Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Utlization of Human Expert Techniques for Detection of Low-Abundant Peaks in High-Resolution Mass Spectra

Grzegorz M Boratyn, *Member IEEE*, Michael L Merchant, and Jon B Klein

Abstract— **Interpretation and classification of mass spectra is usually performed using a list of peaks as their mathematical representation. This fact makes peak detection a bottleneck of mass spectra analysis, since quality and biological relevance of any results strongly depends on the accuracy of peak detection process. Many algorithms utilize intensity to differentiate between peaks and noise and thus bias the detection process to the high abundant peaks. However important information may also be contained in the lower-intensity peaks that are more difficult to discover. We present an algorithm specifically designed for detection of low-abundant peaks.**

I. INTRODUCTION

Mass spectrometry (MS) techniques are being used increasingly for proteomic analysis of biological fluids or tissue samples in order to identify disease-related proteins or peptides that can be used as biomarkers for early diagnosis, disease progression or response to treatment [1], [2], [3], [4], [5], [6], [7].

The MS instruments commonly utilized for analysis of biological samples measure the time-of-flight (TOF) of accelerated ions produced by matrix assisted laser desorption ionization (MALDI) or surface enhanced laser desorption ionization (SELDI), in order to estimate amounts (or intensities) of particles with specific mass-to-charge ratio. A typical mass spectrum, is a rough estimate of mass-to-charge ratio distribution in the analyzed sample. Proteins and peptides are represented in the spectrum by local maxima, often called peaks.

Interpretation and classification of mass spectra is commonly performed using a list of peaks as their mathematical representation. Because quality and biological relevance of statistical analysis, classification, or clustering strongly depends on the quality of analyzed data, peak detection is the bottleneck of the analysis and hence needs to be done with care.

Peaks are often detected as maxima of intensity bins [8], maxima in a local neighborhood higher than the average intensity in a broader neighborhood [6], high intensity values with a given number of progressively lower intensities on both sides of potential peak [9], intensity sequences that correlate with a Gaussian template [10], local maxima of a wavelet-smoothed spectrum [11], or intensities exceeding locally estimated noise level [12], [13].

G. M. Boratyn and M. L. Merchant are with the Kidney Disease Program and Clinical Proteomics Center, University of Louisville, Louisville, KY

J. B. Klein is with Faculty of Department of Medicine, University of Louisville and Veterns Adminidtration Medical Center, Louisville KY

Corresponding author: G. M. Boratyn, phone: 502 852 4586, fax: 502 852 4384, e-mail: greg.boratyn@louisville.edu

Intensity pattern in relatively narrow neighborhood of peak centroid is the most common feature utilized by automatic methods to differentiate between peaks and noise. While abundance is very distinctive feature for peaks that are much higher than the level of noise, the task becomes difficult for low abundant peaks. In high-resolution mass spectra proteins and peptides are represented by a series of isotopic peaks, in case of low-abundant peaks very narrow, similar to noisy spikes in intensity, and thereby difficult to detect.

While detecting peaks, a human expert analyzes a broader neighborhood of peak centroid, searches for isotopic peaks, and similar pattern of local intensities in other spectra. We propose a peak detection algorithm that utilizes those techniques. The presented algorithm combines information about local variance of intensities with likelihood that there are other isotopic peaks.

The paper is organized as follows. The next section presents formal representation of mass spectra and the proposed algorithm. The results of experiments that assess performance of our method are described in Sec. III. Sec. IV provides discussion and conclusions.

II. ALGORITHM DESCRIPTION

A. Representation of Mass Spectrum

Let us assume that we observe n spectra $f_i(t)$ of the following form (similarly to [11], [7]):

$$f_i(t) = B_i(t) + N_i S_i(t) + \epsilon_f(t), \tag{1}$$

where $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, m$ is the clock tic of the TOF detector that is used to compute mass-to-charge ratio M. $B_i(t)$ denotes the systematic artifact, known as baseline. Estimation of baseline is beyond this scope of this paper. Methods that compute baseline can be found in [14], [12], [11]. $S_i(t)$ represents peaks (the true signal), scaled in each spectrum by N_i. Noise is represented by $\epsilon(t)$. We assume that $\epsilon(t)$ is distributed according to the normal distribution $N(0, \sigma_f(t))$ with mean 0 and standard deviation $\sigma_f(t)$, which is a smooth function of t.

