Published in final edited form as:

Proteomics Clin Appl. 2011 August ; 5(0): 440–447. doi:10.1002/prca.201000095.

Improved Signal Processing and Normalization for Biomarker Protein Detection in Broad-Mass-Range TOF Mass Spectra from Clinical Samples

Maureen B. Tracy¹, William E. Cooke², Christine L. Gatlin³, Lisa H. Cazares⁴, Dennis M. Weaver⁵, O. John Semmes⁴, Eugene R. Tracy², Dennis M. Manos^{2,3}, and Dariya I. Malyarenko³

¹William and Mary Research Institute (WMRI), College of William and Mary (CWM), Williamsburg, VA 23187-8795 USA

²Department of Physics, CWM, Williamsburg, VA 23187-8795 USA

³Department of Applied Science, CWM, Williamsburg, VA 23187-8795 USA

⁴Center for Biomedical Proteomics, Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, VA 23501 USA

⁵Saint Leo University, Langley Center, VA 23665 USA

Abstract

Purpose—To demonstrate robust detection of biomarkers in broad-mass-range TOF-MS data.

Experimental Design—Spectra were obtained for two serum protein profiling studies: (1) 2–200 kDa for 132 patients, 67 healthy and 65 diagnosed as having adult T-cell leukemia and (2) 2–100 kDa for 140 patients, 70 pairs, each with matched prostate-specific-antigen (PSA) levels and biopsy-confirmed diagnoses of one benign and one prostate cancer. Signal processing was performed on raw spectra and peak data were normalized using four methods. Feature selection was performed using Bayesian network analysis and a classifier was tested on withheld data. Identification of candidate biomarkers was pursued.

Results—Integrated peak intensities were resolved over full spectra. Normalization using local noise values was superior to global methods in reducing peak correlations, reducing replicate variability, and improving feature selection stability. For the leukemia data set, potential disease biomarkers were detected and were found to be predictive for withheld data. Preliminary assignments of protein IDs were consistent with published results and LC-MS/MS identification. No PSA-independent biomarkers were detected in the prostate cancer data set.

Conclusions and clinical relevance—Signal processing, local SNR normalization and BNA feature selection facilitate robust detection and identification of biomarker proteins in broad-mass-range clinical TOF-MS data.

Correspondence: Maureen Tracy, College of William and Mary, WMRI, c/o Dept. of Physics, P. O. Box 8795, Williamsburg, VA 23187-8795 USA. mbtrac@wm.edu, fax: 757-221-3540.

Conflict of Interest Statement

Authors declare no financial/commercial conflicts of interest.

Keywords

normalization; protein profiling; serum biomarkers; signal processing; TOF-MS

1 Introduction

Protein profiling using TOF-MS is a common tool for biomarker discovery [1 – 3]. Issues remain, however, that can limit the reliability and reproducibility of the results. Variability in peak-intensity measurements can prevent direct comparisons between spectra, and correlations between peaks can cause instability in feature selection and classification. Both of these issues are, in part, due to systematic errors and should be reduced with effective normalization. Various normalization methods are in use including offset removal and/or scaling by global and local factors [4–7] and the use of wavelet decomposition to exclude undesirable signal variations [5]. However, no single method is clearly superior for subsequent comparative analyses [7]. One commonly used method, dividing each spectrum by its total ion current (TIC), has had questions arise about its suitability for comparison of spectra from different clinical groups and the effect of noise [8, 9]. As we show in this paper, TIC normalization can introduce noise contamination that is indicated by strong negative correlations between peaks. This contamination arises from random error in the large amplitude peak measurements that dominate the TIC: while small compared to large peaks, it is large compared to small peaks and can have a disproportionately large effect on the latter during normalization. This, in turn, can lead to erroneous feature selection.

The goal of this paper is to demonstrate the robust detection of protein biomarkers in broad-mass-range TOF-MS data (spectra for which the mass range extends beyond the high-resolution mass-focusing range). We use algorithms previously developed by members of our group for automated signal processing [10–12] and Bayesian Network Analysis (BNA) for feature selection [13]. (Both R and MATLAB versions of the software used are available from the URLs provided in the Sup-Info.) A key step in achieving our goal is to demonstrate effective normalization.

