

Signal conditioning and filtering of SELDI mass spectrometry time series

Dariya I. Malyarenko^{1,3}, William E. Cooke¹, Eugene R. Tracy¹, Haijian Chen^{1,3}, O. John Semmes², Maciek Sasinowski³, Michael W. Trosset¹, Dennis M. Manos¹

¹College of William and Mary, Williamsburg, Virginia 23187-8795; ²Eastern Virginia Medical School, Norfolk, Virginia 23507; ³INCOGEN, Inc., Williamsburg, Virginia 23185

Abstract: Clinical and biological research have used mass spectrometric measurements of peptide changes in body fluids to attempt an early detection of some diseases. Time-of-flight (TOF) mass spectrometry generally detects arrival times of ion signals to determine the ion masses, while Surface Enhanced Laser Desorption Ionization (SELDI) uses preprocessing surface chemistry to enhance the amount of important biological molecules in the TOF sample. All of these methods record the voltage output from ion detectors amplified by electronics designed to provide low noise, wide dynamic range, high sensitivity, and a repeatable response. However, the physical limitations of these detection systems can still introduce instrumental effects in the acquired raw records prior to subsequent statistical data analysis and interpretation. We address these physical limitations and discuss how they can produce unwanted attributes in TOF data records. These unwanted effects can include mass dependent broadening, baseline shifts, peak overlaps, and time-domain jitter. We briefly describe methods* to detect and correct such instrumental effects, and computational algorithms to enhance the mass resolution in corrected raw data, using time series analysis and filtering techniques.

*NOTE: Full account of the analysis and results has been submitted for publication to the Clinical Chemistry, 2004.

Background: The detection of peptide/protein levels in biological samples for clinical diagnostics [1] commonly uses high-sensitivity mass spectrometers (MS) in a linear, time-of-flight (TOF) detection mode. SELDI pre-selects a subset of peptides from a complex biological mixture by a surface processing procedure that concentrates the important molecular species based on their chemical affinity to a pre-activated surface [2]. Then, a short laser pulse releases and gently ionizes these peptides so that they can be separated according to their masses in a time-of-flight tube, with heavier masses traveling slower. The arrival times at a detector create a mass spectra, which is the ion yield as a function of mass. A digitizer records the voltage at the detector, including the response to the arriving ions and any electronic noise. In an ideal spectrum the peak intensity would represent signal strength, and the peak position (arrival time) would indicate the mass. Bio-marker peptides consistent with changes in biological state (e.g., healthy versus cancer [1]) should be evident as differences in the ion yield patterns. The ultimate goal is to classify new records on the basis of pattern similarity to that of a known disease group. However, all of this requires that the signal must be dominated by the mass data, rather than by electronic noise, so that a “peak” really represents the presences of some biological mass.

A variety of problems can reduce the usefulness of the SELDI mass spectra. For instance, the quality of SELDI chip surface may vary from spot to spot, and from chip to chip. Spot reading protocols, including steps between subpositions and laser intensity, may be changed from experiment to experiment, altering the ion yields. Detector electronics may distort recorded signals depending on analyte amount and arrival time. The fit relating the arrival time to an ion mass (*i.e.* the mass calibration) may depend on

the number of calibration peptide peaks and their range [3]. Changing the software parameters and the peak detection thresholds may significantly alter the number of entries, their amplitude and the position in the data matrix. Each of these experimental steps is a potential source of “noise” or biologically irrelevant variation in the data, possibly causing bias or irreproducibility in experiments [3,4]. In addition, SELDI technology limits mass resolution and has large electronic noise associated with its detectors. The net result is unnecessary overlapping of peaks in the recorded spectra, apparent mass jitter, and considerable difficulty in distinguishing mass peaks from background noise. Fortunately, many of these effects can be reduced significantly, by processing the signal using standard time-series background subtraction, calibration and filtering techniques.