The relationship $f_i(M)$ is of larger scientific interest, because masses of differential peaks are utilized for identification of proteins and peptides. The mass-to-charge ratio is a function of TOF. We assume that $M_i(t)$ is of the following form:

$$M_i(t) = A_i + c(t) + \epsilon_z(t), \tag{2}$$

where A_i is the systematic shift with respect to one spectrum in the studied group, $c(t)$ denotes the relationship between M and TOF (model of $c(t)$ can be found in [15]), and

$\epsilon_z(t)$ represents Gaussian noise. Often $M_i(t)$ is estimated by software that operates mass spectrometer. In this paper we assume that each $M_i(t)$ is given.

Prior to detection of peaks all spectra are normalized with respect to their sum of intensities:

$$\hat{f}_i(t) = \frac{f_i(t)}{\sum_{t'=1}^{m} f_i(t')}. \tag{3}$$

This normalization is believed to be valid, because the sum of intensities corresponds to the total number of particles per spectrum that hit the TOF detector [15].

B. Peak Detection

The presented method examines the likelihood that each of the local intensity maximum is a peak. The maxima can be simply detected as locations where the first derivative of spectrum changes sign and the second derivative is negative.

We assume that all spectra belong to one experimental group and we aim at selecting the set of peaks that is representative of the whole group, i.e. peaks that appear in the majority of the studied spectra. Thus we select a target spectrum f_{i^*} that is likely to contain the smallest number of true peaks among the group as follows:

$$i^* = \arg \min_{i=1,\ldots,n} \sum_{t=1}^{m} \hat{f}_i(t). \tag{4}$$

Let M_{ij} denote mass-to-charge of the j-th maximum in the i-th spectrum. The true mass-to-charge value of a potential peak is estimated by averaging the closest intensity maxima from all spectra:

$$\bar{M}_j = \frac{1}{n}\left(M_{i^*j} + \sum_{i=1,\, i\neq i^*}^{n} M_{i\hat{j}} \right), \tag{5}$$

where

$$\hat{j} = \arg \min_{j'=1,\ldots,m} |M_{i^*j} - M_{ij'}|. \tag{6}$$

Thus the set of potential peaks $Q = \{\bar{M}_j\}$ is created. The following subsections present methods that assess the likelihood that each $\bar{M}_j$ is a true peak.

1) Change of intensity variance: The Change of Intensity Variance (CIV) approach is based on the assumption that the intensity variance σ_f is a smooth function of t and M. The CIV method measures change of intensity variance between two consecutive parts of a spectrum. Let us divide the spectra into short disjoint parts f_i^k of length w:

$$f_i^k = (f_i(t_i^k), f_i(t_i^k + 1), \ldots, f_i(t_i^{k+1} - 1)), \tag{7}$$

where $k = 1, 2, \ldots, K$, t_i^k indicates the first element of the k-th part and is given by:

$$\begin{aligned} &t_i^0 = 1,\ t_i^k = t_i, \text{such that} \\ &M_i(t_i + 1) > \min_{i=1,\ldots,n} M_i(t_i^{k-1}) + w, \end{aligned} \tag{8}$$

so that the parts are aligned across spectra with respect to, and have constant length in terms of M. Mass-to-charge ratios are also partitioned along with intensities such that:

$$M^k = \frac{1}{n} \sum_{i=1}^{n} M_i(t^k). \tag{9}$$

Because the noise variance is a smooth function of t, we assume that variance of f_i^k should also change smoothly as a function of k, and sudden changes indicate that the k-th part contains peaks. Small w is required for good discrimination between noise and peaks. However reliable estimation of statistical properties usually requires large number of sample points. Therefore corresponding parts f_i^k from all spectra are combined into one set f^k as follows:

$$f^k = \{f_i^k - \bar{f_i^k}\}_{i=1}^{n} \tag{10}$$

where $\bar{f_i^k}$ represents the mean of f_i^k. Let $var(f^k)$ indicate variance of f^k, then $|\Delta var(f^k)|$ given by:

$$|\Delta var(f^k)| = |var(f^k) - var(f^{k-1})| \tag{11}$$

can be used to measure likelihood that there is a peak between M^{k-1} and M^k. The measure (11) can be computed for each j-th potential peak in Q as follows:

$$V(j) = |\Delta var(f^k)|, \text{ s.t. } M^{k-1} \leq \bar{M}_j \leq M^k. \tag{12}$$

Then the probability that the j-th potential peak in Q is a true peak, given $V(j)$ is given by:

$$P(j|V(j)) = \exp(-(\max_j V(j) - V(j))^2/\sigma_v), \tag{13}$$

where $\sigma_v = 3 \max_j V(j)$ so that $P(j|\max_j V(j)) = 1$ and $P(j|0) \to 0$. Our preliminary experiments revealed that the best w is equal to 2 Da.

