BIOINFORMATICS

Vol. 24 ISMB 2008, pages i407–i413
doi:10.1093/bioinformatics/btn143

Reversible jump MCMC approach for peak identification for stroke SELDI mass spectrometry using mixture model

Yuan Wang[1,2], Xiaobo Zhou[1,*], Honghui Wang[3], King Li[1], Lixiu Yao[2] and Stephen T.C. Wong[1]

[1]Center for Biotechnology and Informatics (CBI), The Methodist Hospital Research Institute, and Department of Radiology, The Methodist Hospital, Weill Cornell Medical College, Houston, TX 77030, USA, [2]School of Electronics, Information and Electrical Engineering, Shanghai Jiao Tong University, China and [3]Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

Mass spectrometry (MS) has shown great potential in detecting disease-related biomarkers for early diagnosis of stroke. To discover potential biomarkers from large volume of noisy MS data, peak detection must be performed first. This article proposes a novel automatic peak detection method for the stroke MS data. In this method, a mixture model is proposed to model the spectrum. Bayesian approach is used to estimate parameters of the mixture model, and Markov chain Monte Carlo method is employed to perform Bayesian inference. By introducing a reversible jump method, we can automatically estimate the number of peaks in the model. Instead of separating peak detection into substeps, the proposed peak detection method can do baseline correction, denoising and peak identification simultaneously. Therefore, it minimizes the risk of introducing irrecoverable bias and errors from each substep. In addition, this peak detection method does not require a manually selected denoising threshold. Experimental results on both simulated dataset and stroke MS dataset show that the proposed peak detection method not only has the ability to detect small signal-to-noise ratio peaks, but also greatly reduces false detection rate while maintaining the same sensitivity.

Contact: XZhou@tmhs.org

1 INTRODUCTION

Stroke is a type of cardiovascular disease which occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot (ischemia stroke) or bursts (hemorrhagic stroke). It is the third leading cause of death in the United States. About 700 000 Americans each year suffer a new or recurrent stroke. Numerous studies have shown that early detection of stroke increases treatment options and improves survival rates. Novel biomarkers of stroke can be discovered by comparing the differences in protein expression profiles between serum or tissue extract samples from stroke patients and normal individuals. Mass spectrometry (MS) is increasingly used to detect disease-related biomarkers from human plasma or serum for early diagnosis, prognosis and monitoring of disease progression or response to treatment (Issaq *et al.*, 2002; Vorderwulbecke, *et al.*, 2005). In surface-enhanced laser desorption and ionization (SELDI; Malyarenko *et al.*, 2005), a biological sample is bound to a pre-coated surface on ProteinChip. The coating enables the ProteinChip surface to bind a particular class of proteins based on their chemical properties. Some proteins in the sample bind to the surface, while others are removed by washing. After that, the samples are analyzed by laser desorption/ionization time-of-flight MS to generate mass/charge profiles of the applied sample.

MS data require complex pre-processing techniques before subsequent statistical data mining analyses can be carried out (Dijkstra *et al.*, 2007). To extract features from the MS data, the first step is peak detection. Peak detection is always performed by a series of substeps such as baseline removal, denoising and peak identification (Baggerly *et al.*, 2004; Fung and Enderwick, 2002; Yasui *et al.*, 2003). Successful baseline removal is very challenging in MS data pre-processing. Once the baseline removal step is performed, it is impossible to recover from the false baseline correction result in the subsequent processing (Baggerly *et al.*, 2004). Similarly, denoising step also leads to a one-way result in the whole preprocessing process. Once a peak is misjudged as noise and removed by denoising algorithm, it will never be found by the subsequent peak identification. Therefore, a good MS data preprocessing method should be able to conduct baseline removal, denoising and peak detection in the same time to avoid generating errors and bias from each individual substep.

Another critical step in peak detection is to successfully remove noise from spectrum. Normally, denoising is performed based on some denoising threshold (Coombes *et al.*, 2005b; Tan *et al.*, 2006). If the threshold is too high, some small peaks are erased along together with noise. And if the threshold is too low, noise may be still left in the denoised spectrum which is detected as false positive peaks in peak detection results. Unfortunately, there is no effective method to automatically choose a proper threshold for different spectra currently. If the denoising threshold is inappropriate, final peak detection result may either have a high false positive rate or a low sensitivity. In order to avoid such a problem, a non-parameter denoising method should be developed in pre-processing MS data.

There are several existing methods for peak detection. Some of them use deterministic method (Fung and Enderwick, 2002; Yasui *et al.*, 2003) to locate peak region and identify peak from background noise. Some use wavelet-based transform such as using discrete wavelet transform to denoising the spectrum (Coombes *et al.*, 2005b; Morris *et al.*, 2005; Randolph and Yasui, 2006) or continuous wavelet-based pattern matching to detect peaks (Du *et al.*, 2006). Also there are statistical and model-based methods (Dijkstra *et al.*, 2006; Wang *et al.*, 2006; Noy and Fasulo 2007). However, most of these peak detection algorithms identify

[*]To whom correspondence should be addressed.

peaks based only on the peak amplitude, ignoring the additional information such as shape of the peaks. Additionally, all of these algorithms perform baseline removal, denoising and peak detection at different stages, which greatly increases the risk of introducing errors from each individual stage. And also some of these algorithms have the problem of having a large amount of false positives in their peak detection results because the manually assigned denoising parameter or signal-to-noise ratio (SNR) threshold is hard to optimize for different spectra.

