RESEARCH ARTICLE

An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers

Yutaka Yasui,[1]* Dale McLerran,[1] Bao-Ling Adam,[2] Marcy Winget,[1] Mark Thornquist,[1] and Ziding Feng[1]

[1] *Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center,*
1100 Fairview Avenue N, MP 702, Seattle, WA 98109-1024, USA
[2] *Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School,*
700 Olney Road, Norfolk, VA 23507, USA

Received 25 July 2002; revised 20 December 2002; accepted 20 December 2002

Discovery of "signature" protein profiles that distinguish disease states (eg, malignant, benign, and normal) is a key step towards translating recent advancements in proteomic technologies into clinical utilities. Protein data generated from mass spectrometers are, however, large in size and have complex features due to complexities in both biological specimens and interfering biochemical/physical processes of the measurement procedure. Making sense out of such high-dimensional complex data is challenging and necessitates the use of a systematic data analytic strategy. We propose here a data processing strategy for two major issues in the analysis of such mass-spectrometry-generated proteomic data: (1) separation of protein "signals" from background "noise" in protein intensity measurements and (2) calibration of protein mass/charge measurements across samples. We illustrate the two issues and the utility of the proposed strategy using data from a prostate cancer biomarker discovery project as an example.

INTRODUCTION

With recent advances in mass spectrometry technologies, it is now possible to study protein profiles over a wide range of molecular weights in small biological specimens [1]. A key research step in translating these technological advancements into clinical utilities is the identification of "signature" protein profiles that distinguish disease states (eg, malignant, benign, and normal) or experimental conditions (eg, treated versus untreated by a drug of interest). For example, a discovery of disease-specific protein profiles could facilitate early detection of the disease and, consequently, contribute importantly towards improving patients' prognosis and survival.

Protein data generated from mass spectrometers have complex features due to complexities in both biological specimens and interfering biochemical/physical processes of the measurement procedure. They are also large in size, generally in the order of tens of thousands of measurement points per sample. Making sense out of such high-dimensional complex data is challenging and necessitates the use of a systematic data analytic strategy. Specifically, there are three major issues in the analysis of mass-spectrometry-generated protein data that need to be resolved effectively by the systematic strategy: (1) separation of protein "signals" from background "noise" in protein intensity measurements; (2) calibration of protein mass/charge measurements across samples; and (3) construction of "signature profiles," as combinations of multiple mass/charge points, that distinguish disease states or experimental conditions.

This paper is concerned with the first two of the three issues in the analysis of mass-spectrometry-generated protein data. We propose a systematic data processing method for separating signals (protein intensity peaks) from noise, and for calibrating mass/charge values of proteins that may fluctuate slightly at random across samples; for approaches to the third data analytic problem, several approaches have been proposed [2, 3, 4, 5, 6].

MATERIALS AND METHODS

In this section, we first describe the prostate cancer biomarker discovery project [4, 5] to illustrate the type of research settings of interest. The project's data are used to explain the signal identification and calibration issues. We then present our proposed data processing method that addresses these issues.

The prostate cancer biomarker discovery project

The Department of Microbiology and Molecular Cell Biology and Virginia Prostate Center of the Eastern Virginia Medical School (EVMS) have been conducting a biomarker discovery project on prostate cancer with a goal of identifying serum protein biomarkers of the disease. This project is part of a large research consortium, the Early Detection Research Network [7, 8], funded by the National Cancer Institute. The basis for the protein-based early detection of cancer is the concept that a transformed cancerous cell and its clonal expansion would result in up- (or down-) regulation of certain proteins; our aim is to identify such early molecular signals of prostate cancer by measuring protein profiles in serum.