We begin by optimizing our signal processing parameters to obtain protein peak data over the full mass range for two data sets. Next, we use BNA to select peaks that have potential for discriminating between the different clinical groups. Since feature selection requires numerous comparisons of peak values from different spectra, it is necessary to provide an effective normalization method. To test the effectiveness of four candidate normalization methods, we used 173 spectra obtained from Quality Control (QC) pooled serum during one of our studies. The rationale for choosing QC data to qualify normalization methods was the expectation that the hallmark of effective normalization of technical replicate data would be that any residual variation would come entirely from random errors. Furthermore, with the removal of systematic error, residual correlations between peaks should be due only to chemical relationships (i.e. ionization states, protein modifications, adducts and neutral losses) that can aid in preliminary protein assignments. (See the 2009 HUPO poster at the URL provided in the Sup-Info.)

The four normalization methods we tested were: a) TIC normalization which, as the most commonly used method, provides a benchmark for the other methods; b) “de-trending” which involves the application of an empirical factor to remove a single observed source of systematic error (a secular decrease in laser desorption power with shot number); c) local SNR (peak signal divided by local noise), to investigate local normalization [4]; and d) global SNR (peak signal divided by global noise) for comparison to local SNR. Below, we will show that, compared to “global” methods (wherein each spectrum is divided by a constant), local SNR yields fewer spurious peak correlations, reduces replicate variability, and leads to more reproducible feature selection in BNA and classification. While local SNR is slightly more involved, we show that the additional effort is repaid; this treatment of our leukemia data yields the most peaks with residual variations and correlations that are consistently associated with the presence or absence of disease. We show how these features allow the construction and testing of a classifier and help make preliminary assignments of protein identities. Finally, we show progress toward the general problem of identification and quantitative verification of biomarkers.

2 Materials and Methods

2.1 Data Sets

Complete raw data sets and processed results are available in R data structures from the ProteomeCommons.org Tranche network. (URLs are provided in the Sup-Info.)

2.1.1 QC Pooled Serum—QC pooled serum was created from sera collected from 360 healthy individuals and was used in both studies described below. Two attributes of our QC data are important to highlight. First, QC spectra obtained from technical replicates of the same pooled serum sample allow characterization of instrumental and sample preparation variability separate from biological variability which is present in clinical data. Since the primary objective of normalization is to remove systematic experimental variability and not biological variability, QC data is ideal for qualifying normalization methods. Secondly, the acquisition of QC and clinical spectra was interspersed to ensure that signal processing models and parameters determined for QC spectra could be applied to clinical data [10, 11, and 14].

2.1.2 Leukemia Study—A subset of TOF-MS spectra from a 2004 NCI leukemia serum protein profiling study [15] was used. Protocols for patient consent (NIH 98N-047) and data acquisition are described elsewhere [13, 15]. Serum samples were purified using IMAC-Cu affinity capture surfaces, robotically spotted, and data was acquired over the range 2–200 kDa using a CIPHERgen PBS II instrument (BioRad Laboratories, Inc., Hercules, CA). Data included 173 replicate spectra from QC pooled serum interspersed with 390 spectra from 132 clinical serum samples obtained from 67 normal and 65 clinically diagnosed adult T-cell leukemia patients (ATL). In addition, 191 spectra from 66 patients which were not identified as belonging to either clinical group (lockbox data) were reserved to test classifiers.

2.1.3 Prostate Cancer Study—A second data set was acquired for a prostate cancer (PCA) serum protein profiling study. Serum samples were obtained from the Virginia

Prostate Center Tissue and Body Fluid Bank. Samples were acquired following informed consent under a protocol approved by the Eastern Virginia Medical School Institutional Review Board. Serum collection protocol details are described elsewhere [14]. Serum samples were purified using IMAC-Cu affinity beads, robotically spotted, and data was acquired over the range 2–100 kDa using a Bruker Ultraflex III (Bruker Daltonics, Inc, Billerica, MA) in linear mode. Details of the optimized acquisition procedure for the broad-mass-range are provided in reference [14]. Data included 86 replicate QC spectra interspersed with 323 spectra from clinical serum samples obtained from 70 patients with, and 70 patients without cancer as confirmed by biopsy. To eliminate non-specific biomarker detection related to serum prostate-specific-antigen PSA levels, each patient pair had matched PSA values: subjects were selected from patient populations with benign prostatic hyperplasia (BPH) and PCA in pairs whose PSA levels were matched to within 0.5 ng/ml over a range of 2–10 ng/ml.