Here we introduce a novel mixture model for stroke MS spectrum data. Based on this model, we designed a non-parameter Bayesian approach to estimate parameters in the model. We believe that the peak detection result is inherently associated with different parameters in this model (peak signal parameters, baseline signal parameters and noise parameters). These parameters are estimated automatically from a reversible jump Markov chain Monte Carlo (MCMC) algorithm simultaneously. Therefore, our peak detection method not only eliminates the risk of introducing errors and bias from baseline correction and noise smoothing but also be able to perform non-parametric denoising to the raw spectrum data. Experimental results show that our method can clearly identify small intensity peaks from relatively large variance of noise while maintaining a low false detection rate (FDR).

2 METHODS

2.1 Mixture model for SELDI spectrum

A SELDI MS spectrum data contains a number of recorded mass to charge ratio (*m/z*) values and observed corresponding intensities. There are three parts of information in a normalized spectrum: peak information generated by biological sample, baseline information from matrix background noise (detected matrix molecules and fragments, dark current and detected air molecules) and random white noise from the MS instrument system. We model the whole MS spectrum as:

$$y_i = \sum_{k=1}^{K} f_k(x_i) + g(x_i) + n_i \quad i = 1,2,\ldots,N, \tag{1}$$

where x_i is the i-th *m/z* in the spectrum and y_i is its corresponding output intensity. N is the length of spectrum with a total number of K peaks. $f(x_i)$ and $g(x_i)$ represent the peak information and baseline information of the spectrum, respectively. $n_i \sim N(0,\sigma^2)$ is the Gaussian random noise with zero-mean and standard deviation σ. $f_k(x_i)$ is the k-th peak signal (Fig. 1). Here, we use radial basis functions to model $f(x_i)$ and use a polynomial function for $g(x_i)$:

$$f_k(x_i) \triangleq A_k e^{-\rho_k (x_i-\mu_k)^2} \quad i = 1,2,\ldots,N; \quad k = 1,2,\ldots,K, \tag{2}$$

$$g(x_i) \triangleq \sum_{p=0}^{P} B_p x_i^p \quad i = 1,2,\ldots,N, \tag{3}$$

where μ_k denotes the location of the k-th peak, A_k denotes its amplitude and ρ_k determines shape of the peak. B_p is the coefficients of polynomial function $g(x_i)$. Integrating Equation (2) and (3) into Equation (1), the whole SELDI MS spectrum can be interpreted as:

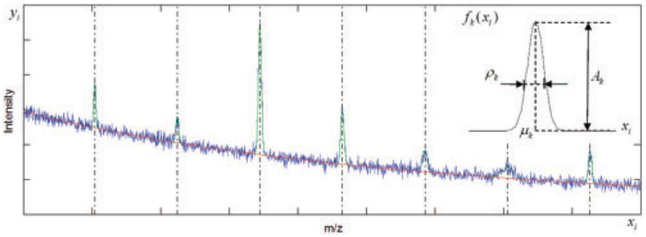

Fig. 1. Red curve indicates baseline of the spectrum $g(x_i)$, green lines indicate peak functions $f_k(x_i)$, $k = 1,2,\ldots,K$. Each peak is determined by its location μ_k, amplitude A_k and shape ρ_k.

$$y_i = \sum_{k=1}^{K} A_k e^{-\rho_k (x_i-\mu_k)^2} + \sum_{p=0}^{P} B_p x_i^p + n_i \quad i = 1,2,\ldots,N. \tag{4}$$

Given $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, the peak detection task of SELDI MS spectrum is to detect the number of peaks K and all the peak locations μ_k in the spectrum. Denote $\mathbf{y} \triangleq [\,y_1 \; y_2 \; \ldots \; y_N\,]^T$, $\mathbf{n} \triangleq [\,n_1 \; n_2 \; \ldots \; n_N\,]^T$, then Equation (4) can be rewritten as:

$$\mathbf{y} = \mathbf{D}(\boldsymbol{\mu}_{1:K}, \boldsymbol{\rho}_{1:K}, \mathbf{x}_{1:N})\boldsymbol{\alpha}_{1:K+2} + \mathbf{n}, \tag{5}$$

where $\boldsymbol{\alpha} = [\,B_0 \; B_1 \; \ldots \; B_P \quad A_1 \; \ldots \; A_K\,]^T$ and

$$\mathbf{D} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^P & f(x_1,\mu_1,\rho_1) & \cdots & f(x_1,\mu_K,\rho_K) \\ 1 & x_2 & \cdots & x_2^P & f(x_2,\mu_1,\rho_1) & \cdots & f(x_2,\mu_K,\rho_K) \\ \vdots & \vdots & & & \vdots & & \\ 1 & x_N & \cdots & x_N^P & f(x_N,\mu_1,\rho_1) & \cdots & f(x_N,\mu_K,\rho_K) \end{bmatrix}.$$

Based on the SELDI MS spectrum data $\{(x_i,y_i)\}_{i=1}^N$, we estimate number of peaks K and the corresponding parameters $\boldsymbol{\alpha}_{1:K+2}$, $\boldsymbol{\mu}_{1:K} \triangleq \{\mu_1,\mu_2,\ldots,\mu_K\}$, $\rho_{1:K} \triangleq \{\rho_1,\rho_2,\ldots,\rho_K\}$ and σ^2.