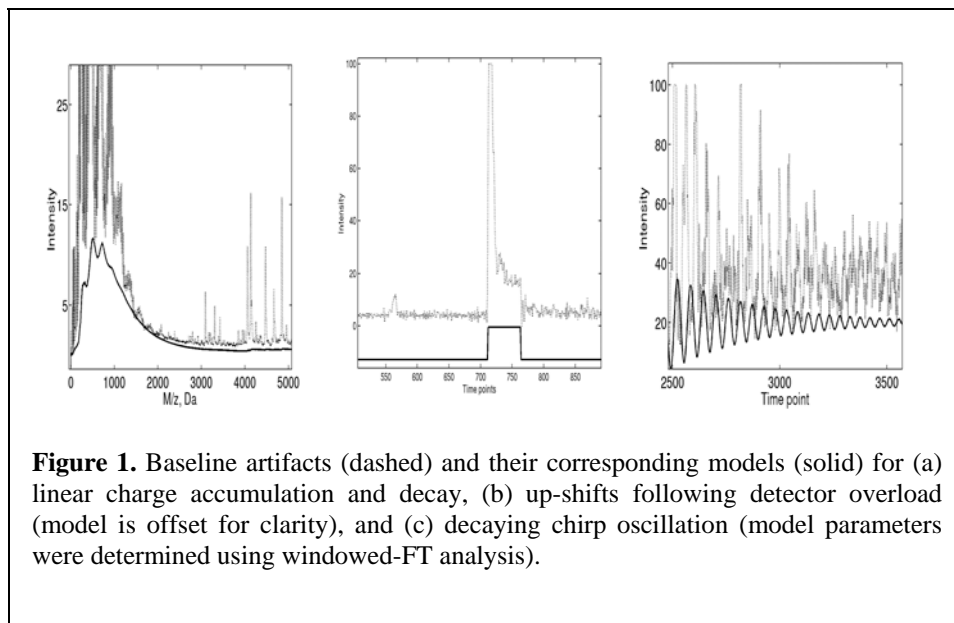
The unprocessed MS spectrum for each patient typically consists of tens of thousands measurements, with the patient group size no more than several hundreds. The denoising of MS data can be a critical preliminary processing step [4], since it provides a way to intelligently reduce the dimensionality of disease classification problem. This step generally influences several aspects of data matrix entries: the presence of an entry (biological relevance and association with a peptide), the column position of the entry (same mass/time variable for all patients), and the value of the entry (comparable scale with other entries in case of time-varying baseline, instrumental broadening, etc.). In what follows, we describe several hardware-related noise sources for SELDI data in relation to data matrix entries, and suggest signal processing approaches to eliminate the noise and enhance feature detection for robust classification.

Methods: We acquired SELDI spectra on a PBS II instrument from blank, hydrophobic and IMAC-Cu ProteinChip® arrays (Ciphergen Biosystems, Inc.) incubated with calibration peptide mixtures or pooled blood serum [1]. We recorded TOF spectra after single and multiple laser shots at different positions on the array spot. We reduced the noise by: (1) modeling and removing linear and non-linear baselines caused by detector saturation and charging effects; and (2) applying a default variable width moving average and rescaling a signal to recover a constant noise level. We reduced peak position jitter by correlating pivot mass peaks [5]. Finally, we used target filters [6] to suppress noise and to enhance peak resolution in the mass focusing range (20 to 12000 Da).

Results and Discussion: The systematic noise for PBS II instrument appears as baseline signal or random variations in the TOF records. Since the signals associated with peptide ions are the sum of the charges arriving at a specified time, they should always be positive. A typical SELDI spectrum is the result of averaging more than a hundred records from individual laser shots taken at different positions on the array spot. In all such spectra, the observed signal peaks usually appear to ride on a relatively smooth background. However, attempts to subtract this background without analyzing its sources frequently produce anomalous negative signals and other errors in the measured peak heights. We found at least three electronics artifacts that systematically contribute to a non-constant background: 1) charge accumulation-decay (Fig. 1a), 2) step-like shifts (Fig.1b), and 3) decaying chirp oscillations (Fig.1c) in response to saturation by highly abundant species (e.g., matrix or polymeric contaminants.)

The linear charge accumulation background correction (Fig. 1a) can be applied to single laser shot data or to data averaged over many laser shots. We found very little variation in model parameters as long as experimental parameters for acquisition are kept constant. However, changes in acquisition parameter settings (e.g., instrumental sensitivity) can alter charge accumulation efficiency and hence, the observed background. Unlike