2.2 Signal Processing

Raw TOF-MS data were processed using automated time-domain tools designed to improve sensitivity and specificity over a broad mass range [10, 11, 14, and 16]. These methods have been shown to increase signal detection sensitivity 10–30-fold for $m/z > 15$ kDa [11] and produce integrated peak intensities, allowing comparative statistical analysis downstream.

For both data sets, QC data was used to optimize the adjustable parameters associated with the signal processing and peak detection models. Detailed user information is available in the documentation provided with the R-packages. In addition, guidance for parameter selection and optimization are provided in the comments in the “OptionsAndParameters” files (available from URLs provided in the Sup-Info.) QC and clinical spectra were thereafter processed using the optimized models and aligned peak lists were generated.

Signal processing for the leukemia study data included (in the order prescribed by the automated tools): (1) pre-alignment using initial conditions recorded by the instrument; (2) fitting and removing a baseline model (exponential decay plus a constant offset) for each spectra; (3) integrative down-sampling using a quadratic equation to ensure constant peak width across the spectra; (4) amplitude rescaling to ensure Gaussian noise (a correction due to the presence of derivative instrumental noise) (5) optimal linear filtering using an asymmetric Gaussian model for the target peak shape to improve signal-to-noise (SNR); (6) pedestal removal, using a moving average of local minima, to reduce correlations between overlapping peaks; (7) peak detection using an automated local noise threshold estimate with respect to local minima for average spectrum; and (8) alignment of individual spectral peaks to the average spectrum. Options for processing steps include some that are commonly used such as alignment to average spectrum [17, 18]. TOF to m/z calibration was achieved by fitting a quadratic formula for m/z to TOF values for known proteins in spectra (protein standards). The calibration was extended to high masses using the albumin peak (66.4 kDa) and its three detectable ionization satellites [12].

The PCA study data was processed using the same methods as the leukemia study data with the following exceptions necessary because of differences between the instruments used: (1) the baseline model was a Gaussian function plus a constant offset; (2) no amplitude

rescaling was required as the high frequency noise was Gaussian; (3) the target peak shape was a symmetric Lorentzian model; and (4) the noise estimate used in peak detection was corrected for square-root baseline-amplitude-dependent Poisson noise.

2.3 Normalization

Prior to the reduction of the clinical data, four normalization methods, TIC, de-trending, local SNR and global SNR, were studied first using QC data from the leukemia study. These normalizations were thereafter applied to clinical data from the leukemia study.

For de-trending normalization, an equation was fitted to TIC values from replicate QC spectra as a function of acquisition index. This equation exhibited an exponential decay and yielded a single factor by which to scale each spectrum. For both local and global SNR, noise was estimated using a Gaussian model and by estimating the SD of amplitude differences between sequential points. To ensure this estimate represented differences due to noise and not the presence of peaks, the SD was estimated assuming knowledge of only the 10% of the distribution closest to a mean value of zero. The peak intensity is thus larger than the local noise and can't be "lost" with local SNR normalization as could occur with TIC normalization.

Three criteria were used to determine the effectiveness of each normalization method: (1) a reduction in peak correlations (without the introduction of erroneous negative correlations); (2) a decrease in peak amplitude CVs; and (3) feature selection stability for clinical data. The first two criteria gauge systematic error reduction and the third criterion is a measure of the reproducibility for our goal of biomarker detection. The normalization method found to be most effective for the leukemia study data, local SNR, was thereafter applied to the PCA study data. For comparison, TIC normalization was also applied as a benchmark to PCA study data.

