Published in final edited form as:

IEEE Trans Signal Process. 2010 November 1; 58(11): 5883–5894. doi:10.1109/TSP.2010.2065226.

Bayesian Peptide Peak Detection for High Resolution TOF Mass Spectrometry

Jianqiu Zhang,

Department of Electrical and Computer Engineering, University of Texas at San Antonio, TX 78249 USA

Xiaobo Zhou,

Texas Methodist Hospital Research Institute, Houston, TX 77030 USA

Honghui Wang,

Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD 20892 USA

Anthony Suffredini,

Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD 20892 USA

Lin Zhang,

School of Information and Electric Engineering, China University of Mining Technology, XuZhou 221116, China

Yufei Huang, and

Department of Electrical and Computer Engineering, University of Texas at San Antonio, TX 78249 USA, and also with the Greehey Children's Cancer Research Institute, Department of Epidemiology and Biostatistics, University Texas Health Science Center at San Antonio, San Antonio TX 78229 USA

Stephen Wong

Texas Methodist Hospital Research Institute, Houston, TX 77030 USA

Jianqiu Zhang: michelle.zhang@utsa.edu; Xiaobo Zhou: xzhou@tmhs.org

Abstract

In this paper, we address the issue of peptide ion peak detection for high resolution time-of-flight (TOF) mass spectrometry (MS) data. A novel Bayesian peptide ion peak detection method is proposed for TOF data with resolution of 10 000–15 000 full width at half-maximum (FWHM). MS spectra exhibit distinct characteristics at this resolution, which are captured in a novel parametric model. Based on the proposed parametric model, a Bayesian peak detection algorithm based on Markov chain Monte Carlo (MCMC) sampling is developed. The proposed algorithm is tested on both simulated and real datasets. The results show a significant improvement in detection performance over a commonly employed method. The results also agree with expert's visual inspection. Moreover, better detection consistency is achieved across MS datasets from patients with identical pathological condition.

Index Terms

Bayesian methods; Markov chain Monte Carlo; mass spectrometry; peptide peak detection; time-of-flight

I. Introduction

In the last ten years or so, mass spectrometry (MS) has increasingly become the method of choice for analyzing complex protein samples. The key breakthrough that turned mass spectrometry into a major tool for proteomics study was the development matrix-assisted laser desorption ionization (MALDI) for ionizing proteins and protein fragments (peptides). MS measures the mass-to-charge ratio (m/z) of a mixture of ions (or charged peptides) and records the number of ions presenting at each m/z value. It shall be noted that a single peptide species has several different isotopic compositions, and it could register several ion peaks at different m/z values. We define ion-level peptide peak detection as the process of identifying individual ion peaks of a peptide. We define peptide-level peak detection as the process of linking all ion peaks associated with a specific peptide.

The output of an MS instrument is a mass spectrum or chart with a series of peaks, each representing a type of ion whose abundance is represented by the height of the peak. The width of peaks is determined by the resolution of the MS instrument. The resolution of MS instruments is measured by full width at half-maximum (FWHM), which is calculated by dividing the m/z value of a peak by its width at half of its maximum height. In low resolution MS spectra (several thousand FWHM), isotopic peaks from the same peptide are convolved together and ion-level peak detection is not possible. Instead, one can only identify the convoluted ion peaks of a peptide at each charge state. In high resolution MS spectra ($\geq 10\,000$ FWHM), isotopic peaks can be resolved and it is possible to detect single ions of peptides. Ion-level peak detection is suited for the detection of low abundant peptides since they usually only register one ion peak at the most abundance charge state and isotope position while other ion peaks are buried in the background. Peptide-level peak detection will miss these single ions of low abundance peptides. It is pointed out in [1] and [2], that low abundance peptides are often more biologically significant and the percentage of low abundance peptides that can be discovered determines the likelihood of success for biomarker discovery. Thus, ion-level peak detection becomes critical in biomarker discovery studies.

Currently most TOF MS can achieve a high mass resolution (10 000–15 000 FWHM), and this improvement can be attributed to better instrument design [3], which enables ion-level peak detection. In spite of this advancement, peak detection algorithms have not been adequately updated. Most algorithms [4]–[8] are designed for low resolution MS data, which cannot be used for high resolution ion-level peak detection. Algorithm in [9] uses high resolution TOF MS, but it uses a Gabor quadrature filter to extract the envelope signal of the same peptide, which essentially converts high resolution data to low resolution data. Kast *et al.* [10] applies a frequency filter to remove the bell shaped background noise in high resolution data. However, peptide ion peaks share the same frequencies with background noise and frequency filters will remove part of peptide ion peaks. There are essentially no algorithms that are specially designed for ion-level peak detection for mass resolutions over 10 000 FWHM. Currently, the general practice in MS data processing is to smooth MS spectra first and then apply an intensity threshold for peak detection [4], [9], [11], [12]. Among many choices of filters, wavelet based filters are widely adopted [4], [13]. We term this processing method as the wavelet-based method/algorithm. The drawback of this commonly adopted processing method is that: it only uses height information for peak detection. The height of low abundance ion peaks often exhibits high variations among samples that are gathered from patients with the same pathological condition. The inconsistency in peak detection will lead to the failure of biomarker discovery, which aims at finding proteins with unique and consistent behavior in patients with the same pathological condition.

In this paper, we address the issue of ion-level peptide peak detection in high resolution TOF MS data. At the considered mass resolution and when biological samples contain a lot of proteins, MS data exhibit unique characteristics that do not exist in low resolution MS data. In Fig. 1, a segment of TOF data is plotted as an example, where peptide peaks have a resolution of 10000 FWHM. Thus, for m/z values of up to 5000 Dalton (atomic mass unit), ion peaks with approximately one Dalton of difference do not overlap. From Fig. 1, we can see that background peptides and chemical noise form periodic bell shaped curves with a period of approximately 1 Dalton. The formation of the periodic background noise curve can be explained by the research in [14], which surveyed the database of all peptides and found that peptide mass distribution forms Gaussian like curves around cluster centers that can be linearly regressed to peptide nominal weights. Peptide nominal weight equals to the total number of its protons and neutrons. The mass deviation from the mass cluster centers can be attributed to varying compositions of neutrons and protons given identical nominal weights as well as the phenomenon of mass defect. Within the approximately 1 Dalton period, we observed that 1) background peaks are usually wider than peptide peaks, 2) background peaks usually have lower intensity, and 3) background and peptide peaks may center at different m/z values. The goal of this project is to find peptide ion peaks that stand out from the background noise peaks. We propose a Bayesian peak detection algorithm that fully explores the differences between peptide ion and background peaks in peak width, height, and center location. Simulation results show that comparing to the wavelet-based method which only utilizes peak height information for peak detection, the proposed algorithm provides significantly higher sensitivity (or true positive rate) to low abundance peptide peaks given the same specificity.

In the proposed algorithm, peak detection is not performed at the peptide level. The reason is that if a peptide is of low abundance, then only one or two isotopic peaks are distinguishable from the background, peak picking algorithms based on isotope pattern will miss such low abundance peaks. The utilization of isotope patterns generally increases specificity with a cost on sensitivity. Since the proposed algorithm is aimed at improving the sensitivity and the application of isotope pattern contradicts this goal, isotope patterns are not used. Charge state deconvolution is also not performed since for low abundance peptides, only partial isotope pattern exists and it is hard to judge the charge state. Charge state deconvolution and deisotoping can be performed after ion-level detection or even after differential analysis to reduce the risk of lumping weak peaks of different peptides together which will cause serious consequences in feature selection. (Note that for high abundance peptides, these two steps can be performed before differential analysis with little risk of error.)

To perform the mathematically intricate Bayesian detection, an algorithm based on Gibbs sampling is developed. The proposed algorithm estimates the posterior distributions of model parameters given the observed data. Particularly, the algorithm produces the *a posteriori* probability on the existence of a peptide ion peak at a given m/z value, which intrinsically normalizes peptide abundance variation among samples, thus providing better detection consistency among MS data samples obtained from different experimental conditions. Furthermore, the algorithm performs data fusion at the ion level by utilizing the posterior distributions of peaks, which overcomes drawbacks of the averaging plus voting approach. The proposed algorithm provides “soft information” (the probability of existence) for peptide ion peaks. Comparing to hard decisions of either yes or no on the existence of peptide peaks, soft information can be better utilized in subsequent processing steps, since mistakes made in hard decisions could propagate. Also, soft information which integrates various information, is more reliable than peak height for feature selection algorithms.