2.2 Bayesian approach and reversible jump MCMC

We assume that the number K and the parameters $\boldsymbol{\theta}_K \triangleq \{\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\rho},\sigma^2\}$ are unknown. Given a set of observations $\boldsymbol{O} \triangleq \{x_1,\ldots,x_N,y_1,\ldots,y_N\}$, our goal is to estimate K and $\boldsymbol{\theta}_K$. Accurately estimating $\boldsymbol{\rho}_{1:K}$ is not an easy job. So we first assume $\boldsymbol{\rho}_{1:K}$ as a constant and $\boldsymbol{\theta}_K$ becomes $\boldsymbol{\theta}'_K = \{\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2\}$. Following Andrieu *et al.* (2001), Bayesian inference is used to estimate the unknown parameters K and $\boldsymbol{\theta}'_K$. Hyper-parameter $\Lambda, \delta^2 \in \mathbb{R}^+$ are introduced and presumed to be independent of each other. We assume σ^2 have a prior distribution of $1/\sigma^2$. δ^2 follows inverse-gamma distribution, i.e. $\delta^2 \sim \mathcal{IG}(2,10)$, and Λ follows gamma distribution, i.e. $\Lambda \sim \mathcal{G}a(1/2,0)$. According to Bayes theorem, the joint posterior distribution can be formalized as:

$$p(K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2|\mathbf{x},\mathbf{y}) \propto p(\mathbf{y}|K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2,\mathbf{x})p(K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2), \tag{6}$$

where, $p(\mathbf{y}|K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2,\mathbf{x})$ is the likelihood and $p(K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2)$ is the prior distribution. The likelihood for mixture model (4) is:

$$p(\mathbf{y}|K,\boldsymbol{\theta}'_K,\mathbf{x}) = (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\cdot}\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{D}\boldsymbol{\cdot}\boldsymbol{\alpha})\right). \tag{7}$$

The prior distribution $p(K,\alpha,\mu,\sigma^2)$ is given by:

$$\begin{aligned} p(K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2) &= p(\boldsymbol{\alpha}|K,\boldsymbol{\mu},\sigma^2)p(\boldsymbol{\mu}|K,\sigma^2)p(K|\sigma^2)p(\sigma^2) \\ &= p(\boldsymbol{\alpha}|K,\sigma^2)p(\boldsymbol{\mu}|K)p(\sigma^2). \end{aligned} \tag{8}$$

Given likelihood and prior distribution, the joint posterior distribution (6) can be obtained as the following expression:

$$\begin{aligned} & p\left(k,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2,\Lambda,\delta^2|\mathbf{x},\mathbf{y}\right) \\ & \propto \left[(2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{D}\boldsymbol{\cdot}\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{D}\boldsymbol{\cdot}\boldsymbol{\alpha})\right)\right] \\ & \times \left|2\pi\sigma^2\Sigma\right|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\alpha}'\Sigma^{-1}\boldsymbol{\alpha}\right) \\ & \times \left(-\frac{1}{\sigma^2}\right)\left[\frac{\mathrm{II}_{\Omega}(J,\boldsymbol{\mu})}{\zeta^J}\right]\left[\frac{\Lambda^k/k!}{\Sigma_{k=0}^{k_{\max}}\Lambda^k/k!}\right], \end{aligned} \tag{9}$$

where $\Sigma^{-1}=\delta^{-2}\mathbf{D}'\mathbf{D}$ and $\mathrm{II}_{\Omega}=(k,\boldsymbol{\mu})$ is the indicator function of the set Ω [1 if $(k,\boldsymbol{\mu})\in\Omega$, 0 otherwise]. One might select the model order K by $\arg\max p(K|x,y)$ with $K\in\{0,1,\ldots,K_{\max}\}$ and also can perform parameter estimation by computing the conditional expectation $E(\boldsymbol{\theta}_K|\mathbf{x},\mathbf{y})$ based on (10) shown subsequently. However, it is difficult to obtain these quantities analytically, as it involves integrals of high-dimension of non-linear functions. Therefore, the reversible jump MCMC (rjMCMC) method was proposed to perform necessary Bayesian computation. The principle of MCMC is to draw random samples from an ergodic Markov chain $(K^{(i)},\boldsymbol{\theta}'^{(i)}_k)_{i\in\mathbb{N}}$ whose equilibrium distribution is the target posterior distribution. The initial value of K is $K_{\max}$. The Markov chain generates $T\gg 1$ sampling points, asymptotically convergent to the posterior distribution. We discard the points resulted from the initial steps, which is so-called in-birth period, and keep the last P steps for the computation. Here we set T = 2000 and P = 1000. Given a test sample $\mathbf{x}_{N+1}$, $\mathbf{y}_{N+1}$ can be then evaluated by:

$$\begin{aligned}\hat{\mathbf{y}}_{N+1}&=\hat{E}\left(\mathbf{y}_{N+1}|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{N+1},\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N\right)\\&=\frac{1}{P}\sum_{i=1}^{P}\mathbf{D}\left(\boldsymbol{\mu}^{(i)},\mathbf{x}_{N+1}\right)E\left(\boldsymbol{\alpha}|K^{(i)},\boldsymbol{\theta}^{(i)},\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N,\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N\right).\end{aligned}\tag{10}$$

The reversible jump MCMC sampler is able to sample directly from the joint distribution and jump between subspaces of different dimensions. A general state–space Metropolis–Hasting (MH) algorithm is performed, in which candidates are proposed according to a set of proposal distributions. The candidates are randomly accepted according to an acceptance ratio that ensures reversibility and thus invariance of the Markov chain with respect to the posterior distribution.

A big advantage of reversible jump MCMC algorithm is that it can change the dimension of search space by performing randomly birth/death/split/merge moves (Green, 1995). In our case, the number of mixtures (corresponding to the number of peaks) is unknown. By using rjMCMC sampler, it can be automatically determined by maximizing $p(K|\mathbf{x},\mathbf{y})$. The convergence of this rjMCMC approach has been proved in Andrieu *et al*. (2001). Although this method is computationally intensive, it provides a robust estimation of unknown true peak signal from noised data. Therefore, it could be able to detect peak accurately from MS data even with strong noise.

2.3 Incorporating estimation of ρ

Note that in the previous rjMCMC algorithm by Andrieu *et al*. (2001), the parameter ρ is treated as constant. However, it could cause a lot of false positives and false negatives in the peak detection result if we set the width of peak as a constant. In this study, instead of using a general ρ to describe shape character for all peaks, we estimate ρ_K for each individual peak. Once K and $\mu_{1:K}$is estimated, the number of peaks and their locations in the spectrum are known. Given K, $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, σ^2, $\mathbf{x}$, $\mathbf{y}$, the conditional posterior density function $g(\rho_j)$ is given by:

$$\begin{aligned}&g(\rho_j)\triangleq p(\rho_j|K,\boldsymbol{\alpha},\boldsymbol{\mu},\sigma^2,\mathbf{x},\mathbf{y})\\&\propto(\sigma^2)^{-h_j}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=\mu_j-h_j}^{\mu_j+h_j}\left(y_i-\sum_{j=1}^{K}A_je^{-\rho_j(x_i-\mu_j)^2}-\sum_{p=0}^{P}B_px_i^p\right)\right)p(\rho_j).\end{aligned}\tag{11}$$

The prior distribution of ρ_j is set to be a uniform distribution between $[\rho_{\min}\ \rho_{\max}]$. h_j is set to be a very small distance, so the peak with location μ_j is the dominate contributor to the signal observed in $[\mu_j-h_j,\mu_j+h_j]$. In the t-th rjMCMC iteration, after obtaining μ_j, a new sample ρ_j^* from a proposal density function $q(\rho_j)$ is proposed, then the sample's probability with respect to the proposal density and the targeting density $g(\rho_j)$ are calculated. The proposal density function $q(\rho_j)$ is given similarly as ρ_j's prior distribution. The probability of accepting the sample is calculated by:

$$r=\frac{g\left(\rho_j^*\right)q\left(\rho_j^{t-1}\right)}{g\left(\rho_j^{t-1}\right)q(\rho_j^*)}.\tag{12}$$

After the number K and other parameters $(\boldsymbol{\mu}_{1:K},\boldsymbol{\rho}_{1:K},\boldsymbol{\alpha}_{1:K+2},\sigma^2)$ are obtained, we can get all peak information (peak numbers, peak locations and peak shapes) in the spectrum. Actually, real peak shape can be of different varieties. Using only one fixed model [Equation (2)] is insufficient to represent all kinds of possible peak signals. However, a combination of mixtures could represent all the different kinds of peaks. Dijkstra *et al*. (2006) claimed that all these small mixtures indicated real peaks. In this study, we still use traditional peak definition by searching the local maximum on estimated peak signal $\sum_{k=1}^{K}A_ke^{-\rho_k(x_i-\mu_k)^2}$. And peak intensity is defined as the distance between peak signal intensity and baseline information intensity at the local maximum.