2.4 Feature Selection Method and Classification

Feature selection was performed on un-normalized and normalized peak-intensity clinical data using BNA [13]. BNA uses mutual information to select peaks that show dependency on clinical group and assigns them to Level 1 in the Bayesian Network (L1 peaks). Further, it detects peaks that are not directly dependent on clinical group but are dependent on the L1 peaks and assigns them to Level 2 in the Bayesian Network (L2 peaks) allowing the detection of "related" clusters. L2 peaks may either be direct products (fragments, ionization states, etc.) of L1 peaks, or can be regulated by disease state "through" L1 peak. This is in contrast to "co-regulated" moieties which would be expected to appear at the same level in the network. Selecting only one peak from a cluster as linked to clinical group improves stability of feature selection and subsequent classification [13]. BNA was performed using replicate averaging, 100 iterations with 10-fold cross validation (subpopulations drawn randomly), and with a high threshold of 3.8 times the Mutual Information baseline for data with randomized clinical group assignments. Selection stability is indicated by selection rate: the maximum selection rate is 1 and would occur only if a peak were found to be dependent on clinical group in all (1000) runs.

Naïve Bayesian classifiers were built (using in-house code based the BNA toolbox) using only data from peaks whose BNA selection rates exceeded 90%. The performance of these classifiers was tested using the lockbox data for three cases: un-normalized, TIC normalized and local SNR normalized data. Processing for lockbox data was identical to processing for the training data (used for feature selection) except that the m/z values for each peak were prescribed to match those detected for the training spectra. The classifiers were also tested using the training data as test data for reference and replicate averaging was used for all cases. Peaks with high selection rates and good classifier performance were considered candidates for disease biomarkers and were used in subsequent protein identification by LC-MS/MS and verification by Western Blot and ELISA. (A brief summary is provided in the Sup-Info. Details will be provided in a forthcoming publication by White et.al.)

3 Results and Discussion

3.1 Leukemia Data Set

3.1.1 Signal Processing—Signal processing of QC and clinical spectra from the leukemia study produced peak lists with 110 and 115 peaks, respectively, over the full mass range. Figure 1 illustrates the average of processed 2–200 kDa QC spectra co-plotted with the average of processed (higher resolution) 1–50 kDa QC spectra. For clarity, the image only shows the range 3–15 kDa over which the peak width is typically constant. Data from the range 1–50 kDa have been previously analyzed and results reported elsewhere [13, 15].

3.1.2 Peak Correlations—Given systematic variation between spectra, such as from laser intensity attenuation, the variation of each peak (between spectra) tracks that of other peaks (within each spectrum) and thus correlation values between peaks are high. Since our goal is to remove systematic error, it would be helpful to know what to expect in the limiting case of its absence. We can do this by considering the peak amplitudes in N technical replicate spectra taken from true QC pooled serum peak amplitudes as N independent identical draws from an underlying population. Assuming the measured peak amplitudes are uncorrelated normal random variables, we show (in the Sup-Info.) that the distribution of correlation coefficients (r) between the purely random peak amplitudes in the N technical replicate spectra is Gaussian with mean = 0 and $SD=1/\sqrt{N}$. Any significant peak correlation values ($r \gg SD=1/\sqrt{N}$) in QC data indicates either uncorrected systematic error, or chemical relationships between peaks.

To illustrate peak correlation statistics, prior to and following normalization, histograms of the correlation coefficients for QC data are shown in Figure 2. (Correlations between specific peaks is shown in the heat maps provided in the Sup-Info.) As seen in the upper left part of Figure 2, before normalization, positive correlations do exist between most peaks indicating systematic errors. Applying de-trending shifts some of the positive correlations to lower values (Fig 2, upper right). TIC reduced some of the high correlations, but also introduced strong negative correlations (Fig 2, bottom left). These correlations are erroneous because the intensity of no one peak should increase while that of another decreases in technical replicate spectra. Only local SNR reduced correlations across the spectra without introducing erroneous negative correlations (Fig 2, lower right). While the distribution of the

correlation coefficients after local SNR normalization remains broader than the theoretical ($SD=0.22$ is almost 3 times the theoretical value of $SD=1/\sqrt{173}$), the approximate Gaussian shape and a mean close to zero (mean=0.08) indicate successful reduction in systematic error. The residual high correlations (tail on the right) are believed to result from true chemical relationships.

Similar effects were observed for clinical data; results were less dramatic since peak correlations were much lower prior to normalization. Histograms and heat maps of correlation coefficients, as well as CV histograms for the clinical data are found in the Sup-Info.