The rest of the paper is organized as follows. In Section II-A, the proposed data model is described. In Section II-B, the proposed algorithm is discussed in detail. In Section III,

results from using both simulated data and eight sets of real data are provided. Finally, conclusion is drawn in Section IV.

II. Methods

A. Modeling of proTOF Data

A total of eight datasets corresponding to samples from eight people with similar pathological conditions are obtained by using PerkinElmer proTOF 2000 O-TOF MALDI mass spectrometer. Each dataset contains three technical replicates that came from three aliquots of the same sample. Each replicate represents the data spectrum ranging from 708.85 Dalton to 4500 Dalton. Refer again to Fig. 1 for an example of the data spectrum in this range. Signals with distinct spectrum structure are observed to align along the m/z coordinate with each signal occupying approximately 1.0005 Dalton interval referred to as a signal period. The 1.0005 Dalton value is obtained based on our observation which is well corroborated in [14], where it indicates that 1.0005 is the mean ratio between peptide weight and peptide nominal weight. The beginning of the MS spectrum is also determined by observation. When applying the proposed algorithm, users are expected to supply this parameter. Three components make up the observed signal in a signal period. The first component is generated by charged peptides hitting the receiver plate forming a peak that is roughly bell-shaped. The bell-shaped peak curve is formed due to the dispersion of arrival time which can be attributed to the variation of initial velocity [15]. Time-of-flight (TOF) is approximately Gaussian in distribution [16]. Thus, the peaks are shaped like a symmetrical bell curve in the time domain. Although in general the conversion between TOF and m/z values is non-linear [15], within the small span of 1 Dalton, the conversion is almost linear. Thus, in high resolution MALDI instruments, the peaks are still roughly bell shaped after converting TOF to m/z values. The peptide signal at m/z value m , $y_s[m]$, is modeled by a Gaussian-like function as

$$y_s[m] = \beta_s e^{-\rho_s^2(m-\mu_s)^2} \quad (1)$$

where β_s models the height of the signal, the inverse of ρ_s models the spread of the signal, and μ_s models the center of the peak on the m/z axis.

The second component is the background noise that is generated by charged chemical impurities, matrix clusters, and low abundance peptides that are completely buried by other background ions. The background ions form similar curves as those of peptides, but the curves are usually wider due to the complex composition of the background ions with almost identical mass. In contrast, the relative purity of peptide ions forms narrower peaks. Background peaks are treated as “chemical noise”, $y_c[m]$, which can be modeled by a similar exponential function

$$y_c[m] = \beta_c e^{-\rho_c^2(m-\mu_c)^2} + d \quad (2)$$

where β_c , ρ_c , and μ_c have similar meanings as their counterparts in the peptide signal model, and d models a random DC element of the chemical noise. This DC element is added to account for a similar effect of baseline in lower resolution MALDI instruments. Such a phenomenon is attributed to charge accumulation in [17]. It shall be noted that if a MALDI instrument with even higher resolution is employed, one might be able to resolve the component ions in the background peaks [18]. Although the type of instruments we

considered have higher resolutions than most other TOF instruments, they still can not reach the required resolution for the analysis of the “chemical noise” yet. It shall be noted that although the content of the chemical noise may not change from one replicate to another, the intensity does vary randomly from one replicate to another. The third component is high frequency thermal noise, which is ubiquitous. The thermal noise is approximated as white Gaussian noise.

In a signal period that is approximately one Dalton wide, apart from thermal noise, a peptide signal may be embedded in chemical noise. Consequently, the data in a signal period is modeled as

$$y[m] = \lambda y_s[m] + y_c[m] + \varepsilon[m] \quad (3)$$

where $\lambda \in \{0, 1\}$ is an indicator random variable, which is one if the peptide signal exists and zero otherwise, and $\varepsilon[m]$ is modeled as i.i.d. zero mean Gaussian noise with standard deviation σ . Additionally, $\theta = [\lambda, \beta_w, \mu_s, \beta_c, \rho_c, \mu_c, d, \sigma^2]^T$ represents the vector of all unknown parameters to be estimated. Note that λ is of primary interest in peptide peak detection. Note that MALDI instruments with resolution around 10 000 FWHM is enough to resolve most ion peaks within one Dalton at the baseline level, but it is still not high enough to resolve multiple peptide peaks with small mass difference within one Dalton. Therefore, most of the time, only one peptide ion peak stands out from the background within one Dalton. It is possible to extend the model to accommodate multiple ion peaks. However, the computational cost will be very high with little benefit in return since the occurrence of multiple ion peaks in one Dalton is very rare. (This case only happens when multiple peptides with identical nominal mass have enough abundance to allow them to stand out from the background and when they have a mutual mass difference that is at least 3.5 times the full width at half maximum of ion peaks. For example, at 1000 Dalton, with a mass resolution of 10 000, this requires a 0.35 Dalton mass difference.) Multiple peak model is more appropriate for instruments even higher resolution.

As an example, the data spectrum in Fig. 1 is fitted by this model and the fitting result is shown as the solid line. It shows that this model is capable of capturing major signal characteristics such as the center location, the width of background and peptide peaks, and the existence of peptide ion peaks.

Although more complex models can fit the signal better, more parameters or parameters with high correlation of such models will pose greater challenge computationally. For example, a Gamma distribution like curve can also fit the data pretty well. But since parameters used to describe a Gamma curve are strongly correlated, it is hard for the proposed Bayesian peak detection algorithm to converge. Since the goal of the algorithm is not to minimize fitting error but to detect peak reliably while minimizing computational requirements, we select the proposed model.

B. Bayesian Peak Detection

1) Goal of Bayesian Peptide Peak Detection—Peak detection can be performed independently on each signal period of approximately 1 Dalton. The goal of peptide peak detection is to determine if there exists a peptide ion signal in a signal period. Given $\mathbf{y}$, a vector of M data samples in a signal period, the objective of Bayesian detection is to obtain the *a posteriori* probabilities (APPs) $p(\lambda|\mathbf{y})$, i.e., the probability that a peptide signal exists given data. Detection of peptide signals can then be made according to the Maximum *a posteriori* criterion using the APPs [19]. The three characteristics of MS data discussed in

Section I are incorporated into the posterior distribution through the selected model. However, the calculation of the APPs requires high dimensional integration of the joint posterior distribution $p(\theta|\mathbf{y})$ over all parameters except λ . In addition, the marginal posterior distributions of signal model parameters are also desirable to estimate the shape of the peptide signal. Given the highly nonlinear nature of data model (3), none of the desired posterior distributions can be obtained analytically. We therefore resort to a Markov chain Monte Carlo (MCMC) sampling solution.

2) A Markov Chain Monte Carlo (MCMC) Sampling Algorithm for Peak

Detection—The MCMC sampling method [19] is used to generate random samples from the full posterior distributions $p(\theta|\mathbf{y})$. By using random samples, the APPs, $p(\lambda|\mathbf{y})$, and the MAP detection solution can be easily estimated according to the theory of Monte Carlo integration. However, since the number of parameters is large, it is difficult and inefficient to sample all nine parameters at once. To overcome the difficulty, a Gibbs sampling scheme is developed. In Gibbs sampling, instead of drawing samples of nine unknown parameters all together, the popular strategy of divide-and-conquer is employed which only samples one or a subset of parameters at a time. The rest parameters are fixed at the sample values obtained from the previous iteration as if they were “true”. For example, the k th parameter group is sampled from the conditional distribution $p(\theta_k|\theta_{-k}, \mathbf{y})$. This sampling process iterates until the underlying Markov chain converges, and the samples afterwards are shown to be distributed according to the marginal posterior distribution $p(\theta_k|\mathbf{y})$. For basics of Gibbs sampling, please refer to [19]. The algorithm of Gibbs sampling of this work can be summarized as the following.