2.4 Initialization

In processing the stroke SELDI MS data, we first use some regular denoising techniques (for example, the wavelet transform-based denoising) to remove noise from the spectra and get roughly peak information (approximate peak numbers and approximate peak width) in the spectra. Then based on the estimation, we set peak shape sampling region $\rho_{\min}$[1] to 3000 and $\rho_{\max}$ to 20 000. Detail sample region h_j is set to 1% of corresponding *m*/*z* of μ_j, $j=1,2,\ldots,K$ (Normally, peak width is small in low *m*/*z* region and large in high *m*/*z* region). $K_{\max}$ is set to twice the approximate peak numbers from previous estimation. $\mu_1,\mu_2,\ldots,\mu_{K_{\max}}$ are initialized from a uniform distribution between $[\min(x_{1:N})\ \ \max(x_{1:N})]$.

2.5 Implementation

Due to computation limitation, it is very slow to estimate all the peak information from a whole spectrum by proposed algorithm. One possible solution is to first focus on a small segment of the spectrum. In fact, our experiment result (Section 3.2) shows that noise is not evenly distributed within the whole spectrum. Noise variance tends to be high (a large σ) in low *m*/*z* region and low (a small σ) in high *m*/*z* region. Instead of performing denoising on the whole spectrum, region-based denoising may smooth the data more precisely. The whole algorithm of detecting peaks from raw SELDI MS data based on the mixture model of Equation (5) can be described in the following steps:

(1) Divide the whole spectrum into segments.[2]

(2) Initialization: i = 0; set $K^{(0)}$ as $K_{\max}$; sample $(\boldsymbol{\mu}^{(0)}_{1:K},\boldsymbol{\rho}^{(0)}_{1:K})$ according to their prior distributions.

(3) *while* $i\leq T$[3]
- Random a decision and make move (birth/death/merge/split/update/random walk) based on the decision, i.e. perform the reversible jump process. $K^{(i)},\boldsymbol{\mu}^{(i)}_{1:K}$ are updated from the move according to Equation (9).
- For each $\mu_j^{(i)}$, from its neighborhood $[\mu_j^{(i)}-h_j^{(i)},\ \mu_j^{(i)}+h_j^{(i)}]$, and update $\rho_j^{(i)}$ using Equations (11) and (12).

[1] In our program, we first convert *m*/*z* of the process spectrum to [0 1], so the random sampler can work at a proper data range. ρ is related to this transformation.

[2] The memory requirement of the rjMCMC algorithm is $K\times N^2$. K is the number of Gaussian mixtures (peaks) and N is the length of segment.

[3] T is the maximum length of Marko Chain.

```
    • Update $\boldsymbol{\alpha}^{(i)}_{1:K+2}, \sigma^{2(i)}$ according to Equation (9).
    • $i \leftarrow i+1$.
  end
```

3 RESULTS

We first use simulated spectra generated from MS spectrum model in Equation (4) to test the performance of the proposed peak detection method in Section 2. We also use simulated spectra generated from Coombes *et al*. (2005a) to compare the performance between the wavelet-based method and our proposed method. Then we apply our peak detection method on stroke SELDI MS data.

3.1 For simulation

In order to test the performance of our peak detection method, we first use the simulated spectra. The spectra were generated according to the SELDI data model described in Equation (4) and we adjusted the parameters properly to make the simulated data have the same characteristics as stoke SELDI MS data received from the experiment. Stroke data has several characteristics: peaks in the low *m*/*z* region tended to have a small peak width which looked sharp and peaks in the high *m*/*z* region tended to have a large peak width which looked broad; sometimes peaks were too close and together they combined to a wider peak; noise intensity was decreased while *m*/*z* was increasing and peak intensity generally followed a decreasing trend from low *m*/*z* to high *m*/*z* (Vestal and Juhasz, 1998). Based on the characteristics above, we generated two datasets of simulated spectrum. Each dataset had a total number of 100 simulated spectra. The real peak number was 35 in dataset A and was 50 in simulated dataset B. Peak location μ was randomly generated from 2000 to 12 000 Daltons (Da), with exponentially decreased probability of appearance. Peak area parameter ρ was restricted to [2000 20 000] in dataset A and [5000 30 000] in dataset B, both with an exponentially decreased tendency. Zero mean Gaussian noise was generated with a variance from 0.4 to 0.08 in dataset A and from 0.6 to 0.1 in dataset B. Peak intensity was randomly generated from 0.5 to 30, with 80% probability in [0.5 5] and 20% in [5 30].

Figure 2 is one of the spectra from simulated dataset A. It is obvious that the wavelet-based peak detection method has many false detection results with confusing large noise into peaks, and it misses some small peaks like the one with 3913, 3962 and 7661 Da. In contrast, our peak detection method successfully detects all correct peaks with only one false result. And peak intensity is more precise as compared to the wavelet-based method. Detail of peak detection result from our proposed method and undecimated discrete wavelet transform (UDWT) method are shown in Table 1.

Simulated dataset A has fewer peaks than dataset B, so the peak overlapping chance is relatively higher in the second dataset. Also, noise variance is higher in the second dataset, so it makes peaks in simulated dataset B more difficult to be detected. From Table 1, we can see that both peak detection methods have a good sensitivity in the simulated dataset A. The performance of wavelet-based peak detection method highly relies on the cutoff wavelet coefficient threshold, and it is very hard to balance between the sensitivity and FDR. Often, in order to be more sensitive, wavelet-based method has to suffer a large number of false positives in their detection result. However, our proposed method is able to achieve a good sensitivity while keeping the false detection at a very low rate. In dataset B, due to presence of strong noise, our peak detection method still maintains a good sensitivity and a relatively low FDR. Yet the performance of wavelet-based method drops greatly with lower sensitivity and significantly increasing of FDR. Although we set the SNR much larger, FDR of the result is still beyond acceptable limits.