In this project, serum samples of 386 subjects were retrieved from the serum repository of the EVMS Virginia Prostate Center, approximately equally from four disease groups: late-stage prostate cancer ($N = 98$); early-stage prostate cancer ($N = 99$); benign prostatic hyperplasia (BPH) ($N = 93$); and normal controls ($N = 96$). The four disease groups were defined as follows: prostate cancer cases had a positive biopsy that was staged A or B (early-stage) or C or D (late-stage) and had a prostate specific antigen (PSA) concentrations greater than 4 ng/ml; BPH patients had PSA values between 4 ng/ml and 10 ng/ml, low PSA velocities, and at least two negative biopsies; and normal controls were aged 50 or older (ie, the same age range as cancer and BPH patients), had a PSA level less than 4 ng/ml, and had a normal digital rectal exam.

Each of the retrieved serum samples was assayed at the EVMS for protein expression by the surface-enhanced laser desorption/ionization (SELDI) ProteinChip Array technology [1, 2, 3, 4, 8, 9, 10, 11, 12, 13] of Ciphergen Biosystems, Inc, 6611 Dumbarton Circle, Fremont, CA 94555. The SELDI technology is a time-of-flight mass spectrometry with a special ProteinChip® Array whose surface captures proteins using chemically or biologically defined protein-docking sites. Proteins are captured on the chip surface, purified by washing the surface, and crystallized with small molecules called "matrix" or "energy-absorbing molecules" whose function is to absorb laser energy and transfer it to proteins. Energized protein molecules fly away from the surface into a time-of-flight tube where the time for the molecules to fly through the tube is a function of the molecular weight and charge of the protein. A detector at the end of the tube measures the "intensity" of proteins at each discrete time of flight and outputs about 48 000 data points of time of flight, intensity pairs. Each discrete time of flight corresponds uniquely to a ratio of the molecular weight of a protein to the number of charges introduced by the ionization. SELDI output, therefore, produces about 48 000 data points of mass/charge, intensity pairs. Our analyses used 16 898 data points per sample covering the mass/charge range of 2 000–40 000. Figure 1 shows an example of SELDI output from the first subject in the normal control group of the prostate cancer biomarker discovery project.

The two data analytic issues

The first data analytic issue is the mathematical definition of "peaks" in protein intensity, that is, the identification of "signals" separated from "noise" in protein intensity measurements. Defining protein intensity peaks mathematically provides two advantages. First, it reduces the dimensionality of data from tens of thousands of data points to a more manageable size (eg, less than a thousand data points). Second, perhaps more importantly, it clarifies the interpretation of "signature" protein profiles, the end products of the analysis. Specifically, protein intensity peaks and their heights at certain mass/charge values indicate the presence and the approximate amount of corresponding proteins or peptides in the specimen being analyzed. Without this data reduction step, the "signature" protein profiles that we obtain may not have a clear interpretation: a signature profile can be any pattern of protein intensity measurements and may not correspond to any protein intensity peaks.

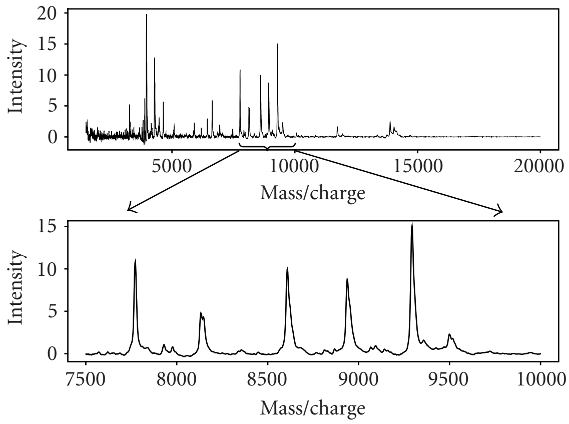

FIGURE 1. An example of SELDI output: the top panel shows the protein intensity measures (y-axis) in the range of mass/charge values below $20,000$ (x-axis) and the bottom panel zooms into a subregion covering 7 500–10 000 mass/charge values.