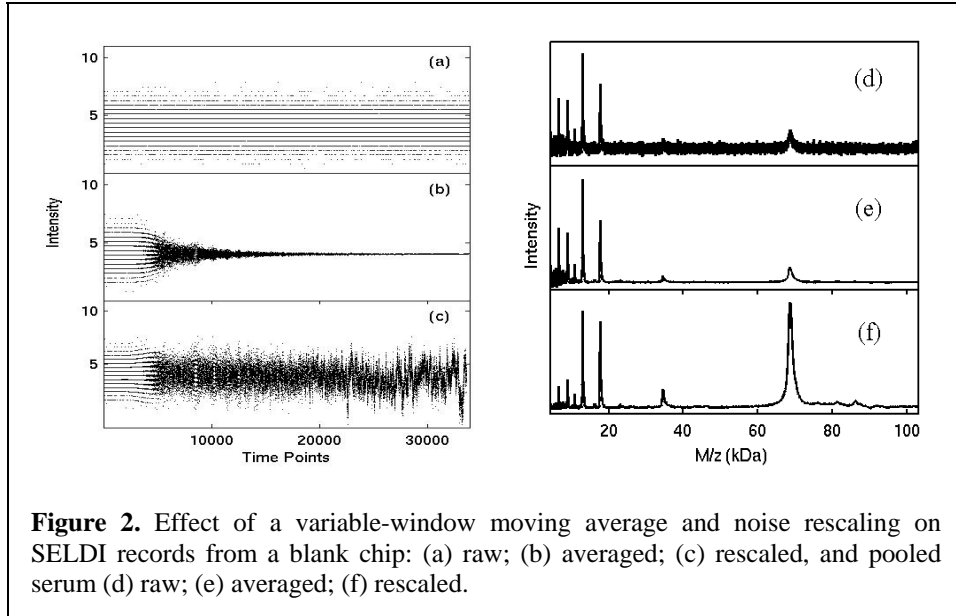
manufacturer's default algorithm (variable-width hull fitting), our baseline model has very few parameters, which have physical interpretation, and does not mistakenly eliminate slowly varying structure in broad overlapping peaks as background.



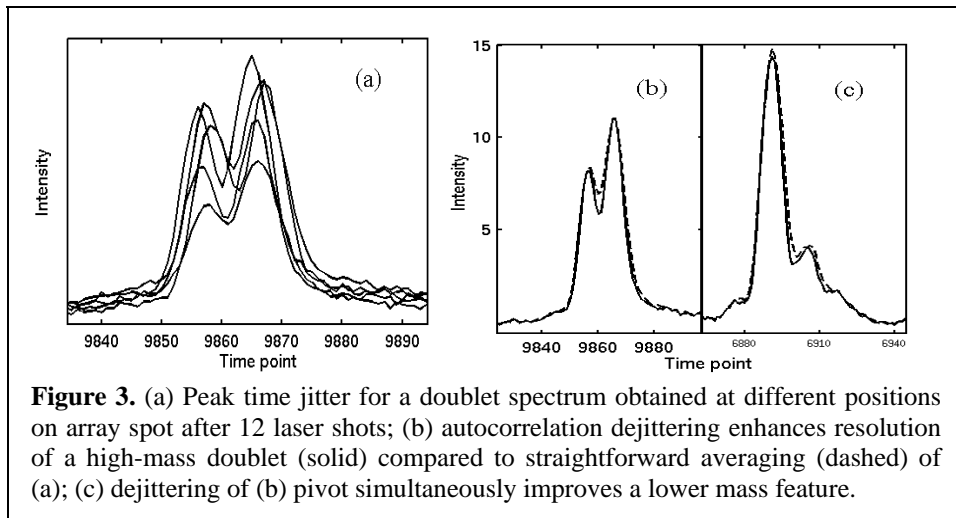
Two other contributions to the background are nonlinear and result from detector overload events most severe in the mass region below 3 kDa. Following an overload event, the nominal background level shifts by a large amount and stays shifted for a time that depends on the duration and the height of the overload (Fig 1b). Avoiding detector saturation can eliminate this artifact. It can be also corrected on a shot-by-shot basis during acquisition, if shifts (Fig. 1b) are subtracted before multiple shot averaging is done. Another nonlinear effect is that the output signal oscillates, or rings, for thousands of times steps following an overload event (Fig. 1c). Our analysis showed that oscillation frequency increases linearly as the ringing decays, which indicates that the electronic response is complex and nonlinear. These decaying coherent oscillations can be very difficult to remove from the baseline since they have both positive and negative components. When the overload cannot be avoided or nonlinear baseline cannot be modeled and subtracted, the data in the low mass range may not be fit for further statistical analysis. In general, baseline modeling and subtraction step impacts determination of relevance and correct value for the entry in the data matrix.