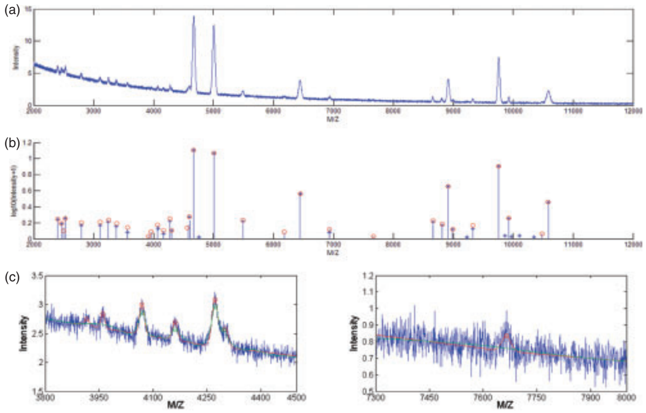

Fig. 2. (**a**) Raw spectrum. (**b**) Peak detection result by wavelet-based method (blue star) and our proposed method (red circle). Real peak is shown in blue spike. (**c**) Close look about the denoised spectrum by wavelet-based (green) and our proposed method (red), with circle point out detected peaks.

Table 1. Result on simulated data

	SNR	Sensitivity	FDR
Dataset A			
Proposed method	0.3	0.909	0.047
Proposed method	0.6	0.894	0.028
UDWT (low threshold)	1.5	0.868	0.262
UDWT (low threshold)	3	0.805	0.145
UDWT (high threshold)	1.5	0.822	0.129
UDWT (high threshold)	3	0.782	0.077
Dataset B			
Proposed method	0.5	0.881	0.079
Proposed method	1	0.839	0.064
UDWT (low threshold)	4	0.731	0.542
UDWT (low threshold)	10	0.693	0.375
UDWT (high threshold)	4	0.689	0.154
UDWT (high threshold)	10	0.659	0.120

Low threshold in UDWT is set to 4 and high threshold is set to 10.

We also test our proposed method on the simulated spectra from Coombes *et al*. (2005a). Figure 3 is a close look of one simulated spectrum with *m*/*z* from 8000 to 12 000 Da. Figure 3a and b are denoising results from the wavelet-based method with low wavelet coefficient thresholds and high wavelet coefficient thresholds, respectively. Clearly, a small threshold allows wavelet denoising to preserve more details from the raw spectrum, but it keeps many low frequency noises in the denoised result, which will turn into false-positive peaks in detection. Using a high threshold in wavelet denoising suppresses low frequency noise but it also suppresses peaks with small intensity that leads to a low sensitivity. Figure 3c is the denoising result from our proposed method. While maintaining the ability to detect small peaks, our algorithm successfully removes all kinds of noise. It can also be noticed that

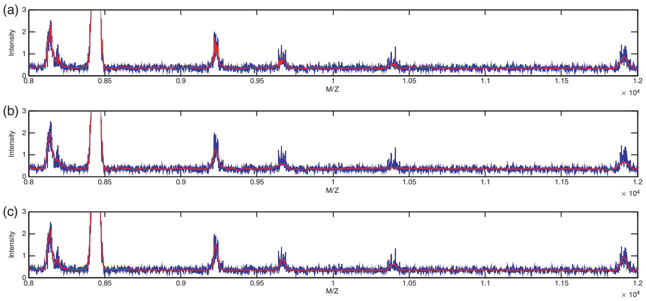

Fig. 3. (**a**) Result by the wavelet-based peak detection method with low threshold. (**b**) Result by the wavelet-based peak detection method with high threshold. (**c**) Result by our proposed method.

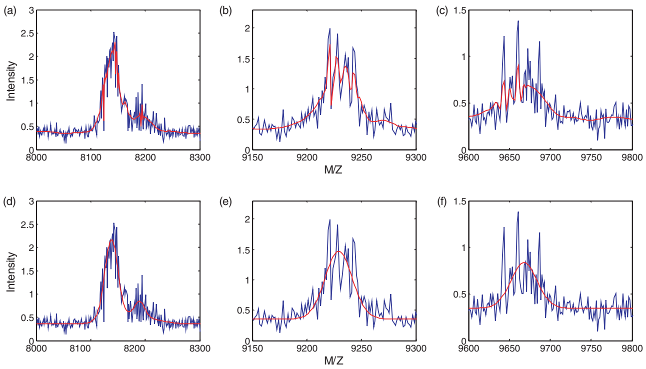

Fig. 4. (**a**–**c**) Result by the wavelet-based peak detection method. Raw spectrum is shown in blue. Denoised spectrum is shown in red. (**d**–**f**) Result by our proposed method.

by using a hard threshold cutoff in the wavelet-based denoising, denoised peak signal tends to have a lower intensity than true peak signal. In contrast, by well modeling the MS spectrum, true peak shape and intensity are well preserved after denoising. A detail comparison of the denoising result from both wavelet-based method and our proposed method is shown in Figure 4. We can see that although being able to detect small peak signals, wavelet method also keep much high frequency noise in the denoised signal.