To illustrate this first data analytic issue by a concrete example, consider the protein intensity measurements shown in the bottom panel of Figure 1. There are five large peaks (signals) that are visually evident in the plotted range of 7 500–10 000 mass/charge values. There are, however, other smaller peaks that are less evident as to whether or not they represent signals. A good mathematical definition of peaks would capture, at minimum, the five clear peaks, and possibly, less evident ones, but would also have a high level of specificity such that the rest of the data points would not be identified as peaks.

The second data analytic issue is the calibration of mass/charge measurements across multiple samples. This issue can be explained clearly by an example. Figure 2 shows the SELDI output from the first four subjects in the normal control group of the prostate cancer biomarker discovery project. In the left panel of Figure 2, it appears that, at least, the five visually apparent peaks, including the one marked by "×," and some less-evident peaks are aligned well in the direction of the mass/charge axis across the four samples. When we zoom into the small region near the peak marked by "×" (the right panel of Figure 2),

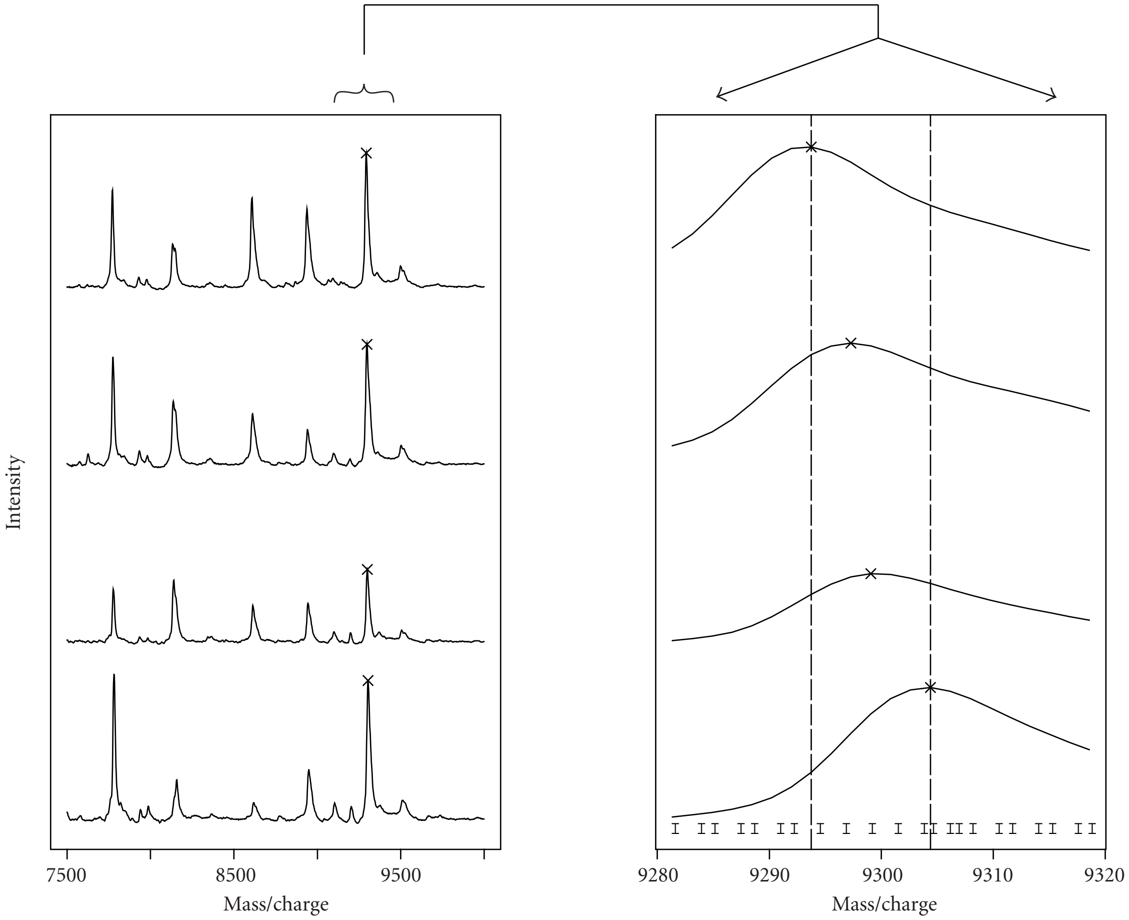

FIGURE 2. A set of SELDI output from four subjects: the left panel shows the protein intensity measures in the range of 7 500–10 000 mass/charge values and the right panel zooms into a small region corresponding to a visually apparent peak. The right panel shows slight shift in the mass/charge value of an identical peak across the subjects.