Algorithm

The Gibbs Sampler for Peak Detection

Iterate the following steps and for the j th iteration

- 1 Draw a sample of σ^2 from $p(\sigma^2|\mathbf{y}, \theta_{-\sigma^2}^{(j-1)})$
 - 2 Draw samples of $\theta_1 = [\beta_s, \beta_c, d]$ from $p(\theta_1^{(j)}|\mathbf{y}, \theta_{-\theta_1})$
 - 3 Draw samples of $\theta_2 = [\rho_c, \mu_c]$ from $p(\theta_2|\mathbf{y}, \theta_{-\theta_2}^{(j)})$
 - 4 Draw samples of $\theta_3 = [\rho_s, \mu_s]$ from $p(\theta_3|\mathbf{y}, \theta_{-\theta_3}^{(j)})$
 - 5 Draw a sample of λ from $p(\lambda|\mathbf{y}, \theta_{-\lambda}^{(j)})$
-

where θ_{-x} represents a subset of θ excluding x and the superscript $\theta^{(r)}$ denotes a sample of θ drawn at iteration r . It has been shown that the samples will converge to the desired full posterior distribution. Convergence assessment will be discussed in detail later. The samples taken after convergence can be used to estimate the target parameters. Particularly, given J samples obtained after convergence, $p(\lambda = 1|\mathbf{y})$ can be approximated according to Monte Carlo approximation as

$$p(\lambda=1|\mathbf{y}) \approx \sum_{j=1}^J \lambda^{(j)} / J \quad (4)$$

which is the frequency of ones in the J samples of λ and peak can be detected based on the MAP criterion if $p(\lambda = 1|\mathbf{y})$ is greater than a threshold. Note the selection of threshold will affect the false alarm rate and the detection rate. We can see that the key for the Gibbs sampling is to derive the expression of the required conditional distributions and then draw samples from these distributions. Due to the nonlinearity of model (3), the computational complexity of the Gibbs sampling algorithm is nontrivial. For details of the derivation steps of the algorithm, see the Appendix.

Convergence and the Number of Samples: The proposed MCMC method will converge after an initial burn in period. To monitor the convergence effectively and automatically, a scheme described in [19, sec. 11.6] is adopted. Specifically, five Markov chains are run in parallel for each of the nine parameters in the model, the between- and within-sequence variances are calculated after every 100 iterations. A score based on the two variances is then evaluated. When the score is smaller than a predefined threshold, the Markov chains are considered to have converged and the iteration process terminates. The first half of the sampled sequence is discarded as the burn-in segment.

3) Bayesian Data Fusion—The goal of data fusion is to integrate data replicates to improve the detection accuracy. From a Bayesian perspective, the aim is to obtain the posterior distribution of parameters given all replicates, which can be expressed as

$$p(\theta|\mathbf{y}_{1:K}) \propto p(\mathbf{y}_K|\theta)p(\theta|\mathbf{y}_{1:K-1}) \quad (5)$$

where $\mathbf{y}_{1:K}$ is the collection of K replicates. Equation (5) suggests a sequential data fusion scheme that integrates a new replicate by recycling the posterior distribution given previously integrated replicates as the prior distribution. Specifically, the proposed Gibbs sampling algorithm is first applied to the first replicate. After this, the sample means and variances of θ_2 and θ_3 are passed on as the parameters of the respective prior distribution when integrating the second replicate. Also, the APP of λ , $p(\lambda|\mathbf{y}_1)$, is used as the prior for processing the second replicate. The height information is not passed due to the fact that concentrations of peptides are different in different replicates. The same is preformed sequentially to integrate the third, fourth, ..., until the K replicate. After all K replicates are processed sequentially, samples obtained from the Gibbs sampling algorithm will follow the distribution $p(\theta|\mathbf{y}_{1:K})$.

The Bayesian method provides a natural framework for integrating the replicates. The targeted posterior distribution $p(\theta|\mathbf{y}_{1:K})$ provides the model parameter distribution based on ALL replicates $\mathbf{y}_{1:K}$. It happens that this distribution can be calculated in a step by step fashion. It shall be pointed out that the order of processing should not affect the result theoretically if the computer is of infinite precision. For example, suppose there are two replicates and one of the replicate spectrums shows a peptide peak clearly and in the other spectrum the peptide peak has low intensity and resembles a peak produced by thermal noise. If the first replicate is processed in the beginning, it will produce a probability measure indicating the existence of a peak (i.e., LLR of λ will be greater than zero). Then this value will be added to the LLR of λ of the second replicate which is ambivalent about the existence of the peak (LLR of λ is around zero). The overall value on the LLR of λ will be positive. If the order is switched, the end result should still be the same.

However, numerically, there could be a difference in processing order. For example, from one replicate, the LLR of λ is very small and is considered to be negative infinity due to limited precision of the computer, then in subsequent steps, it is not possible to change this

“strong belief.” To address such a problem, we process replicates with the smallest over all ion counts first because with a smaller signal-to-thermal-noise ratio, the produced probability measures on the parameters are not numerically extreme. While such a step can not totally wipe out the problem, it reduces the chance of the numerical problem. Also, we stress that the selection of data fusion order only affects rare cases and is an extra cautionary step.

III. Results

A. Tests on Simulated Systems

The proposed algorithm is tested on several sets of simulated MS data. The simulated datasets are generated based on the signal model described in (3). Parameters related to chemical noise in the model are generated randomly from the Gaussian distributions with different mean and variance to account for the fact that β_c, μ_c, ρ_c vary on the whole range of m/z value. Specifically, $\beta_c \sim \mathcal{N}(20, \beta_{cd})$, $\mu_c \sim \mathcal{N}(0.50025, 0.01^2)$, and $\rho_c \sim \mathcal{N}(5.5, 0.01^2)$. For peptide signal, it is decided that $\mu_s \sim \mathcal{N}(0.50025, 0.05^2)$. Other parameters change significantly at different m/z values, and therefore different combinations of $\beta_{cd}, \rho_s, \beta_s, \sigma^2$ are selected to test the performance of the algorithm under different combinations of these parameters. The typical set of values are $\beta_{cd} = 2$, $\rho_s = 20$, and $\sigma^2 = 2.2$. Also, $\beta_s = 0$ or ($\beta_s = 5$; when $\beta_s = 0$, there is no peptide ion signal and when $\beta_s = 5$, a weak peptide signal is generated. Higher values of β_s are also tested and peptide peaks are almost always detected. So we do not include those simulations here.

First, the fidelity of the proposed algorithm in estimating the APPs for weak peaks is investigated. Since the APPs of λ can be considered as the combined information of peak height, peak width, and peak location, they can be used as features for biomarker discovery to improve the performance. Thus, it is important to estimate the APPs accurately. In Fig. 2, the histogram of the APPs $p(\lambda = 1|\mathbf{y})$ is plotted for $\beta_s = 0$ (no peptide signal) and $\beta_s = 5$ (weak peptide signal). The histogram is constructed based on 800 independent trials. It can be observed that when weak peaks exist, the histogram of $p(\lambda = 1|\mathbf{y})$ pronounces at 1 and when there is no signal, the distribution centers around 0.15. It is noted that the APPs of the case of $\beta_s = 0$ are more spread out and do not concentrate on value zero since it is generally difficult to pronounce non-existence than existence. Thermal noise can produce very weak and narrow peaks that behave like peptide signals. The result reflects this and the APP distribution is more dispersed when there are no peptide signal. If the variance of the thermal noise reduces, the APP distribution will be narrower.

This demonstrates the algorithm’s ability in providing faithful estimates of the APPs even for weak peaks.

Next, the efficacy of data fusion is examined. To this end, three replicates are produced to mimic the real datasets. The ROC curves of the detection results for using one, two, and three replicates are plotted in Fig. 3. The corresponding false positive rate (FPR) and true positive rate (TPR) of each point on a curve is obtained by setting a decision threshold on $p(\lambda|\mathbf{y})$ in 800 independent trials. It shows that the Bayesian data fusion scheme behaves as expected and the integration of three replicates greatly improves the detection performance. We also alter the order of data fusion and we have not observed any noticeable changes in the resulted ROC.