2) Isotopic peaks: Existence of other Isotopic Peaks (IP) is one of the main feature of true peaks utilized by human experts. Therefore incorporating this information into the detection process should result in better performance. Because of natural occurrence of isotopic elements, a peptide in the mass spectrum is represented by a series of peaks with masses increased by $1, 2, 3, \ldots$ [Da]. Depending on the mass and abundance, 3 to 8 isotopic peaks can usually be detected by human eye in the mass spectrum.

Let $P(I_j)$ denote the probability that there is one isotopic peak on the left and one to the right side of $\bar{M}_j \in Q$. This probability can be computed as follows:

$$\begin{aligned} P(I_j) = \frac{1}{2|Q|} \sum_{l=1}^{|Q|} P(l)(P(D|M_j - M_l - 1) + \\ + P(D|M_j - M_l + 1)), \end{aligned} \tag{14}$$

where $|Q|$ represents the number of local maxima in spectrum f_{i^*}, $P(l)$ denotes the probability that the $\bar{M}_l \in Q$ is a true peak, computed, for example, as $P(l|V(l))$ with (13), $P(D|M)$ is the probability that M is the correct location of potential isotopic peak and is given by:

$$P(D|M) \exp(-(M^2/\sigma_I), \tag{15}$$

where σ_I should be close to standard deviation of ϵ_z.

Assuming statistical independence of the following events: 1) the $\bar{M}_j \in Q$ is a true peak given $V(j)$, and 2) there are other peaks at locations $M_j - 1$ and $M_j + 1$, the probability that both events are true is given by:

$$P(j|V(j), I(j)) = P(j|V(j))P(I(j)). \tag{16}$$

Thus the $P(j|V(j))$ is updated by the information provided by $P(I_j)$. The resulting probability $P(j|V(j), I(j))$ can be again utilized in (14) as $P(l)$ and updated again with the new $P(I_j)$ so that isotopic peaks of isotopic peaks are also taken under consideration. Hence we propose the following iterative procedure for updating the probability that the j-th maximum is a true peak:

$$\begin{aligned} P^0(j) &= P(j|V(j)) \\ P^r(j) &= P^{r-1}(j)P(I_j), \end{aligned} \tag{17}$$

where $P^r(\cdot)$ denotes the value of $P(\cdot)$ in the r-th step of the algorithm. In each r-th iteration the likelihood $P^{r-1}(j)$ is updated by the information about isotopic peaks at locations $M_j - r$ and $M_j + r$. If a majority of peptides in the spectrum are represented by a series of $2r^*$ isotopic peaks, then according to $P^{r^*+1}(j)$ peaks are indistinguishable from noise. Therefore the detection performance increases for $r = 0, 1, \ldots r^*$, and decreases for steps $r^* + 1$, $r^* + 2, \ldots$. Hence the best detection performance is yielded by $P^{r^*}(j)$. The lower bound for the number of isotopic peaks in the mass spectrum depends on mass and abundance of the chemical compound. Thus parameter r^* can be estimated theoretically based on mass range of the spectrum, or experimentally utilizing a spectrum that contains known masses.

III. Experimental Results

The proposed peak detection procedure was examined on simulated mass spectra generated by the virtual mass spectrometer described in [15]. Spectra for the following masses were generated: 900, 950, 1001, 1023, 1050, 1200, 1250, 1400, 1503, 1600, 1750, 2000 [Da]. The number of molecules was chosen as 500 for all masses. The resulting centroid intensity for mono-isotopic peaks are equal to 275 and 255 for the smallest mass and largest mass, respectively. Each of the simulated peptides is represented by a series of 5 to 10 isotopic peaks. Other parameters of the virtual mass spectrometer were set to the following values: length of the drift tube = 1 m, acceleration voltage = 20000V, mean initial velocity = 350 m/s, standard deviation of initial velocity = 5 m/s, time resolution = 4E-09 s. Software defaults were used for all other parameters.