however, it becomes apparent that there is slight shift across the four samples with respect to the mass/charge values of the peak: there is an approximately 10-dalton shift between the first subject at the top and the fourth subject at the bottom. The measurement points in this small region are indicated by "|" at the bottom of the right panel of Figure 2: the 10-dalton shift corresponds to a 6-measurement-point shift in the mass/charge axis.

Although the magnitude of the shift is small, the inconsistent mass/charge values across samples for an identical peak present a challenge in data analyses: an identical peak is labeled by different mass/charge values across samples. Consider, in the right panel of Figure 2, the protein intensity at the mass/charge value of the peak from the first subject, that is, approximately at $9,294$ daltons. At this mass/charge value, the protein intensity for the second subject is nearly equal to the intensity value at the peak. For the third and fourth subjects, however, the protein intensities at the mass/charge value are approximately half and one quarter, respectively, of the intensities at the peaks. Thus, even though the magnitude of shifting is small, assessments of protein intensities by the original mass/charge values of the SELDI output would be greatly misleading.

The proposed solutions to the two data analytic issues

Our proposal for the first data analytic issue, the mathematical definition of "peaks," is to define peaks by judging, at each mass/charge point, whether or not the protein intensity at that point is the highest among its nearest $\pm N$-point neighborhood set, nearest in the mass/charge-axis direction; if it is the highest, that point is defined as a peak. We initially considered various values of N and chose $N = 10$ by trial and error in order to be on the inclusive side in classifying peaks; see also our previous discussion on the selection of N under a slightly different setting [6]. If a more conservative definition is preferred, the value of N can be increased.

Our proposal for the second data analytic issue, the calibration of mass/charge measurements across samples, is to replace the original mass/charge values of all peaks with a set of calibrated mass/charge values. To describe our algorithm of the proposed method, we first introduce some terms/concepts that are helpful. We call a range of potential mass/charge shifting from a mass/charge point as the "*window of potential shift*" for that mass/charge point; and refer to the set of calibrated mass/charge values as the "*new mass/charge set*." Note that the window of potential shift for a mass/charge point, say P, contains the mass/charge point P itself. Based on quality-control experiments by the manufacturer of SELDI machines, it is known that the window of potential shift for a mass/charge point is approximately ±0.1–0.2% of the mass/charge value of that point; we used 0.2% in the current analysis.

The algorithm is initiated by calculating, at each mass/charge point, the total number of peaks, *in all samples*, that are within the window of potential shift for the mass/charge point. The mass/charge point that has the highest total number of peaks (summing over all samples) within its window of potential shift is entered into the new mass/charge set as a calibrated mass/charge value, and all the mass/charge points that are within the window of potential shift for this point are removed from the subsequent steps of the algorithm. Then, the above procedure is repeated (ie, finding the point, from the remaining points, that has the highest total number of peaks within its window of potential shift, entering it into the new mass/charge set, and removing the mass/charge points that are within its window of potential shift from the subsequent steps of the algorithm) until all peaks are exhausted from every sample.

The end product of this repeated procedure is the new mass/charge set. The final step of the algorithm is to construct a calibrated dataset that consists of protein intensity measures of each sample that correspond to the points in the new mass/charge set. For each sample, i, and for each point in the new mass/charge set, j, we propose to take the maximum protein intensity measure of the sample i, among the protein intensity measures corresponding to the window of potential shift for the point j, as the protein intensity measure Y_{ij} at the calibrated mass/charge point j. The final calibrated dataset is $\{Y_{ij}\}$ whose elements represent protein intensity measures indexed by the sample number i and the calibrated mass/charge value j.

