bernal[+], Charles M. Guttman*

[+]Mathematical & Computational Sciences Division, and *Polymers Division
National Institute of Standards and Technology, Gaithersburg, MD 20899-8541 USA

Today astonishing amounts of high-quality mass-spectral data are available at the press of a button. Still, quantitative data analysis is very often performed “by hand”. For example, it is not uncommon for the analyst to spend significant time inspecting mass spectra and selecting and integrating apparent peaks. This process is time consuming, is inherently subject to person-specific bias, and is not repeatable. Motivated by this, we seek to develop a fully-automated data analysis algorithm capable of rapid and reproducible processing of mass spectra containing large numbers of peaks. In addition to being computationally fast, the algorithm should be free of user input and/or parameter selection to save operator time and to remove operator bias. This algorithm would be extremely beneficial to NIST’s goal of creating a synthetic-polymer Standard Reference Material for absolute molecular mass distribution measurement by matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) where operator bias must be eliminated. Toward this end we present a mathematical algorithm that, given mass spectra of both pure matrix (to represent measurement background) and matrix with polymer analyte, quickly and robustly approximates the location and the area beneath all peaks in the analyte mass spectrum with no parameter selection or operator judgment. The numerical implementation employs only reproducible mathematical operations and yields consistent results on a variety of computer platforms, compilers and computing environments. This method requires no assumptions or knowledge regarding peak shape, size or distribution nor does it require preprocessing of the data, i.e. smoothing.

The numerical method consists of three parts (Figure 1). In the first part statistical parameters are automatically calculated from both spectra based on estimating the background noise (chemical, electronic, etc.) that is inescapable in MALDI (or any other) mass spectral measurement. Our experience suggests that the nature and magnitude of the noise contaminating the MALDI-MS output cannot be predicted solely from the experimental conditions; therefore, blind application of smoothing and/or filtering algorithms may unintentionally remove information from (or add “information” to) the data.

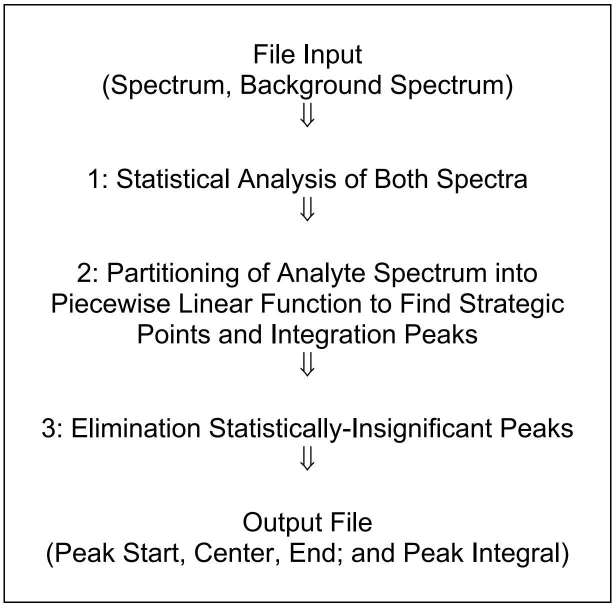
After this preprocessing phase is complete a piece-wise linear approximation to the data is constructed using a non-linear programming method [1]. This second part can be viewed as constructing an L_2 approximation to an L_1 fit [2]. Given a data set of N points, a collection of strategic points is selected. 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 which a point with greatest orthogonal distance is again selected. This numerical scheme is performed until the greatest orthogonal distance to any end-point-connecting line segment drops beneath a threshold value accurately estimated from the data. This method is an example of a multistage algorithm [3] where the first portion of the algorithm requires the selection of *strategic* points. The second phase of the algorithm requires the solution of an optimization problem, specifically, locating strategic point heights (or adjusting strategic y-axis values associated with strategic x-axis values) that minimize the sum of orthogonal distance from raw data. This problem is a nonlinear (and non-quadratic) optimization problem that can be accomplished quickly using a modern nonlinear programming algorithm (e.g. [1]).

In the last part the resulting strategic points of the piecewise linear approximation define a set of function extreme points corresponding to peak maxima and peak minima (which are used to define peak limits and peak maxima). The integration of the function between peak limits is calculated using a simple trapezoidal rule. (Clearly a more sophisticated integration strategy could be substituted with little effort.)

Finally, the algorithm output appears to be robust. Small changes in the data (e.g. numerically adding random noise) seem to have little effect on the output. This makes it an excellent candidate where uniformity is desired, for example as in an interlaboratory comparison [4].

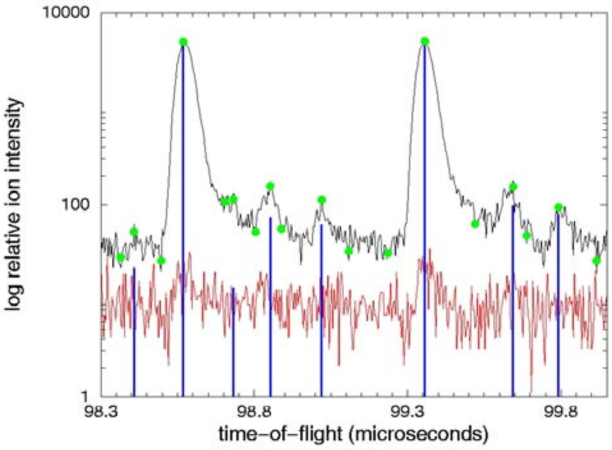
Figure 2 shows a mass spectrum of polystyrene cationized with silver, $[M+Ag]^+$. The main peak spacing is 104 u. The small intermediate peaks are from matrix adducts. (Note the log scale for ion intensity). This calculation, which revealed almost one hundred peaks in the full spectrum, took approximately 10 s on a desktop personal computer.

Figure 1 Algorithm Flow Chart



File Input
(Spectrum, Background Spectrum)
⇓
1: Statistical Analysis of Both Spectra
⇓
2: Partitioning of Analyte Spectrum into Piecewise Linear Function to Find Strategic Points and Integration Peaks
⇓
3: Elimination Statistically-Insignificant Peaks
⇓
Output File
(Peak Start, Center, End; and Peak Integral)

Figure 2 Data (black), Background (red), Strategic Points (green) and Relative Peak Areas(blue).



References:

[1] P. T. Boggs, A. J. Kearsley and J. W. Tolle, ”A practical algorithm for general large scale nonlinear optimization problems”, SIAM Journal of Optimization **9**(3) (1999) 755

2] I. Barrondale and F. D. K. Roberts, “An Improved Algorithm for Discrete L1 Approximation”, SIAM Journal of Numerical Analysis **10** (1993) 839

[3] C. M. Guttman, A. J. Kearsley and W. E. Wallace “A Numerical Method for Mass Spectral Data Analysis”, Applied Mathematics Letters, submitted

[4] C.M. Guttman, S.J. Wetzel. W.R. Blair, B.M. Fanconi, J.E. Girard, R.J. Goldschmidt, W.E. Wallace, and D.L. Vanderhart, “NIST-Sponsored Interlaboratory Comparison of Polystyrene Mass Distribution Obtained by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry”, Analytical Chemistry **73** (2001) 1252

E-mail: ajk@nist.gov (Anthony J. Kearsley)
URL: http://polymers.msel.nist.gov