Once the linear and nonlinear background effects are eliminated the remaining noise in SELDI PBS II system appears to be random fluctuations in the detector (Fig. 2a). These can be reduced significantly by averaging – either over many laser shots, or over many time points. SELDI-TOF detector samples at constant time intervals so that any individual mass peak will be distributed over many dwell times, the number increasing for higher masses. This reduces the signal to noise ratio with growing mass since the inherent random fluctuations per dwell time are constant. Thus, for the high masses,

integrating or averaging over many dwell times is a natural correction for random noise that may not reduce the resolution.



The default Ciphergen PBS II signal processing routine uses a variable width moving average filter in the mass domain to gain this increase in the high-mass signal to noise. Figure 2b shows the effect of the manufacturer's default moving average on the noise from 2a. The rapid decrease in the amplitude is typical of the noise reduction due to the averaging of a random process over a longer time window. Figure 2c shows a renormalization of the averaged noise from Figure 2b, by the square root of the window size in time points. This rescaling results in a nearly constant amplitude noise level, although moving average does introduced some structure into resulting noise. When the same moving average and rescaling procedure is applied to the SELDI record of a pooled serum sample (Fig. 2d-f), peak picking becomes easier, since it is not necessary to re-calculate the noise level at different masses to find a local noise level. This procedure helps broad high-mass peaks become clearly visible above the noise level, as is evident from the example in Figure 2f for the singly charged albumin peak (68 kDa) and doubly-



charged albumin (34 kDa). Furthermore, the relative intensities of rescaled features provide better measures of relative ion yields over a broad mass range with nearly constant noise level.

Since TOF measurements relate the mass of an ion to the time it takes to reach the detector, the calibration that relates the arrival time to the actual ionic mass is a crucial characterization step. The errors in arrival times would produce the mass calibration errors and possible misalignment of features in a data matrix. We found that when laser moves between sub-positions on the array spot, we observed significant apparent shifts in the peak timing for the same feature (Fig. 3a). This is due to the laser beam moving over a relatively large distance from position to position, altering the net flight time for the ions of the same mass, which are extracted from different heights of randomly piled crystals on the spot surface. We corrected these shifts by introducing a sub-position dependent time shift derived by maximizing the cross-correlation between each individual trace and the average trace over the narrow doublet near the time point 9860 (8464 Da). When these corrected traces were then averaged, they produced a spectrum with improved resolution, as shown in Figure 3b. In addition, Figure 3c shows that the same time shifts can change the mass location of the peak at 3444 Da (time point 6290), and simultaneously improve the mass resolution there.

The real benefit of using this correlative procedure [5] to correct for calibration changes between sub-positions is that it does not require a known mass. It needs only a single feature with relatively narrow structure. This auto-calibration procedure provides an obvious improvement in correcting peak location and intensity, which make the entries in the data matrix more accurate. In addition to autocalibration during acquisition, a similar approach can be utilized for detecting and correcting shifts between different spots and arrays, and hence for proper alignment of variables in the data matrix.

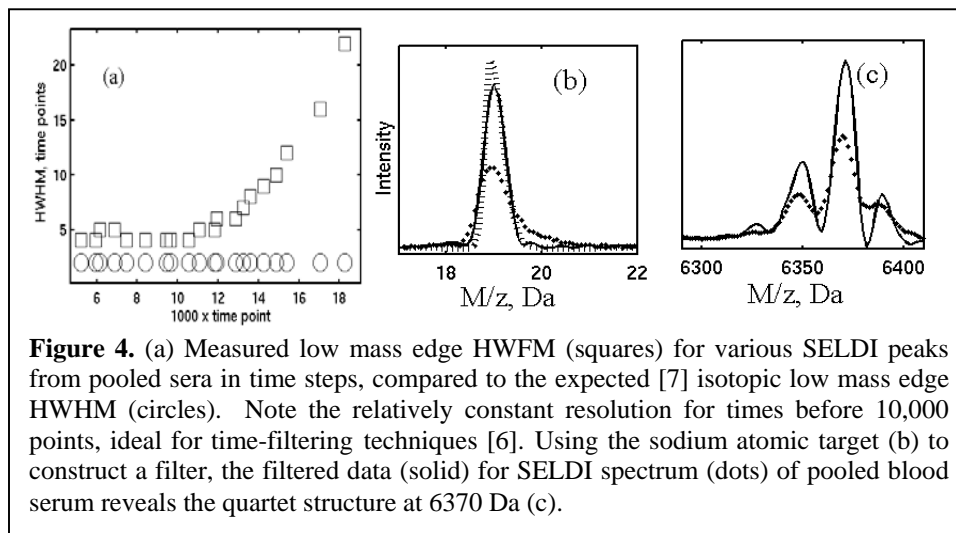


Figure 4. (a) Measured low mass edge HWHM (squares) for various SELDI peaks from pooled sera in time steps, compared to the expected [7] isotopic low mass edge HWHM (circles). Note the relatively constant resolution for times before 10,000 points, ideal for time-filtering techniques [6]. Using the sodium atomic target (b) to construct a filter, the filtered data (solid) for SELDI spectrum (dots) of pooled blood serum reveals the quartet structure at 6370 Da (c).