An application of the calibrated dataset to biomarker discovery

To illustrate the utility of the calibrated dataset produced by the proposed method, we applied it to a construction of "signature profiles" of disease states in the prostate cancer biomarker discovery project. The 386 serum samples of the project were separated by a stratified random sampling into "test data" (a total of 60, 15 samples from each of the four disease states: late- and early-stage prostate cancer; BPH; and normal) and "training data" (the remaining 326 samples). The training data were used to construct a calibrated dataset, from which signature profiles were derived for classifying the three disease states of interest (ie, cancer, BPH, and normal). To test the performance of the derived signature profiles independently from the training data, the test data was used as follows. First, a calibrated *test* dataset $\{Z_{ij}\}$ was constructed by setting, for each test sample i and each calibrated mass/charge value j that appears in the derived signature profiles, the maximum protein intensity measure Y_{ij} within the window of potential shift for j as Z_{ij}. Second, the signature profiles derived from the training data were applied to the calibrated test dataset to classify each test sample into the three disease states. Finally, the classification errors in the test data were assessed comparing the classified disease states with the true disease states. The stratified sampling that created the training and test data was conducted by a statistician (DM) who received the data from the EVMS laboratory, and the disease states of the test samples were blinded to all data analysts.

In the analysis of the training data, the protein-intensity measure, Y_{ij}, for sample i at calibrated mass/charge value j, was transformed because of its heavily skewed distribution. The transformed protein intensity measure, T_{ij}, is given by

$$T_{ij} = \ln\left(Y_{ij} - c_j + 1\right) - s_i, \tag{1}$$

where c_j is the minimum protein intensity measure at the calibrated mass/charge value j among all training samples, which makes the logarithmic transformation possible, and s_i is the mean value of $\ln(Y_{ij} - c_j + 1)$ of the sample i across all calibrated mass/charge values. The subtraction of s_i aims to remove the sample-specific mean protein intensity since a number of sample-specific factors could modify the amounts and measurements of proteins across samples. The same transformation was also employed in the analysis of the test data $\{Z_{ij}\}$.

The signature profiles (classifiers of the disease states) were constructed using two logistic regression models: one for classifying cancer/BPH versus normal and the other classifying cancer versus BPH. The two logistic regression analyses used the respective disease states as outcome variables and transformed protein intensity measures $\{T_{ij}\}$ as potential covariates, selecting only those with significant associations with the disease states (at $P = .0001$ and $P = .0005$ levels, resp.) by a forward variable selection method.

RESULTS AND DISCUSSION

Results

Figure 3 shows the peaks identified by the proposed peak identification method. Our simple mathematical definition of peaks captured both visually apparent and some less-evident peaks. The number of peaks identified per sample was similar across the four disease states: the median (range) of 469 (361–571), 463 (389–596), 467 (367–555), and 444 (390–559) for the groups of late-stage

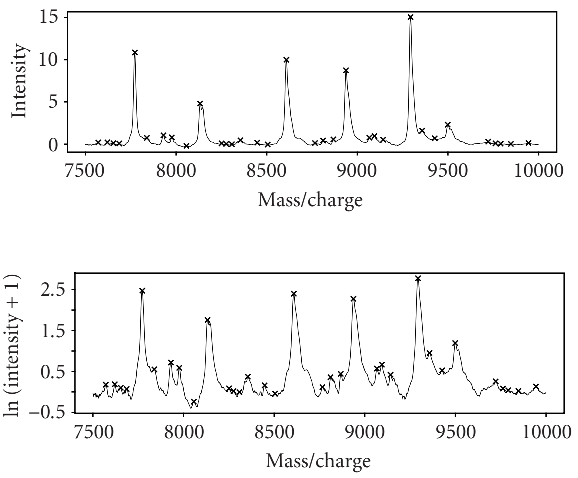

Figure 3. Peaks identified by the proposed peak identification method, marked by "×," in the range of 7, 500–10, 000 mass/charge values, corresponding to the bottom panel of Figure 1. The top panel shows the original protein intensity measures in the y-axis, while the bottom panel shows transformed protein intensity measures in the y-axis for ease in examining the peaks.

cancer, early-stage cancer, BPH, and normal control, respectively.