As a final caution about TIC, and to further illustrate the error that can result from its use, Figure 3 provides intensity scatter plots for 9.29 and 66.3 kDa peaks, prior to normalization and with TIC and local SNR normalization. While the two peaks initially had low correlation (0.28), TIC normalization introduced strong negative correlation (−0.93). Local SNR reduced the correlation below the initial value (to 0.11).

3.1.3 Peak CVs—The second measure of normalization performance considered was peak CV. Figure 4 provides histograms of CVs calculated for all detected peaks prior to and after normalization for QC data. Although TIC normalization reduces CV considerably, local SNR reduces it further, indicating that local SNR provides the best peak intensity reproducibility of the methods compared. A few high outliers remain, as they also do for TIC, which may result from poor alignment for peaks in regions with high noise.

3.1.4 BNA Feature selection stability—The final measure for normalization performance was feature selection stability. Table 1 lists peaks that had selection rates greater than 90% for each normalization method, along with the cross validation error rates (mean $\pm$ SD). Local SNR resulted in five peaks, which was the highest number of consistently selected peaks. One peak, 11.5 kDa, had high selection rates for all normalization methods (except global SNR with 75%). The selection rate for the others depended on normalization method. Furthermore, several peaks (11.7, 11.9 and 12.1 kDa) were determined by the BNA to always be dependent on 11.5 kDa; this cluster has been previously documented [1, 13, 15, 19]. It is important to note that our normalization does not suppress chemical correlations. For the cluster described above, for all but one minimal exception, normalization increased the correlations between 11.5 and 11.7, 11.9 and 12.1 kDa peaks. Local SNR produced the greatest increase in these correlations values. Figures that illustrate the stability of L1 and 11.5 kDa dependent L2 peaks, the Bayesian network, and the correlations described above are provided in the Sup-Info.

3.1.5 Classification—Naïve Bayesian classifiers were built using peaks with BNA selection rates over 90%. Three, one, and five peaks were used for un-normalized, TIC normalized and local SNR normalized data, respectively. In addition to the peak m/z values and BNA cross-validation error rates, Table 1 includes classification error rates for training data (used as test data) and test data. Classification error rates for test data normalized using local SNR and TIC did not exceed the predicted (cross-validation) error rates indicating strong predictive power. The classification error rate for un-normalized data exceeded the

BNA cross-validation error rate. Test data normalized using local SNR had the lowest error rate.

3.1.6 Preliminary Assignments and Protein ID—Peaks with high selection rates (discriminating power) and good classifier performance (predictive power) are considered potential biomarkers [13]. Several such peaks from this study and the 1–50 kDa study [13] were selected for identification. This is the subject of the future paper (White, et al., 2011) and only selected results are presented here.

The 11.5 kDa peak that was consistently linked to clinical group for all normalization methods is part of the 11.7 kDa cluster that has been LC-MS/MS identified as serum amyloid A and verified by both Western blot and ELISA as being up-regulated in the serum of ATL patients. This ID has confirmed the preliminary assignment from peak-cluster correlations (due to des-Arg modification [1]) and BNA (as well as BNA of higher-resolution spectra [13]).

A second peak, 20.5 kDa, was also consistently linked to clinical group when using local SNR normalization. The 20.5 kDa peak is a known truncation fragment (as is 20.8 kDa) of 21.1 kDa which was identified as Retinol-binding protein 4. This identification was supported by BNA in which the 20.8 kDa peak was always linked to the 20.5 kDa peak, and by correlation analysis for all normalization methods. High correlations were observed between the 20.5 kDa peak and both the 20.7 and 21.0 kDa peaks in normal patient samples (0.947–0.976 and 0.779–0.887, respectively). However, in leukemia patient samples these correlations dropped (to 0.655–0.773 and 0.206–0.483, respectively) which could suggest a different fragmentation pattern for the diseased state.

3.2 Prostate Cancer Data Set

TIC and Local SNR normalization applied to the prostate study data produced results that were consistent with those observed for the leukemia study data. Detailed results are presented in the Sup-Info. Feature selection using BNA and t-tests did not yield potential biomarkers. Given that the study design excluded sensitivity to serum PSA level and that serum is not a proximal fluid, this result is understandable. This indicates, as a minimum, our methods did not produce false positives.