Based on the simulation results, we can draw a receiver operating characteristic (ROC) curve of sensitivity versus FDR. From the ROC curve (Fig. 5) we can see that both the wavelet-based method using a low threshold and our method could achieve a maximum sensitivity of nearly 95%, i.e. both methods have the potential ability to detect nearly all peaks. Because relatively small peaks will be smoothed in denoising, the wavelet-based method using a high threshold can only reach a maximum sensitivity of ~80%, which means it could only detect 80% of the peaks in simulated spectra. We can also notice that by using an integrated peak detection method (performing baseline estimation, denoising and peak identification together); our algorithm can greatly reduce FDR while maintaining the same sensitivity.

3.2 For stroke data

We used high-throughput SELDI technique to screen 48 stroke patients and 32 healthy control patients to discover any potential new biomarkers for the diagnosis of ischemic stroke. To increase the coverage of proteins in SELDI protein profiles, the blood samples were fractionated with HyperD Q (anion ion exchange) into six fractions. The protein profiles of fractions 1, 3, 4, 5 and 6 were acquired with two SELDI Chips: IMAC and CM10. One 96-well anion exchange resin plate was used to fractionate samples into six discrete fractions (pH 9 + flow through, pH 7, pH 5, pH 4, pH 3 and organic wash) as previously described by Koopmann *et al.* (2004). Fractionation has been shown to greatly increase the number of proteins that can be resolved. ProteinChip arrays were analyzed utilizing a ProteinChip Reader, model PBSIIc (Ciphergen Biosystems Inc., Fremont, CA, USA). Protein spectra were externally calibrated using the All-in-One Protein Standard II (Ciphergen Biosystems Inc.) consisting of seven calibrants between 7 and 147 kDa. Data was collected between 0 and 200 kDa with the region between 2 and 20 kDa optimized.

First we use blank samples without protein sample to test the ability of our algorithm to do baseline estimation and denoising. Figure 6 is the spectrum using ProteinChip CM10 washed by the first fraction method. It can be seen that MS data <2000 Da is highly contaminated by irregular chemical noise and other noise from the system. So in our experiment, we only use data >2000 Da and discard the data at <2000 Da. Red line is the baseline information estimated by our method. Noise variance is shown below. It is clear that noise is not evenly distributed in the whole spectrum. Noise variance gradually decreases as *m*/*z* is increased. So denoising based on the whole spectrum is hardly able to correctly balance the false alarm result and low sensitivity. Therefore, subregion-based

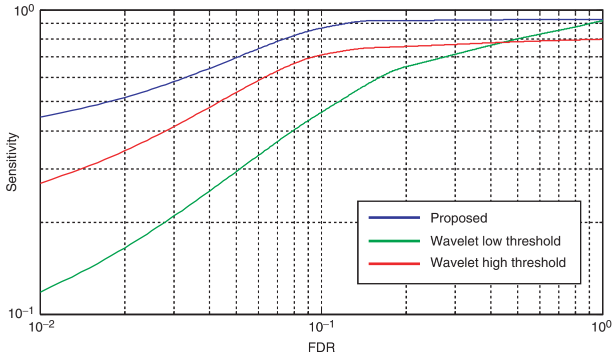

Fig. 5. ROC curve on simulated dataset. Axis is log-transformed.

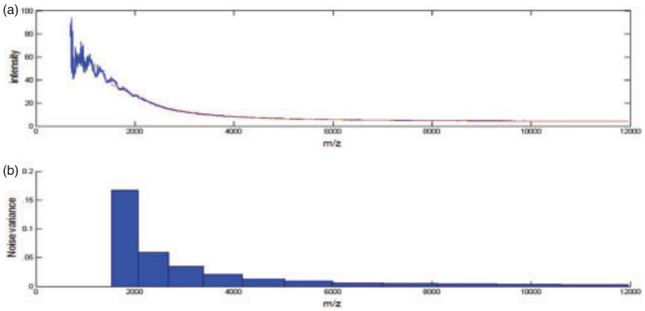

Fig. 6. (**a**) Raw blank spectrum (blue) and baseline information estimated by our proposed peak identification method (red). (**b**) Estimated noise variance.

denoising that estimates noise more accurately is clearly of greater advantage. Using a blank spectrum, we can also find that low-level polynomials (<3) are enough to fit baseline information.

The ultimate goal of biomarker discovery is to identify the locations of peaks where peak intensities have strong contrast between individuals from the control group and the disease group. Unfortunately, we do not have the prior information on whether or not the detected peak is a real peak generated by the biological sample. In order to compare the performance of our peak detection method with others, we consider both possible peaks and high confidence peaks manually selected by expert from the raw spectra as standard peaks for the evaluation. In the stroke MS data from the first fractionated samples profiled on CM10 ProteinChip arrays, we identified total of 32 high-confidence peaks and 198 possible peaks from 2000 to 100 000 Da in a total number of 80 spectra.