To demonstrate the effectiveness of the propose Bayesian algorithm, we compare the proposed algorithm with a commonly used MALDI processing method. To perform data fusion, we select the method of averaging because with an intra-experiment CV of less than 10% among the dataset, the intensities are similar and averaging is the best practice

according to [11]. Averaging produces a spectrum that has much smaller thermal noise variance. There exists a lot of options for subsequent processing steps: parametric model based methods such as match filtering [20], or on generalized inversion [21]. However these methods can not be applied since they do not model the chemical noise. In [11], it is pointed out that the easiest way to find peak is to smooth the raw spectrum and take those local maxima which exceed a threshold [4], [9], [12]. Thus, after averaging, we perform smoothing to get rid of high frequency thermal noise and then we apply different intensity thresholds to detect peptide peaks. There are many methods for high frequency noise removal, and a popular method is using wavelet denoising [4], [13]. Here we refer to this method as the wavelet based method. Note that the method presented in [10] does not perform smoothing after subtracting frequency components of the chemical noise. Instead, the peaks buried in noise are magnified and then a threshold is applied for detecting peaks. Since such a method is not adopted widely, it is not included in our comparison.

The selection of threshold affects the detection performance. If a threshold is set low, then TPR increases at the expense of increased FPR. When the threshold is set high, TPR decreases and so does FPR. To implement the wavelet smoothing, the Matlab de-noising function “wden” is used. Threshold selection rule is set to be universal and soft thresholding is used. Rescaling is based on a single estimation of level noise based on first level coefficients. Wavelet decomposition is performed at level 2, and “sym8” wavelet is used. The wavelet-based method is tuned before it is applied. The detailed explanation of the wavelet algorithm implemented in Matlab can be found at: <http://matlab.izmiran.ru/help/toolbox/wavelet/wden.html>. Finally various thresholds are applied for detection.

The ROC curves of the proposed algorithm are compared with those of the wavelet based algorithm for four different cases. In case one, all parameters are set to their typical values except that $\beta_{cd} = 1.5$, which is set to elucidate the effect of reduced chemical noise height variation on the algorithms. In case two, parameters are set to their typical values. In case three, while all other parameters are set to their typical values, we set $\rho_s = 30$ instead of $\rho_s = 20$, which reduces the spread of peptide signals. In case four, noise variance σ^2 is reduced to 1.8 and we set $\beta_{cd} = 1.5$ to examine the effect of the reduced additive noise variance on the tested algorithms. The results are summarized in Fig. 4.

From Fig. 4, it is apparent that our proposed algorithm performs better than the wavelet based algorithm in all four cases. Comparing case 1 and 2, we can see that there is a noticeable difference between the ROC curves in the wavelet based method, while the difference between the ROC curves of the proposed algorithm is very small. This means that the variation in chemical noise height almost does not affect the performance of the proposed algorithm since the algorithm does not rely on the height of signal for peak detection solely. Comparing case 2 and case 3, we can see that as peptide signal becomes narrower, the performance of both algorithms degrades. In such cases, it is easier to confuse the narrow peptide signal with high frequency thermal noise. The wavelet algorithm deteriorates more than the proposed algorithm. By comparing case 1 and case 4, we can see that by reducing the variance of additive noise, the proposed algorithm improves quite significantly while the wavelet algorithm does not change since additive noise is mostly removed after wavelet de-noising. Overall, the proposed algorithm is much more robust and outperforms the wavelet based algorithm on simulated datasets.

B. Test on Real Data

The proposed algorithm is applied to 8 sets of real pTOF MS data as described in Section II-A. We performed visual inspection of a segment (861–864 Dalton) in Fig. 1 according to the expert prior knowledge and annotated the peak identities in each Dalton of the segment.

Due to noise, the signal in 863–864 cannot be confidently identified; the peak might be an isotope of a peptide at m/z 861. In Fig. 5, we showed the detection results in terms of the posterior $p(\lambda = 1|Y)$. The results agree very well with visual inspection. More significantly, the posterior probabilities also captures the uncertainties in the expert's inspection. In the first two Daltons, the algorithm detects the existence of peaks almost surly. Yet, the algorithm detected a peak in 863–864 with $p(\lambda = 1|Y) = 0.88$; which means that there is clearly uncertainty in deciding a peak but the algorithm considers the signal to be more likely a peak while the visual inspection also inclined to declare a peak. In the last Dalton, the algorithm agrees with the expert opinion by declaring chemical noise with the probability of 0.94. Posterior probabilities reflect expert knowledge very well.

Next, we compare the proposed with the wavelet based algorithm. The wavelet based algorithm as described in Section III-A is applied. The noise characteristics in the real data is very similar to our simulated data and identical parameter settings for the “wden” function in Matlab are used as that in Section III-A. Threshold selection rule is set to be universal and soft thresholding is used. Rescaling is based on a single estimation of level noise based on first level coefficients. Wavelet decomposition is performed at level 2, and “sym8” wavelet is used. The wavelet-based method is tuned before it is applied. We study the residue signal after applying the “wden” smoothing function. It is assumed that the residual signal represents thermal noise and should have no correlation with the smoothed signal. We evaluate different combinations of parameter settings provided by “wden” in Matlab and the select the parameter setting that yields the lowest normalized correlation between the residual and the smoothed signal. Fig. 6 demonstrates that the smoothing is effective in removing all high frequency variations while ion-level peptide signals are preserved.

When applying our proposed algorithm, the detection threshold is set as 0.8, i.e., when $p(\lambda = 1|Y) > 0.8$, we consider that the peak exists. A small segment (952–968 Dalton) is first selected and visual inspection is performed to determine if signal peaks are present. Therefore, detection call is based on human expertise and experience, which is considered fairly reliable for this small segment. In Figs. 7 and 8, three replicates of data as well as the fitted result by the proposed algorithm are plotted together. For the wavelet algorithm, the fitted curves are plotted by the solid lines and the identified peaks are indicated by stems marked by “o”. For the proposed algorithm, since it can estimate the complete peptide signal shape, the envelopes of the detected peptide signal are marked by *. In Fig. 7, we compare the result of the wavelet based algorithm with a high threshold (34) to the proposed method. It is shown that peaks at 956–957, 967 Dalton are missed by the threshold while the proposed method does not. In Fig. 8, we show the result when the threshold is lowered to 30. At this level, the peaks at 956 and 967 Dalton missed by the higher threshold are admitted. But the lower threshold admits a new noise peak at 968 Dalton and the peptide signal at 957 Dalton is still missed. This result demonstrates that the proposed Bayesian algorithm is much more versatile in detecting signals with different intensity and different noise variances.

Next, we compare the reproducibility of peak detection by the proposed algorithm and the wavelet based algorithm. The number of peaks found per spectrum is not an adequate measurement on the quality of peak finding algorithms [4]. It is important to detect peaks consistently. In addition, a good peak detection algorithm should also be able to reproduce its findings obtained under the same conditions. In [4], the reproducibility of peak detection of different algorithms is tested by comparing the number of detected peaks that are present in all replicates of identical aliquotes. In our experiment, we examined the consistency of the algorithm in detecting a peptide peak across eight datasets. The rationale of this experiment is that since eight datasets are obtained from patients with similar pathological conditions, there must exists a set of common peptides which are indicative of their healthy/diseased

state. Note that this assumption justifies the usage of Mass Spectrometry as a disease diagnostic and for biomarker discovery tool. We argue that the ability of consistently detecting the presence/absence of peaks among a group of datasets should be the ultimate test on the performance of a peak detection algorithm, because more consistent peaks one can find, the higher the probability of discovering biomarkers. It shall be pointed out that with limited resolution, even if a peptide peak is found consistently at a m/z value within a dataset, it does not warrant that the peak is registered by the same peptide, and further investigation by a biologist is needed. The algorithm merely provides a list of candidate peptide peaks for further investigation.