4 Concluding Remarks

Automated signal processing tools were used to resolve integrated peak intensities over 2 – 200 kDa and 2 –100 kDa spectra obtained, respectively, from two clinical serum protein profiling studies. We have demonstrated a normalization method (SNR based on local noise) that removes systematic peak intensity variation across spectra without introducing erroneous correlations. Furthermore, this local method was found to be superior to methods based on division by a single factor per spectrum in improved outcome of automated BNA for selection of discriminating features. Residual correlations remaining after normalization contain meaningful information about chemical relationships that can aid in preliminary assignments of protein IDs.

Peaks consistently linked to clinical group, i.e. potential biomarker proteins, were detected in clinical data from the leukemia study using BNA. Naïve Bayesian classifiers based on these peaks did not exceed predicted error limits for lockbox data normalized using TIC and local SNR indicating strong predictive power. One consistently selected peak, 11.5 kDa, was shown to be part of a cluster with 11.7 kDa, based on persistent high correlations and BNA detection of dependencies, confirming previous preliminary assignments [13]. This cluster was identified using LC MS/MS as serum amyloid A and has been verified using ELISA and Western Blot tests. No PSA-independent biomarkers were detected in the prostate cancer data. By its design, the PCA study may provide a baseline test-case against false discovery of PSA-independent biomarker detection in serum.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH grant CA126118 and ARRA supplement award 3R01CA126118-04S1 from the Advanced Proteomics Platforms and Computational Sciences Program within the Clinical Proteomics Initiative of the National Cancer Institute. We thank INCOGEN, Inc for maintaining the database containing the leukemia study data and for the use of the VIBE software package used to access the database. We greatly appreciate Karl Kuschner's consultation on BNA and Krista White's work on protein identification.

Abbreviations

ATL	adult T-cell leukemia
BNA	Bayesian Network Analysis
BPH	benign prostatic hyperplasia
L1	BNA Level 1 – links to clinical group
L2	BNA Level 2 – links to L1 peaks
PCA	prostate cancer
PSA	prostate-specific-antigen
QC	quality control pooled serum
r	correlation coefficient
SNR	signal-to-noise
Sup-Info	Supporting Information

References

1. Hortin GL. The MALDI-TOF Mass Spectrometric View of the Plasma Proteome and Peptidome. Clin. Chem. 2006; 52:1223–1237. [PubMed: 16644871]
2. Lu M, Faull KF, Whitelegge JP, He J, et al. Proteomics and Mass Spectrometry for Cancer Biomarker Discovery. Biomark. Insights. 2007; 2:347–360. [PubMed: 19662217]