The virtual mass spectrometer produces an ideal mass spectrum with no noise. Mass-to-charge ratios, corresponding to all peak centroids present in the spectrum were labeled. An artificial set of mass spectra was generated according to (1) and (2), similarly to [7], with the following parameters: $B_i(t) = N(178, 34)$, $N_i = N(1, 0.3)$, $\epsilon_m = N(0, 0.02)$, several values of n and $\sigma_f(t)$ were utilized. The baseline is assumed to be flat, because the non-linear baseline can be estimated and subtracted with baseline correction algorithms. $\sigma_f(t)$ is assumed to be constant. The parameters were estimated from MALDI mass spectra, obtained from blood serum, collected from patients that took part in the coronary arteries disease study in the University of Louisville. The mass spectra were acquired in the Clinical Proteomics Center, University of Louisville.

The peak detection performance was assessed and compared to the *Cromwell* package [1] that implements Undecimated Discrete Wavelet Transform on the mean spectrum (MUDWT) denoising and peak detection, described in [7].

The performance of both methods was assessed by calculating the area under the Receiver Operating Characteristic Curve (ROC). Because both methods return a list of peak centroids rather than classify each element of the spectrum, we modified counting of true positives, true negatives, false positives, and false negatives to make them suitable for computation of ROC. A detected peak is considered a true positive if the smallest difference between its location and locations of the labeled peaks is smaller that the accuracy, here set to 0.02% of M location (similarily to our mass spectrometer). Thus the number of true positives TP is given by the number of correctly detected peaks. We compute the number of false positives FP as the number of incorrectly detected peaks:

$$FP = R - TP, \tag{18}$$

where R is the number of peaks returned by a peak detection method. On the other hand, the number of false negatives FN is computed as the number of true peaks that were not detected:

$$FN = L - TP, \tag{19}$$

where L is the number of labeled true peaks. We compute the number of true negatives TN as number of all TOF clock ticks in the spectrum that were neither returned as peaks nor labeled:

$$TN = m - R - FN. \tag{20}$$

The experiments were repeated 50 times for several values of number of spectra n and standard deviation of noise σ_f. The resulting areas under ROC curves (AUC) are summarized in Tables I and II. The columns of Table I present number of generated spectra (n), standard deviation of noise (σ_f), mean AUC computed for the CIV method ($P(j|V(j))$, mean AUC computed for the MUDWT algorithm, and statistical significance of the fact that CIV outperforms MUDWT, expressed as the paired t-test P-value. Similarly, the columns of Table II show the mean AUC computed for the IP method (with $r^* = 4$), where probabilities are initialized with CIV (CIV-IP), mean AUC computed for the MUDWT, and significance of the fact that the CIV-IP method outperforms MUDWT. Values of both tables are computed for the same data sets. The performance of CIV and isotopic peaks methods that is at the level of 0.95 significantly better than the AUC of MUDWT is shown in bold in both tables.

As expected, the AUC's presented in Tables I and II increase with n and decrease as σ_f increases. Table I reveals

[1] available at http://bioinformatics.mdanderson.org/software.html

TABLE I

AREAS UNDER ROC CURVE COMPUTED FOR PEAK DETECTION WITH THE CIV AND MUDWT METHODS

n	σ_f	CIV	MUDWT	P-value
10	20	**0.9598**	0.9490	< 0.001
20	20	**0.9637**	0.9557	< 0.001
40	20	0.9655	0.9631	0.107
60	20	**0.9720**	0.9662	0.003
10	30	**0.9591**	0.9399	< 0.001
20	30	**0.9611**	0.9478	< 0.001
60	30	0.9567	0.9600	0.905
10	40	**0.9543**	0.9318	< 0.001
20	40	**0.9530**	0.9410	< 0.001
40	40	0.9523	0.9498	0.136
60	40	0.9418	0.9525	0.999

TABLE II

AREAS UNDER ROC CURVE COMPUTED FOR PEAK DETECTION WITH THE CIV-IP, MUDWT METHODS

n	σ_f	CIV-IP	MUDWT	P-value
10	20	**0.9707**	0.9490	< 0.001
20	20	**0.9714**	0.9557	< 0.001
40	20	**0.9698**	0.9631	< 0.001
60	20	**0.9744**	0.9662	< 0.001
10	30	**0.9650**	0.9399	< 0.001
20	30	**0.9685**	0.9478	< 0.001
60	30	**0.9643**	0.9600	0.030
10	40	**0.9628**	0.9318	< 0.001
20	40	**0.9652**	0.9410	< 0.001
40	40	**0.9658**	0.9498	< 0.001
60	40	**0.9577**	0.9525	0.017

that CIV outperforms MUDWT for small n. Averaging larger number of spectra results in cleaner signal and thus performance of MUDWT is in those cases similar or better than CIV. Note also that P-values in the right-most column increase no only with n, but also with σ_f. Thus CIV is more sensitive to noise than MUDWT.