To conduct a fair comparison between two different algorithms, we require a fixed total number of the detected peaks for each algorithm, since an algorithm can always generate a meaningless “consistency” result when either the false positive rate or the false negative rate is allowed high. In this experiment, the segment from 1008.85 to 1308.85 Dalton is chosen for demonstration. The reason for choosing only a segment is because the threshold for the wavelet-based algorithm needs to be adapted to local chemical noise levels. Running on a smaller segment generally benefits the wavelet algorithm. For each dataset, the detection threshold of the Bayesian algorithm is set to 0.8, i.e., if $p(\lambda|y_{1:3}) > 0.8$ for a particular signal period, then a peptide peak is detected. For a given signal period l , we represent the number of datasets that are determined to contain a peptide peak as $C_l \in [0,8]$. $C_l = 0$ indicates that no peptide signal exists in any datasets in the signal period l . If $C_l = 8$, the peptide peak exists in all eight datasets for the given signal period l . We calculate C_l s for the 300 signal periods from 1008.85 to 1308.85. In Fig. 9, we plot the histogram of C_l for the proposed algorithm. As a comparison, we also test the wavelet algorithm and plot the corresponding histogram of C_l . Note that the height threshold is selected such that the averaged number of detected peptide peaks per dataset is approximately the same as that of the proposed algorithm at $p(\lambda|y_{1:3}) > 0.8$. If an algorithm has better consistency, the higher frequency should be expected for $C_8 = 8$ and $C_l = 0$. From Fig. 9, it is obvious that the proposed algorithm returns more consistent results, i.e., the proposed algorithm identifies almost twice as many signal periods with $C_p = 8$ and $C_p = 0$ than the wavelet algorithm. This shows the great improvement of the proposed algorithm in terms of consistency than the wavelet algorithm. We also simulated the case when we set $p(\lambda|y_{1:3}) > 0.6$ as the threshold, the proposed algorithm still has much better consistency.

In a follow-up work [22], we performed feature selection and classification based on the probabilistic (soft) information obtained through the algorithm we proposed here and compare it to the wavelet based algorithm on a total of 16 datasets, six of which are from diseased patients and the rest from a control group. Using the Bayesian soft information, a 100% sensitivity and a 91% specificity are achieved. Using similar feature selection and classification algorithms based on the peak intensity (obtained after the wavelet based algorithm), a sensitivity of 66% and a specificity of 83% are achieved. This shows the effect of the proposed peak detection algorithm in down stream processing steps.

IV. Discussion and Conclusion

We present a parametric model for high resolution TOF MS data. Unlike existing peak detection approaches, where only peak height information is considered, the proposed model characterizes multiple features of peptide signal and chemical noise, thus lending information to better identify low peptide ion peaks. To fully take advantage the model, a Bayesian MCMC sampling algorithm is developed, which incorporates the prior knowledge of peptide signals and chemical noise, and produces probabilistic information about the existence of peptide signals. The proposed algorithm is based on a rigorous statistical framework and it integrates several replicates better than ad hoc “data fusion” methods such

as voting. The proposed algorithm also combines thermal noise removal, chemical noise removal and peptide peak detection in one processing step whereas typical MS processing algorithms separate them in three steps, which is suboptimal.

The algorithm has been validated on simulated datasets, where the comparison with commonly used methods shows much improved detection performance. The proposed Bayesian algorithm is also applied to eight sets of real pTOF data. The results show that our algorithm provides significantly greater consistency. Consistency in peak detection has important implications on classification and feature selection, and is crucial for reliable biomarker discovery.

Peptide ion peak detection is just the first yet important step towards MS data-oriented biomarker discovery and disease diagnosis. In a follow up work to this project, we demonstrated the effectiveness of the proposed algorithm in reducing classification error.

In this work, we did not utilize generalized isotope patterns for peak detection. However, in other processing steps, such pattern information can be utilized. For example, in the feature selection step, it is possible to utilize isotope patterns to build reliable classifier. For example, if one isotope peak seems to express differentially among healthy and diseased patient but other isotope peaks that belongs to the same peptide do not, then one can exclude the peak from the list of possible features for classification.

The discussion of charge state estimation is also left out in this paper. The result of this peak detection algorithm can be feed into feature selection algorithms for differential analysis or peptide/protein identification algorithm in which the issue of estimating charge state can be addressed.

Since the mass drift considered in this paper is very small, we do not address the issue of m/z alignment. If the mass drift is large, then conventional m/z alignment methods should be applied before peak detection.

The proposed algorithm is computationally complex. However, it has a structure that is compatible with parallel processing. Thus, the complexity can be managed. Given the good performance of the proposed Bayesian algorithm in peptide peak detection, extending the proposed method to other MS technologies including LC/MS will also be promising.

Acknowledgments

The authors would like to thank Dr. L. Sapp in PerkinElmer, Boston, MA, for acquiring all pTOF MS data.

The work of J. Zhang is funded by a San Antonio Life Sciences Institute-Research Enhancement grant and supported by a SALS research enhancement grant and an award G12RR013646 from the National Center For Research Resources. The work of X. Zhou is funded in part by The Methodist Hospital Scholarship Award. The work of X. Zhou and S. Wong are also funded in part by NIH Grants R01LM08696, R01LM009161, and R01AG028928. The work of Y. Huang is supported by NSF Grant CCF-0546345. The work of L. Zhang is supported by the Chinese Scholarship Council.

Biographies

Jianqiu (Michelle) Zhang received the Ph.D. degree in electrical engineering from the State University of New York at Stony Brook in 2002.

From 2002 to 2007, she worked as an Assistant Professor in the Department of Electrical and Computer Engineering, University of New Hampshire. Since 2007, she has been with the Department of Electrical and Computer Engineering at the University of Texas at San

Antonio (UTSA), where she is currently an Assistant Professor. She has been a Visiting Professor at the Greehey Children's Cancer Institute and the Department of Epidemiology and Biostatistics at the University of Texas Health Science Center at San Antonio. Her current research interest is in the field of proteomic mass spectrometry, Bayesian statistical signal processing methods, bioinformatics, biomarker discovery, and classifications. She has been working on problems of LC/MS peak detection, quantification, FTMS signal correction, and MALDI imaging for prostate and kidney cancer research. Her current research is funded by a San Antonio Life Sciences Institute Research Enhancement grant.

Dr. Zhang was a recipient of the Best Paper Award of the *IEEE Signal Processing Magazine*.

Xiaobo Zhou received the B.S. degree from Lanzhou University, Lanzhou, China, in 1988 and the M.S. and the Ph.D. degrees from Peking University, Beijing, China, in 1995 and 1998, respectively, all in mathematics.

From 1988 to 1992, he was a Lecturer at the Training Center in the 18th Building Company, Chongqing, China. From 1992 to 1998, he was a Research Assistant and Teaching Assistant with the Department of Mathematics at Peking University, Beijing, China. From 1998 to 1999, he was a Post doctoral Fellow in the Department of Automation at Tsinghua University, Beijing, China. From 1999 to 2000, he was a Senior Technical Manager with the 3G Wireless Communication Department at Huawei Technologies Company, Limited, Beijing, China. From February 2000 to December 2000, he was a Postdoctoral Fellow in the Department of Computer Science at University of Missouri-Columbia. From 2001 to 2003, he was a Postdoctoral Fellow in the Department of Electrical Engineering at Texas A&M University, College Station. From 2003 to 2007, he was first a Research Fellow and then a faculty member with the Harvard Center for Neurodegeneration and Repair in Harvard Medical School and Radiology Department in Brigham & Women's Hospital. Since June 2007, he has been an Associate Professor of The Methodist Hospital Research Institute affiliated with Cornell University, Ithaca, NY. His current research interests include image

and signal processing for medical imaging analysis, cellular imaging analysis, neuroinformatics, bioinformatics for genomics and proteomics, In-silicon cancer stem cell microenvironment modeling, and drug response modeling for integrated multiscale systems biology research. He is funded by a number of NIH grants. He made original contributions to the cellular image analysis, genomics and phosphoproteomics analysis.

Honghui Wang received the B.S. degree from Wuhan University, China, in 1982, the M.S. degree from the Institute of Chemistry, Academia Sinica, Beijing, China, in 1986, and the Ph.D. degree in chemistry from University of Louisville, Louisville, KY, in 1992.

He is currently working as a Staff Scientist in the Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD. His research interests include clinical proteomics, assay and method development, and biomarker discovery and validation.