3. Huijbers A, Velstra B, Dekker TJA, Mesker WE, et al. Proteomic Serum Biomarkers and Their Potential Application in Cancer Screening Programs. *Int. J. Mol. Sci.* 2010; 11:4175–4193. [PubMed: 21151433]
4. Meuleman W, Engwegen JYMN, Gast MW, Beijnen JH, et al. Comparison of normalization methods for surface-enhanced laser desorption and ionization (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics.* 2008; 9:88. [PubMed: 18257918]
5. Randolph TW. Scale-based normalization of spectral data. *Cancer Biomark.* 2006; 2:135–144. [PubMed: 17192066]
6. Sauve, AC.; Speed, TP. Proceedings of the Genomic Signal Processing and Statistics workshop. Baltimore, MD, USA: 2004 May 26–27. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. <http://stat-www.berkeley.edu/users/terry/Group/publications/Final2Gensips2004Sauve.pdf>
7. Borgaonkar SP, Hocker H, Shin H, Markey MK. Comparison of Normalization Methods for the Identification of Biomarkers Using MALDI-TOF and SELDI-TOF Mass Spectra. *OMICS.* 2010; 14:115–126. [PubMed: 20141334]
8. Cairns DA, Thompson D, Perkins DN, Stanley AJ, et al. Proteomic profiling using mass spectrometry – does normalizing by total ion current potentially mask some biological differences? *Proteomics.* 2008; 8:21–27. [PubMed: 18095358]
9. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, et al. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics.* 2005; 5:4107–4117. [PubMed: 16254928]
10. Malyarenko DI, Cooke WE, Adam B, Malik G, et al. Enhancement of Sensitivity and Resolution of Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometric Records for Serum Peptides Using Time-Series Analysis Techniques. *Clin. Chem.* 2005; 51:65–74. [PubMed: 15550476]
11. Malyarenko DI, Cooke WE, Tracy ER, Drake RR, et al. Resampling and deconvolution of linear time-of-flight records for enhanced protein profiling. *Rapid Commun. Mass Spectrom.* 2006; 20:1670–1678. [PubMed: 16637003]
12. Malyarenko DI, Cooke WE, Bunai CL, Manos DM. Automated assignment of ionization states in broad-mass matrix-assisted laser desorption/ionization spectra of protein mixtures. *Rapid Commun. Mass. Spectrom.* 2010; 24:138–146. [PubMed: 19960505]
13. Kuschner KW, Malyarenko DI, Cooke WE, Cazares LH, et al. A Bayesian Network Approach to Feature Selection in Mass Spectrometry Data. *BMC Bioinformatics.* 2010; 11:177. [PubMed: 20377906]
14. Gatlin-Bunai CL, Cazares LH, Cooke WE, Semmes OJ, Malyarenko DI. Optimization of MALDI-TOF MS detection for enhanced sensitivity of affinity-captured proteins spanning a 100 kDa mass range. *J. Proteome Res.* 2007; 6:4517–4524. [PubMed: 17918874]
15. Semmes OJ, Cazares LH, Ward MD, Qi L, et al. Discrete serum protein signatures discriminate between human retrovirus-associated hematologic and neurologic disease. *Leukemia.* 2005; 19:1229–1238. [PubMed: 15889159]
16. Gatlin CL, White KY, Tracy MB, Wilkins CE, et al. Enhancement in MALDI-TOF MS analysis of the low molecular weight human serum proteome. *J. Mass. Spectrom.* 2011; 46:85–89. [PubMed: 21190259]
17. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.* 2005; 21:1764–1775. [PubMed: 15673564]
18. Karpievitch YV, Hill EG, Smolka AJ, Morris JS, et al. PrepMS: TOF MS data graphical preprocessing tool. *Bioinformatics.* 2007; 23:264–265. [PubMed: 17121773]
19. Tolson J, Bogumil R, Brunst E, Beck H, et al. Serum protein profiling by SELDI mass spectrometry: detection of multiple variants of serum amyloid alpha in renal cancer patients. *Laboratory Investigation.* 2004; 84:845–856. [PubMed: 15107802]

Statement of Clinical Relevance

Application of optimized signal processing tools leads to more than an order of magnitude increase in sensitivity, accuracy and mass-range for the current TOF-MS-based protein detection strategies used in high-throughput clinical proteomics. A normalization technique based on local noise levels across spectra, together with Bayesian network feature selection facilitates robust detection and identification of protein biomarkers in clinical data.

Figure 1.
Average of processed QC spectrum for the leukemia study acquired over two mass ranges.

Figure 2.
Histograms for peak correlation coefficients (r) for all peaks detected in QC data from the leukemia study, prior to and following normalization using TIC, de-trending, global SNR and local SNR. (173 spectra, 110 peaks)

Figure 3. Scatter plots of 66.3 versus 9.29 kDa peak intensities, prior to and after TIC and Local SNR normalization of QC data from the leukemia study. (173 spectra)

Figure 4. Peak CV histograms for all peaks detected in QC data from leukemia study, prior to and following normalization using TIC, de-trending, global SNR and local SNR. (173 spectra, 110 peaks)

Table 1

Variable selection and classification error for leukemia study data.

Normalization	BNA ^(a) L1 Peaks (kDa)	BNA ^(a) Error Rate (mean +/- SD)	Naive Bayes Classifier ^(b) Error Rate for	
			Training as Test Data	Test Data
None	11.5, 22.6, 25.8	0.1079 +/- 0.0179	0.0606	0.3030
De-Trended	11.5	0.0992 +/- 0.0160		
TIC	11.5	0.1386 +/- 0.0177	0.0758	0.1061
Global SNR	25.8	0.1020 +/- 0.0167		
Local SNR	11.5, 20.5, 25.8, 78.8, 132.1	0.0917 +/- 0.0157	0.0455	0.0909

^(a) Selection Rate > 90% for 100 iterations, 10-fold cross validation (132 patients)

^(b) Training data from 132 patients and test data from 66 patients.