Figure 4 shows calibration of one of the visually apparent peaks in Figure 2 by the proposed method. The first four samples of the normal control group had peaks that were slightly shifted in the direction of the mass/charge axis around 9 300 mass/charge value. The calibrated mass/charge value corresponding to these peaks was 9 306.2 and, as shown in Figure 4, the four previously shifted peaks are now lined up at this calibrated mass/charge value with the original protein-intensity measures being used as $\{Y_{ij}\}$. The intensity values at 9, 306.2 mass/charge value changed from 7.87, 9.90, 6.40, 15.18 before the calibration to 8.61, 10.73, 7.04, 15.59 after the calibration in the four samples.

After the calibration, the number of mass/charge values in the new mass/charge set was 957. This represents a considerable reduction of data from 16 898 points per sample to 957 (5.7% of 16 898) in the range of 2 000–40 000 mass/charge values. Figure 5 shows the number of mass/charge values in the new mass/charge set according to their values. The number in the new mass/charge set was highest in the smallest values of mass/charge and monotonically decreased for larger mass/charge values.

The calibrated dataset was then used in the logistic regression analysis to construct signature profiles of three disease states (cancer, BPH, and normal). With the 957 log-transformed protein intensity measures as potential covariates, the forward selection method identified four calibrated mass/charge values that were significantly associated with the classification of cancer/BPH versus normal at $P = .0001$ level. Similarly, seven calibrated mass/charge values were selected into the model for the classification of cancer versus BPH at $P = .0005$ level. Note that the four and seven selected mass/charge values were those at which some samples showed peaks in protein intensity. Based on fitted probabilities from the two logistic regression models, the 60 test data samples were classified into the three disease states. Of the 60 test data samples, 55 (91.6%) were correctly classified, suggesting the high utility of the calibrated dataset, even with this simple classifier construction method using the standard forward selection logistic regression analysis.

Table 1. Test data classification results of the logistic-model-based classifiers constructed using the calibrated training dataset.

Predicted by models	True disease state		
	Cancer	BPH	Normal
Cancer	27	2	0
BPH	1	13	0
Normal	2	0	15

Discussion

Previously in this project, the peak identification and mass/charge value calibration were performed manually, taking a significant amount of human effort and time [4]. The proposed data processing method was motivated to automate the human processing of the SELDI output by the use of computers. It mimics the steps of the previous manual processing and aims to eliminate potential human errors in dealing with the high-dimensional complex data. The excellent performance in classification shown in Table 1 suggests the high utility of the data produced by the proposed data processing method.

There are important advantages in separating the data processing stage, as proposed here, from the subsequent signature profile construction stage. First, the proposed method can be applied with complete blinding to the disease states of samples, ensuring an unbiased data processing across samples. Neither the peak identification nor the mass/charge-axis calibration of our proposed method requires knowledge of the disease state of each sample. Second, by separating the two stages that have distinct data analytic issues, we are able to consider various targeted approaches for resolving the stage-specific data analytic issues.