Anthony Sutfredini received the B.A. degree from Boston University in 1973 and the M.D. degree from the University of Connecticut School of Medicine in 1979.

Following a residency in internal medicine at the Medical College of Virginia, he completed fellowships in critical care medicine at the University Health Center of Pittsburgh and later in the Critical Care Medicine Department, Clinical Center, National Institutes of Health,

Bethesda, MD. He has been a Senior Investigator in the Critical Care Medicine Department, NIH, since 1989. His research interests include regulatory mechanisms of acute inflammation in pneumonia, sepsis, and septic shock.

Lin Zhang received the B.Sc. degree and the Ph.D. degree in communication and information systems from the China University of Mining and Technology, Xuzhou, Jiangsu, in 2002 and 2007, respectively.

She is currently a Lecturer at the China University of Mining and Technology. Her research interests includes classification and feature selection.

Yufei Huang received the Ph.D. degree in electrical engineering from the State University of New York at Stony Brook in 2001.

From 2001 to 2002, he worked as a Postdoctoral Researcher in the Department of Electrical and Computer Engineering, the State University of New York at Stony Brook. Since 2002, he has been with the Department of Electrical and Computer Engineering at the University of Texas at San Antonio (UTSA), where he is currently an Associate Professor. He has been a Visiting Professor at the Center of Bioinformatics, Harvard Center for Neurodegeneration & Repair. He is currently also an Adjunct Professor of the Greehey Children's Cancer Institute and Department of Epidemiology and Biostatistics at the University of Texas Health Science Center at San Antonio. His expertise is in the area of genomics signal processing, statistical modeling, and Bayesian methods. He currently focuses on high throughput biological data integration, gene networks discovery, micro-RNA target identification, and LC-MS data analysis.

Dr. Huang was a recipient of the National Science Foundation (NSF) Early CAREER Award in 2005, the Best Paper Award of 2006 Artificial Neural Networks in Engineering Conference, and the 2007 Best Paper Award of the *IEEE Signal Processing Magazine*. His research has been supported by the NSF, the National Institute of Health, and the Air Force Office of Scientific Research. He has been an organizer of several workshops and special sessions on genomic signal processing. He is an Associate Editor of the *IEEE Transactions on Signal Processing*, the *EURASIP Journal on Bioinformatics and Computational Biology*, and the *International Journal Machine Learning and Cybernetics*.

Stephen Wong received the B.E.E.E. (Hons.) degree from the University of Western Australia, Perth, in 1983 and the M.Sc. and Ph.D. degrees in computer science from Lehigh University, Bethlehem, PA, in 1989 and 1991, respectively. He received his executive education from the MIT Sloan School of Management, the Stanford University School of Business, and Columbia University School of Business.

He was the Director of the Center for Bioinformatics, Harvard Center of Neurodegeneration and Repair (HCNR), and Director of the Functional and Molecular Imaging Center, and an Associate Professor of Radiology, Harvard Medical School and Brigham & Women's Hospital. He is currently the Vice-Chair of Radiology, Director of Bioinformatics Core, The Methodist Hospital Research Institute and Cornell University, Ithaca, NY. His research theme has been focused on the application of advanced technology to pragmatic biomedical problems and is based on the belief that problems of importance involve the interplay between theory and application. He is a hybrid scientist. He has published over 200 peer-reviewed papers and holds six patents in biomedical informatics. He also served on NIH and NSF scientific review panels. He has broad research and development experience worldwide for two decades with HP, Bell Labs, ICOT-Japanese 5th Generation Computer Systems project, Philips Electronics, Charles Schwab, and UCSF. His earlier research involves the pioneering work in optical time domain reflectometer and optical networks, think jet (first inkjet) automation, 1 MB DRAM, and VLSI factory automation before moving into the fields of bioinformatics and medical imaging. He was a key member of the UCSF PACS effort, founded the product development departments of Philips Medical Systems, and directed the Web trading development and re-architecturing of Schwab.com, one of the largest secured eCommerce sites.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003; 422(6928):198–207. [PubMed: 12634793]
2. Veenstra T, Conrads T, Hood B, Avellino A, Ellenbogen R, Morrison R. Biomarkers: Mining the biofluid proteome. *Mol Cellular Proteomics*. 2005; 4(4):409–418.
3. PerkinElmer. prOTOF 2000 MALDIO-TOF MALDI orthogonal time of flight mass spectrometer. PerkinElmer Life and Analytical Sciences Product Brochure [Online]. Available: http://las.perkinelmer.com/Content/relatedmaterials/brochures/bro_highcontentproteomics.pdf
4. Coombes K, Tsavachidis S, Morris J, Baggerly K, Hung M, Kuerer H. 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(16):4107–4117. [PubMed: 16254928]
5. Du P, Kibbe W, Lin S. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*. 2006; 22(17):2059. [PubMed: 16820428]
6. Morris J, Coombes K, Koomen J, Baggerly K, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*. 2005; 21(9):1764–1775. [PubMed: 15673564]
7. Yasui Y, McLerran D, Adam B, Winget M, Thornquist M, Feng Z. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *J Biomed Biotechnol*. 2003; 4(2003):242–248. [PubMed: 14615632]
8. Wang Y, Zhou X, Wang H, Li K, Yao L, Wong S. Reversible jump MCMC approach for peak identification for stroke SELDI mass spectrometry using mixture model. *Bioinformatics*. 2008; 24(13):i407. [PubMed: 18586741]
9. Yu W, Wu B, Lin N, Stone K, Williams K, Zhao H. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Comput Biol Chem*. 2006; 30(1):27–38. [PubMed: 16298163]
10. Kast J, Gentzel M, Wilm M, Richardson K. Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *J Amer Soc Mass Spectrom*. 2003; 14(7):766–776. [PubMed: 12837599]
11. Hilario M, Kalousis A, Pellegrini C, Muller M. Processing and classification of protein mass spectra. *Mass Spectrom Rev*. 2006; 25(3):409–49. [PubMed: 16463283]
12. Yasui Y, Pepe M, Thompson M, Adam B, Wright G Jr, Qu Y, Potter J, Winget M, Thornquist M, Feng Z. A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostat*. 2003; 4(3):449.
13. Shao X, Leung A, Chau F. Wavelet: A new trend in chemistry. *Acc Chem Res*. 2003; 36(4):276–283. [PubMed: 12693925]
14. Wolski W, Farrow M, Emde A, Lehrach H, Lalowski M, Reinert K. Analytical model of peptide mass cluster centres with applications. *Proteome Sci*. 2006; 4(1):18. [PubMed: 16995952]
15. Coombes K, Koomen J, Baggerly K, Morris J, Kobayashi R. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Inform*. 2005; 1(1):41–52.
16. Dass C. Principles and practice of biological mass spectrometry. *Chem BioChem*. 2002; 3:1155–1160.

17. Malyarenko D, Cooke W, Adam B, Malik G, Chen H, Tracy E, Trosset M, Sasinowski M, Semmes O, Manos D. 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. *Clinical Chem.* Nov.2004 51:65–74. [PubMed: 15550476]
18. Krutchinsky A, Chait B. On the nature of the chemical noise in MALDI mass spectra. *J Amer Soc Mass Spectrom.* 2002; 13(2):129–134. [PubMed: 11838016]
19. Gelman, A. *Bayesian Data Analysis*. Boca Raton, FL: CRC Press; 2004.
20. Palmblad M, Buijs J, Håkansson P. Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. *J Amer Soc Mass Spectrom.* 2001; 12(11):1153–1162. [PubMed: 11720389]
21. Mohammad-Djafari A, Giovannelli J, Demoment G, Idier J. Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems. *Int J Mass Spectrom.* 2002; 215(1–3):175–193.
22. Zhang, L.; Zhang, J.; Zhou, X.; Wong, H.; Huang, Y.; Liu, H.; Wong, S. Feature selection and classification of prOTOF data based on soft information. *Proc. Int. Conf. Machine Learning Cybernetics*; Shanghai, China. Jul. 12–15, 2008; p. 4018-4023.