All P-values in Table II show statistical significance, hence the CIV-IP method outperformed MUDWT in all experiments. The P-vale increases for the largest n and σ, however the difference in performance is still significant.

IV. CONCLUSIONS

A novel approach to detection of low-abundant peaks in high-resolution mass spectra that mimics the techniques utilized by human expert was presented in this paper. The performance of the proposed algorithm was assessed with simulated mass spectra and compared with existing peak detection method. The proposed methodology outperformed the existing algorithm on the simulated set of mass spectra.

The simulated mass spectra did not represent a real mass spectrum as accurately as those in [7]. Our goal was to assess the performance of the presented algorithm specifically for low-abundant peaks. Experiments with more representative simulation and real mass spectra will follow. However, even if the proposed method does not perform as well on more general mass spectra, it could be a good addition to an existing peak detection method that does not detect low-intensity peaks very well. Combination of two or more methods such that each is specialized in detection of certain types of peaks would yield better performance than each of those methods separately.

REFERENCES

[1] P. R. Srinivas, S. Srivastava, S. Hanash, and G. L. Wright Jr. Proteomics in early detection of cancer. *Clinical Chemistry*, 47:1901 – 1911, 2001.

[2] B. L. Adam, Y. Qu, J. Davis, M. Ward, M. Clements, L. Cazares, O. J. Schellhammer, Y. Yasui, Z. Feng, and G. L. Wright Jr. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62:3609 – 3614, 2002.

[3] T. P. Conrads, M. Zhou, E. F. Petricoin III, L. Liotta, and T. D. Veenstra. Cancer diagnosis using proteomic patterns. *Future Drugs*, 4:411 – 420, 2003.

[4] E. P. Diamandis. Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clinical Chemistry*, 8:1272 – 1278, 2003.

[5] J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin. Proteomic applications for the early detection of cancer. *Nature Reviews*, 3:267 – 275, 2003.

[6] Y. Yasui, M. Pepe, M. L. Thompson, B.-L. Adam, G. L. Wright, Jr., Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4:449 – 463, 2003.

[7] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764 – 1775, 2005.

[8] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L. C. Xiao, and K. R. Coombes. A comprehensive approach to the analysis of matrix-assisted laser dessorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3:1667 – 1672, 2003.

[9] M. A. Rogers, P. Clarke, J. Noble, N. P. Munro, A. Paul, P. J. Shelby, and R. E. Banks. Proteomic profiling of urinary proteins in renal cancer by syrface enhanced laser desorption ionization and neural-network analysi: identification of key issues affecting potential clinical utility. *Cancer Research*, 63:6971 – 6983, 2003.

[10] D. Bylund, R. Danielsson, G. Malmquist, and K. E. Markides. Chromatographic aligment by wrapping and dynamic programming as a pre-processing tools for parfac modeling of liquid chromatography mass spectrometry data. *Journal of Chromatography A*, 961:237 – 244, 2002.

[11] J. S. Morris K. A. Baggerly M. C. Hung H. M. Kuerer K. R. Coombes, S. Tsavachidis. 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. Technical report, Department of Biostatistics, University of Texas M. D. Anderson Caner Center, 2004.

[12] G. A. Satten, S. Datta, H. Moura, A. R. Woolfitt, M. da G. Carvalho, G. M. Carlone, B. K. De, A. Pavlopoulos, and J. R. Barr. Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20:3128 – 3136, 2004.

[13] L. A. Higgins. Mass spectral data processing using a novel peak detection algorithm improves results interpretation. In *Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics*, page avaliable on CD, 2004.

[14] M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3(9):1692 – 1698, 2003.

[15] K. R. Coombes, J. M. Koomen, K. A. Baggerly, J. S. Morris, and R. Kobayashi. Understanding the characteristics of mass spectrometry data through the use of simulation. Technical report, Department of Biostatistics, University of Texas M. D. Anderson Caner Center, 2004.