A limitation in our proposed method for the calibration of mass/charge values is that the calibration procedure depends on the dataset being analyzed, that is, the sets of calibrated mass/charge values from multiple datasets may not agree. This is perhaps not a critical issue at the biomarker discovery stage of research, such as the prostate cancer biomarker discovery project discussed here, since the data analysis for the biomarker discovery would use a single dataset. If the data are measured using multiple SELDI machines, either at one laboratory or at multiple laboratories, the dataset-dependent nature of

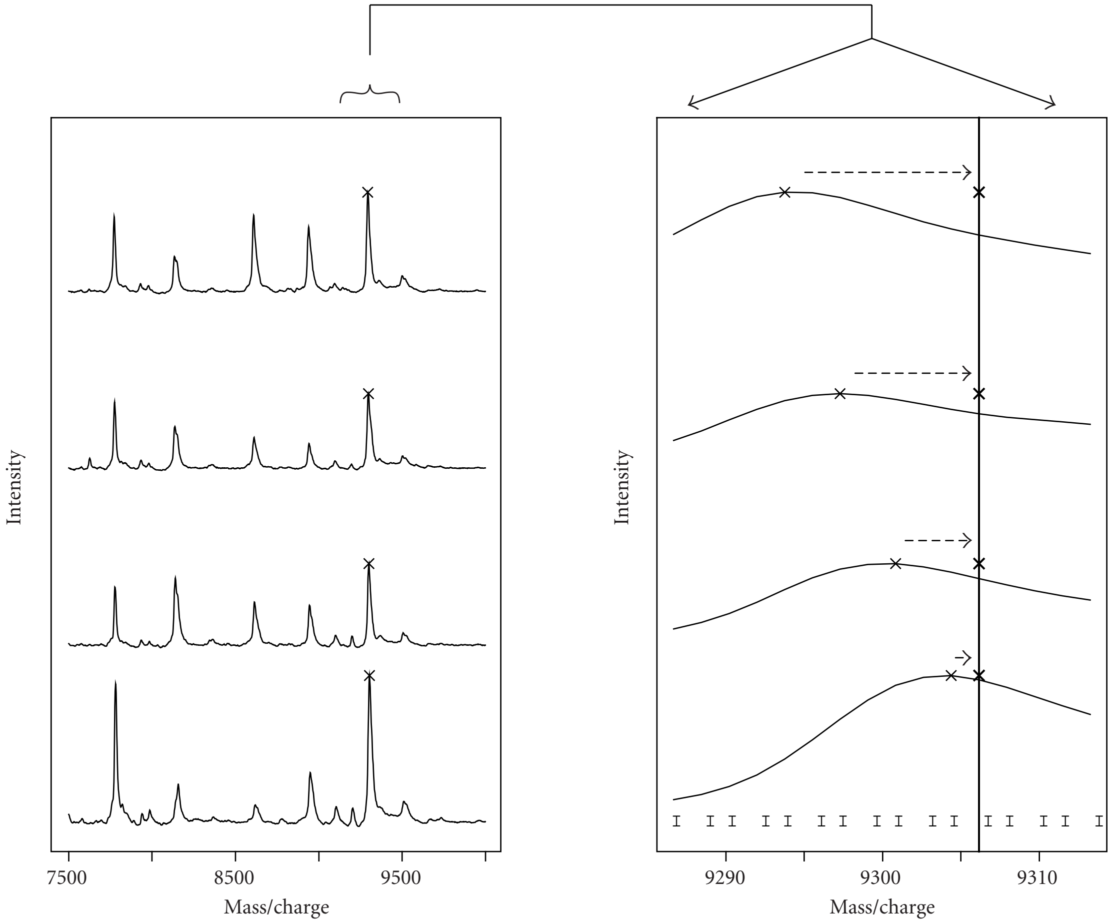

Figure 4. Calibration of a visually apparent peak in Figure 2 by the proposed method.