Appendix

Here, we explain the detailed derivations in each Gibbs sampling step. For simplicity, we drop the iteration superscript (j) in the following discussion.

Sample σ^2 From $p(\sigma^2|\mathbf{y}, \theta_{-\sigma^2})$

In order to obtain the conditional posterior on σ^2 , we first derive the likelihood function

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{m=1}^M (2\pi\sigma^2)^{-\frac{M}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{y}[m] - \lambda \mathbf{y}_s[m] - \mathbf{y}_c[m]\|^2} \\ &\propto (\sigma^2)^{-M/2} \exp(-1/(2\sigma^2) \|\mathbf{y} - \lambda \mathbf{y}_s - \mathbf{y}_c\|^2) \end{aligned} \quad (6)$$

where M is the total number of observations within a given signal period, $y_s[j]$ and $y_c[j]$ are estimated peptide and chemical noise signals that can be obtained by plugging in estimated parameters to (1) and (2). $\mathbf{y}_s = [y_s[1], \dots, y_s[M]]$, and $\mathbf{y}_c = [y_c[1], \dots, y_c[M]]$ are vector representations of the estimated peptide and chemical noise signals within one signal period. Equation (6) is derived based on the fact that $p(\mathbf{y}|\theta)$ is a multivariate Gaussian distribution with mean $\lambda \mathbf{y}_s + \mathbf{y}_c$, and the additive noise is zero mean Gaussian with variance σ^2 .

Then the conditional posterior distribution of σ^2 can be evaluated as

$$\begin{aligned} &p(\sigma^2|\mathbf{y}, \theta_{-\sigma^2}) \\ &\propto p(\mathbf{y}|\theta)p(\sigma^2) \\ &\propto (\sigma^2)^{-M/2} \exp(-1/(2\sigma^2) \|\mathbf{y} - \lambda \mathbf{y}_s - \mathbf{y}_c\|^2) (\sigma^2)^{-1} \\ &\propto IG(M/2, \|\mathbf{y} - \lambda \mathbf{y}_s - \mathbf{y}_c\|^2/2) \end{aligned} \quad (7)$$

where in (7), we selected the inverse gamma density function as the prior distribution, $p(\sigma^2) \propto (\sigma^2)^{-1}$. We realize that the overall conditional posterior distribution is an Inverse Gamma density function. In this case, $p(\sigma^2|\mathbf{y}, \theta_{-\sigma^2})$ has an analytical form and samples can be taken directly from it.

Sample θ_1 From $p(\theta_1|\mathbf{y}, \theta_{-\theta_1})$

The three parameters in $\theta_1 = [\beta_s, \beta_c, d]^\top$ are special in the sense that given other parameters, $\theta_{-\theta_1}$, they become linear with respect to model (3). However, the conditional distribution takes different forms for different λ .

The conditional posterior density can be expanded as

$$p(\theta_1|\mathbf{y}, \theta_{-\theta_1}) \propto p(\mathbf{y}|\theta)p(\theta_1|\theta_{-\theta_1}) \quad (8)$$

$$\propto p(\mathbf{y}|\theta)p(\beta_c, d|\theta_{-\theta_1})p(\beta_s|\theta_{-\theta_1}) \quad (9)$$

$$\propto p(\mathbf{y}|\theta)p(\beta_c, d)p(\beta_s|\lambda) \quad (10)$$

where (9) follows since the prior knowledge on chemical noise height β_c and DC level d is independent of the prior knowledge on the height of the peptide signal β_s , (10) follows since β_c, d are independent from the rest of model parameters, and the prior distribution on β_s depends on the value of the indicator random variable.

When $\lambda = 1$, i.e., when the sample obtained at previous iteration indicates that there exists a peptide signal, we can write the observations $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{H}\theta_1 + \boldsymbol{\varepsilon} \quad (11)$$

where $\mathbf{H} = [\mathbf{h}_s, \mathbf{h}_c, \mathbf{1}_{M \times 1}]$ is a $M \times 3$ matrix obtained by concatenating three column vectors $\mathbf{h}_s$ and $\mathbf{h}_c$, and $\mathbf{1}_{M \times 1}$ together. Here, $\mathbf{h}_s = e^{-(\rho_s)^2(m-\mu_s)^2}$, $\mathbf{h}_c = e^{-(\rho_c)^2(m-\mu_c)^2}$, and $\mathbf{1}_{M \times 1}$ is a vector in which every element equals to one. With such a formulation, it can be shown that, provided with Gaussian prior distributions $p(\theta_1|\lambda = 1) = p(\beta_c, d) * p(\beta_s|\lambda = 1)$, the conditional posterior density of θ_1 is also Gaussian, whose the mean and the covariance matrix are the Bayesian estimates of these parameters. The mean can be shown to be

$$\mu_{\theta_1|\mathbf{y}, \theta_{-\theta_1}} = \mu_{\theta_1} + \sum_{\theta_1} \mathbf{H}^\top (\mathbf{H} \sum_{\theta_1} \mathbf{H}^\top + \sum_{\boldsymbol{\varepsilon}})^{-1} (\mathbf{y} - \mathbf{H}\mu_{\theta_1}) \quad (12)$$

and the covariance matrix becomes

$$\sum_{\theta_1|\mathbf{y}, \theta_{-\theta_1}} = \sum_{\theta_1} - \sum_{\theta_1} \mathbf{H}^\top (\mathbf{H} \sum_{\theta_1} \mathbf{H}^\top + \sigma^2 \mathbf{I})^{-1} \mathbf{H} \sum_{\theta_1}. \quad (13)$$

where $\sum_{\theta_1}$ and μ_{θ_1} are the mean and diagonal covariance matrix of the prior Gaussian distribution $p(\theta_1|\lambda = 1)$ and $\mathbf{I}$ stands for an identity matrix.

In our simulations, we selected prior distribution parameters as $\mu_{\theta_1} = [\text{Max}(\mathbf{Y}) - 18 - 5, 18, 5]$ where $\text{Max}(\mathbf{Y})$ stands for the maximum value of the observation vector during the

considered signal period, and $\Sigma_{\theta_1} = \text{diag}\{25, 16, 16/9\}$, where $\text{diag}\{\cdot\}$ stands for the diagonalization of the vector. These values are obtained by visually inspecting the data. More elaborate methods for selecting these priors are possible which may impact the efficiency of sampling. However, in our experiment, the impact is not significant if we change the parameters for the prior distributions slightly. These parameters will need to be adjusted for different datasets. It shall be noted that the conditional posterior distribution is mainly determined by the observations but not these prior distributions given that the number of observations is large enough. In our simulations, for a given datasets to be processed, we use priors estimated from other datasets obtained in the same experiment. In this way we avoid using the same information twice when calculating the posterior probabilities.

When $\lambda = 0$, i.e., the previous sampling step indicates that no peptide signal exists, then the observation $\mathbf{y}$ carry no information about β_s . Therefore, the conditional posterior distribution is now expressed as

$$\begin{aligned} p(\theta_1|\mathbf{y}, \theta_{-\theta_1}, \lambda=0) &= p(\beta_c, d|\mathbf{y}, \theta_{-\theta_1}, \lambda=0)p(\beta_s|\lambda=0) \\ &\propto p(\mathbf{y}|\theta)p(\beta_c, d)p(\beta_s|\lambda=0). \end{aligned} \quad (14)$$

where $p(\beta_s|\lambda=0)$ is the prior of β_s given $\lambda=0$. Note that if the same conjugate Gaussian priors as those in the case of $\lambda=1$ are chosen again for β_c and d , $p(\beta_c, d|\mathbf{y}, \theta_{-\theta_1})$ is still the Gaussian distribution, whose mean and covariance matrix are defined in (12) and (13). From this Gaussian distribution, samples of β_c and d can be easily taken. Meanwhile, the prior distribution $p(\beta_s|\lambda=0)$ seems to have no physical meaning and can be set up arbitrarily. However, according to [19], the selection of this prior distribution affects the biasness given to the case $\lambda=0$. In our algorithm, we set $p(\rho_s|\lambda=0)$ as a uniform distribution between 0 and the maximum element of $\mathbf{y}$.