Peak size and location can be determined more accurately if the resolution of the records is enhanced. We found that target filter deconvolution [6] can significantly enhance the resolution in SELDI TOF data. The idea behind this filtering technique is to fit a part of the record that has a single peak on a noisy background, which has a shape characteristic of the instrument function and also has a high signal to noise ratio. We can use it to create

a filter that maps the instrument function into a wavelet of arbitrary shape (target). Typically, the desired shape is a symmetric peak with a narrower line shape. After designing the best-fit filter, we can then apply it to the entire mass-focusing region, assuming a constant instrumental function and stationary noise. This type of an approach was warranted in a TOF spectrum, since the instrumental resolution of peaks from masses as small as sodium (23 Da) up to approximately 12 kDa is nearly constant in time (Fig. 4a). Thus, the monoisotopic sodium peak could be used as a target for filter construction (Fig.4b). Figure 4a also shows that this filter would not be appropriate outside the mass focusing range since the instrumental function changes rapidly.

The net effect of resolution enhancement via target filtering in the mass focusing range is that some spectra that appeared to have slight shoulders on large peaks can be resolved to identify clear satellite peaks (Fig.4c). These peaks then can be easily detected and interpreted as chemical adducts or neutral losses for the parent peptides. Interestingly, after deconvolution the majority of small peaks in the pooled serum spectra that we obtained from IMAC-Cu chip could be identified with sodium and sinapinic acid adducts or water and ammonia neutral losses of a relatively small number of distinct peptides.

Conclusions: Typical unprocessed SELDI-TOF-MS data show contribution from many electronics artifacts related to acquisition protocols and technology limitations. If not taken into account and eliminated by data processing before classification, these artifacts may bias the data or limit experimental reproducibility. The time series techniques that we applied to SELDI-TOF data before any peak identification procedure, can significantly improve the data to make the feature identification process simpler and more robust. Filtering SELDI-TOF data in the time domain significantly improves the data reproducibility to enhance data interpretation and to ease the comparison among data sets for diagnostics and classification applications. The resulting improvements for data matrix would include 1) reduction of irrelevant data entries (dimension reduction) by noise suppression and association of higher resolution features with biological species, 2) proper alignment of entries for the same variables by peak position calibration and resolution enhancement, and 3) correction of values to represent relative peptide yields free of electronics artifacts and noises by baseline elimination, peak deconvolution, and noise suppression.

Acknowledgements: This work was supported by VA CTRF Fund #IN2002-03, and Phase I SBIR grant from the NCI CA101479. We thank Dr. Bao-Ling Adam and Dr. Malik Gunjan for assistance with acquisition of calibration data for SELDI PBSII instrument. We are grateful to Dr. Stacy Moore and Dr. Scott Weinberger from CIPHERGEN, Inc. for their help in clarifying instrumental specifications and parameters.

1. Adam B-L, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL Jr. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res* 2002; 62:1609-14.
2. Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* 2000; 21(6):1164-1167.

3. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF Protein patterns in Serum: Comparing Data sets from Different Experiments. *Bioinf Adv Access* 2004; Jan:1-9.
4. Coombes KR, Fritsche HA, Clarke C, Chen J-N, Baggerly KA, Morris JS, Xiao L-C, Hung M-C, Kuerer HM. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem* 2003; 49(10):1615-1623.
5. Nicola AJ, Gusev AI, Proctor A, Hercules DM. Automation of Data Collection for Matrix-Assisted Laser Desorption /Ionization Mass Spectrometry Using a Correlative Analysis Algorithm. *Anal Chem* 1998;70:3213-3219.
6. Robinson EA, Treitel S. *Statistical communication and detection*. London: Griffin, 1967:pp249-283.
7. Senko MW, Beu SC, McLafferty FW. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *J Am Soc Mass Spectrom* 1995; 6:229-233; (<http://prospector.ucsf.edu/ucsfhtml4.0/msiso.htm>).