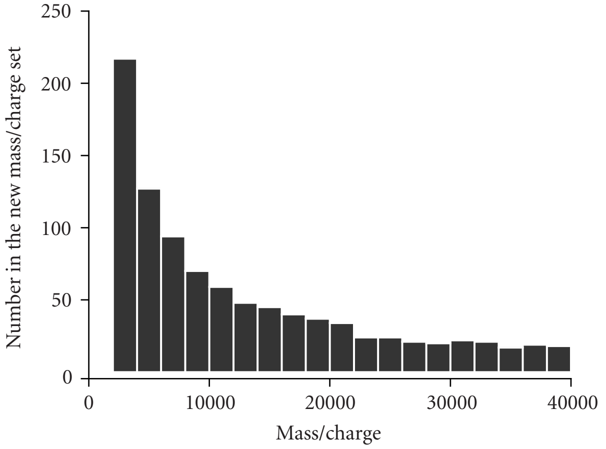

Figure 5. Number of mass/charge points in the new calibrated mass/charge set according to their mass/charge values.

the proposed method needs to be considered carefully. If datasets generated by multiple SELDI machines are combined, additional sources of variations are introduced (eg, between machine and laboratory variations). The "window of potential shift" of a mass/charge value in the combined dataset would, therefore, be expected to be larger than the ±0.1–0.2% considered here, that was based on the quality-control data of the SELDI manufacturer. It is necessary to either use a wider window of potential shift or add an extra step in the method to minimize the additional variation when combining datasets generated by multiple SELDI machines.

A potential improvement of the proposed method is to make the mathematical definition of peaks such that it copies closely the thought process of experienced mass spectrometry experts in identifying peaks. Although our simple definition appears reasonable and functional, it would certainly enhance the method if a peak identification procedure similar to that of experts could be implemented. For example, a learning algorithm can be applied to a large dataset that is read and peak-identified by experts so that details of experts' procedures may be recognized for possible implementation in the mathematical definition.

In summary, we proposed an automatic data processing procedure for the peak identification and mass/charge-axis calibration problems for mass-spectrometergenerated protein measures. Our procedure is easy to implement and appears to work effectively as evidenced in the excellent classification performance by the resulting calibrated dataset. There are points to improve in the proposed method and additional issues to resolve when it is applied to multiple datasets generated by more than one SELDI machine. We hope to resolve these issues in our future research.

ACKNOWLEDGMENT

This research was supported by Grant U01-CA86368 from the National Cancer Institute.

REFERENCES

[1] Srinivas PR, Srivastava S, Hanash S, Wright GL Jr. Proteomics in early detection of cancer. *Clin Chem*. 2001;47(10):1901–1911.

[2] Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359(9306):572–577.

[3] Ball G, Mian S, Holding F, et al. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*. 2002;18(3):395–404.

[4] Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*. 2002;62(13):3609–3614.

[5] Qu Y, Adam BL, Yasui Y, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem*. 2002;48(10):1835–1843.

[6] Yasui Y, Pepe M, Thompson ML, et al. A Data-Analytic Strategy for Protein-Biomarker Discovery: Profiling of High-Dimensional Proteomic Data for Cancer Detection. *Biostatistics* 2003;4(3):449–463.

[7] Srivastava S, Kramer BS. Early detection cancer research network. *Lab Invest*. 2000;80(8):1147–1148.

[8] Verma M, Wright GL Jr, Hanash SM, Gopal-Srivastava R, Srivastava S. Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers. *Ann N Y Acad Sci*. 2001;945:103–115.

[9] Wright GL Jr, Cazares LH, Leung SM, et al. Proteinchip® surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostatic Dis*. 1999;2(5/6):264–276.

[10] Rubin RB, Merchant M. A rapid protein profiling system that speeds study of cancer and other diseases. *Am Clin Lab*. 2000;19(8):28–29.

[11] Adam BL, Vlahou A, Semmes OJ, Wright GL Jr. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics*. 2001;1(10):1264–1270.

[12] Paweletz CP, Trock B, Pennanen M, et al. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis Markers*. 2001;17(4):301–307.

[13] Rosty C, Christa L, Kuzdzal S, et al. Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology. *Cancer Res*. 2002;62(6):1868–1875.

* Corresponding author.
E-mail: yyasui@fhcrc.org
Fax: +1 206 667 5977; Tel: +1 206 667 4459