Sample θ_2 From $p(\theta_2|\mathbf{y}, \theta_{-\theta_2})$

It can be shown that the conditional posterior distribution $p(\theta_2|\mathbf{y}, \theta_{-\theta_2})$ is known up to a normalizing constant as

$$\begin{aligned} p(\theta_2|\mathbf{y}, \theta_{-\theta_2}) &\propto (\sigma^2)^{-M/2} e^{-\frac{1}{2\sigma^2} \sum_{m=1}^M \left(y^c[j] - \beta_c e^{\rho_c(m[j] - \mu_c)^2} \right)^2} p(\theta_2) \\ &= g(\theta_2) \end{aligned} \quad (15)$$

where $\mathbf{y}^c = \mathbf{y} - \lambda\mathbf{y}_s - d\mathbf{I}_{M \times 1}$ is the chemical noise without the DC element, $p(\theta_2)$ is the prior distribution of θ_2 . The prior distribution is set to be a two dimensional uniform distribution with lower limits set to be zero as these parameters can only be positive. The upper limit of ρ_c is set to be 7 since it is observed that the spread of 99% of the area under chemical noise peak is approximately $6/(\sqrt{2}\rho_c)$, which is never smaller than 0.6 Dalton. The upper limit of the center of the chemical noise is set to be 1 above the starting m/z value of the signal period since the resolution is greater than one Dalton and the center of the peak can never deviate from the starting m/z by one Dalton in the m/z range considered in this paper (≤ 4500 Da). Direct Sampling from (15) is not possible due to the unknown constant. Alternatively, we employ a single Metropolis-Hastings sampling step. To this end, we first obtain a candidate sample θ_2^* from a proposal density $q(\theta_2)$, then θ_2^* is accepted as a sample from $p(\theta_2|\mathbf{y}, \theta_{-\theta_2})$ with probability

$$r = \min \left\{ \frac{g(\theta_2^*) q(\theta_2^r)}{g(\theta_2^r) q(\theta_2^*)}, 1 \right\}. \quad (16)$$

The proposal density is set to be a two dimensional Gaussian distribution, whose mean is equal to sample θ_2^{j-1} of the last Gibbs sampling iteration, and whose covariance matrix is the sampled covariance of Σ_{θ_2} obtained from the previous 1000 Gibbs sampling steps and updated every 1000 steps.

Sample $\theta_3 = [\rho_s, \mu_s]^T$ From $p(\theta_3|\mathbf{y}, \theta_{-\theta_3})$

The next group of parameters to be sampled are the mean and the spread of the peptide signal peak. Just like θ_1 , the conditional distribution is different conditioning on different values of λ . Given $\lambda = 1$, i.e., when the previous sampling step indicates that there exists a peptide signal, $p(\theta_3|\mathbf{y}, \theta_{-\theta_3})$ can be only known up to a normalizing constant

$$\begin{aligned} & p(\theta_3|\mathbf{y}, \theta_{-\theta_3}) \\ & \propto (\sigma^2)^{-M/2} e^{-\frac{1}{2\sigma^2} \sum_{j=1}^M \left(\mathbf{y}^s[j] - \lambda \beta_s e^{\rho_s(m[j] - \mu_s)^2} \right)^2} p(\theta_3) \\ & = g(\theta_3) \end{aligned} \quad (17)$$

where $\mathbf{y}^s = \mathbf{y} - \mathbf{y}^c - d\mathbf{I}_{M \times 1}$ is the estimated peptide signal peak given $\theta_{-\theta_3}$, and $p(\theta_3)$ is the prior distribution on θ_3 . $p(\theta_3)$ is also set to be a two dimensional uniform distribution over the space for $0 < \mu_s < 1$ and $4 < \rho < 45$. The limits of ρ_s are chosen based on the fact that the spread, or approximately $6/(\sqrt{2}\rho_s)$, is greater than 0.1 Dalton and smaller than 1 Dalton. Again, θ_3 cannot be sampled directly and a single Metropolis-Hastings sampling step is employed. The acceptance probability is defined similarly as that in (16) but with θ_3 replacing θ_2 . The proposal density $q(\theta_3)$ is also chosen to be a two dimensional Gaussian distribution, whose mean equals to θ_3^{j-1} from the last Gibbs sampling iteration, and the covariance matrix is the sample covariance during previous 1000 Gibbs sampling steps, which is updated every 1000 steps.

When $\lambda = 0$, i.e., the previous sampling step indicates that no peptide signal exists for the given signal period, $p(\theta_3|\mathbf{y}, \theta_{-\theta_3}) = p(\theta_3)$ and samples of θ_3 are taken from the prior distribution as described above.

Sample λ From $p(\lambda|\mathbf{y}, \theta_\lambda)$

Note that $p(\lambda|\mathbf{y}, \theta_\lambda)$ is a Bernoulli distribution, which can be sampled directly. It can be shown that

$$\begin{aligned} & p(\lambda|\mathbf{y}, \theta_{-\lambda}) \propto p(\mathbf{y}|\theta)p(\theta|\lambda)p(\lambda) \\ & \propto (\sigma^2)^{-M/2} e^{-1/(2\sigma^2) \|\mathbf{y} - \lambda \mathbf{y}_s - \mathbf{y}^c\|^2} p(\rho_s, \theta_3|\lambda)p(\lambda) \end{aligned} \quad (18)$$

where $p(\theta|\lambda)$ is proportional to $p(\rho_s|\lambda)$ since all the parameters except ρ_s in $\theta_{-\lambda}$ are independent of λ *a priori*. From (18), the log-likelihood ratio (LLR) of λ can be calculated as

$$LLR_\lambda = -\frac{1}{2\sigma^2}(|\mathbf{y} - \mathbf{y}_s - \mathbf{y}_c|^2 - |\mathbf{y} - \mathbf{y}_c|^2) + \ln \frac{p(\rho_s|\lambda=1)}{p(\rho_s|\lambda=0)} + \ln \frac{p(\lambda=1)}{p(\lambda=0)}. \quad (19)$$

and then $p(\lambda|\beta_s, \rho_s, \mu_s, \beta_c, \rho_c, \mu_c, d, \sigma, \mathbf{y})$ can be derived as $p(\lambda = 1|\mathbf{y}, \boldsymbol{\theta}_{-\lambda}) = 1/(1 + e^{-LLR_\lambda})$ and $p(\lambda = 0|\mathbf{y}, \boldsymbol{\theta}_{-\lambda}) = 1 - p(\lambda = 1|\mathbf{y}, \boldsymbol{\theta}_{-\lambda})$. Note that a uniform prior is assumed for $p(\lambda)$ and the last term in (19) is zero when processing the first replicate. When processing subsequent replicates, $p(\lambda)$ is set as the $p(\lambda|\mathbf{Y}_{1:k-1})$, where $\mathbf{Y}_{1:k-1}$ stands for observations of the signal from previously processed replicates. The prior distribution on parameters related to the peptide signal ρ_s differ given different values of λ , and we use the same prior distributions as described following (12), (13), and (14).

Fig. 1. Plot of a segment of prOTOF MALDI data. The annotation is according to expert knowledge.

Fig. 2. Histogram of $p(\lambda = 1|y)$ obtained from 800 independent trials.

Fig. 3.
The ROC curves for using different number of data replicates.

Fig. 4. The ROC comparisons of the proposed algorithm with the wavelet base algorithm under different simulation settings.

Fig. 5.

The posterior probabilities of the indicator variable $p(\lambda = 1|Y)$ for each 1.005 Dalton for the m/z range of Fig. 1. The curves of the replicates are also shown. The posterior probabilities agree very well with the expert.

Fig. 6.
Wavelet smoothing of the averaged spectrum.

Fig. 7. The comparison of the detection and curve fitting results by the proposed and wavelet based algorithms. The detection threshold for the wavelet based algorithm is set at 34.

Fig. 8.

The comparison of the detection and curve fitting results by the proposed and wavelet based algorithms. The detection threshold for the wavelet based algorithm is set at 30.

Fig. 9.

Comparisons of different simulations to test the robustness of the algorithm. M/z value = [1008.85, 1308.85] Bayesian threshold = 0.8 Wavelet threshold = 28.12.