Since the major task of an automatic peak detection method is to locate all high confident peaks and possible peaks in the spectrum while not introducing false results, we compared the FDR between our proposed method and the wavelet-based peak detection method at each different sensitivity level. Sensitivity of high confidence peaks varied from 80% to 100% which stands for three conditions where 80/90/100% of manually identified high confidence peaks are correctly identified by peak detection method. We also studied another three conditions where a peak detection method could identify 70/80/90% of manually identified possible peak locations. FDR is calculated by the proportion of false-positive results and total positive results. From Table 2, we can see that when identifying high confidence peaks, both the wavelet-based and our proposed peak detection method have reasonable results if they only used to recognize 80% of high confidence peaks, and both methods will tend to more false-positive results as criteria becomes more strict. When it is required to detect 100% of high confidence peaks, wavelet-based method result in 22% FDR, which means 22% of their final results were false results; while in the same circumstance; only 14% of the results from our proposed method were incorrect. According to previous experiences of biomarker identification, high confidence peaks are often insufficient to differentiate a disease group from the control group. Peaks related to potential biomarkers are usually found in possible peak locations. When it is required to detect 90% of possible peaks from raw spectrum data, the wavelet-based method will result in 65% false positives, which suggests that each time wavelet-based method identifies a correct peak, it will produce approximately two false alarm results ($65\%/(1-65\%)\approx 2$). All of these false-positive results will put a huge burden on the following classifiers. By using an integrated processing and robust denoising techniques, our proposed peak detection method could greatly reduce the FDR while maintaining the same sensitivity.

Figure 7 presents a close view of aligned spectra from control group with *m*/*z* range from 3300 to 4200 Da. In this region, some obvious peaks together with some irregularly appearing peaks are highly contaminated by noise. In order to find reproducible peaks, we use the wavelet-based peak detection method and our proposed method to process the raw spectra and estimate the probability of whether a peak appears in each 5 Da. From the figure we can see that our proposed method works more effectively than wavelet-based method. The proposed method could clearly find out the locations of highly reproducible peaks while suppressing the interference from noise.

Table 2. FDR on stroke data

Method	High confident peaks			Possible peaks		
	80%	90%	100%	70%	80%	90%
Wavelet-based method	4%	16%	22%	39%	48%	65%
Proposed method	0%	9%	14%	19%	27%	38%

Comparison of FDR both from the wavelet-based and our proposed method at fixed sensitivity. There are 32 high confident peaks and 198 possible peaks from 2000 to 10 000 Da in a total number of 80 spectra.

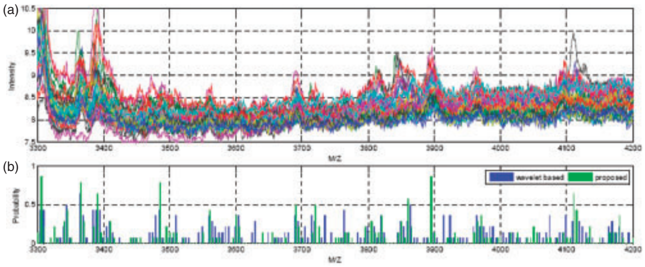

Fig. 7. (**a**) Raw MS data with simple baseline correction and alignment for better viewing. There are 32 spectra from control group. Each spectrum is from different person in control group. (**b**) Peak reproducibility estimation by the wavelet-based peak detection method (blue) and our proposed method (green) in this article.

Figure 8 is a 3D plot of 48 spectra from the disease group with a *m*/*z* range from 7300 to 8000 Da. The spectra were already processed with simple baseline removal and alignment for better viewing. Manually, we identified five possible peak locations in this region (7345, 7472, 7650, 7773 and 7858 Da). Peak detection results both from the wavelet-based method and our proposed method are shown under the spectra. The wavelet denoising threshold is 10. SNR is set at 8 for wavelet based method and 4 for our proposed method. From the results, we can see that peak detection results by our method are highly consistent through all spectra. All possible peaks have been successfully detected with only a small number of false-positive peaks in the final result. Yet, peak detection results from wavelet-based method are more ‘noisy’. For small intensity peaks as located in 7345 and 7472 Da, real peaks are more likely smoothed by wavelet denoising, whereas for small intensity peaks as located in 7650, 7773 and 7858 Da, some large noise are still left in the denoised spectra and become false positive in the final detection result.

4 CONCLUSION

Accurate peak detection from SELDI MS spectra is the most crucial part in SELDI-based biomarker identification. By proposing a novel mixture model, we can interpret the spectrum data more appropriately. We use a Bayesian approach to estimate parameters from the proposed mixture model, and use MCMC to perform Bayesian inference. By introducing a reversible jump method, we can automatically estimate the number of mixtures in the model. Compared to previous method, our method does not need to separate peak detection into substeps; therefore, minimizing the

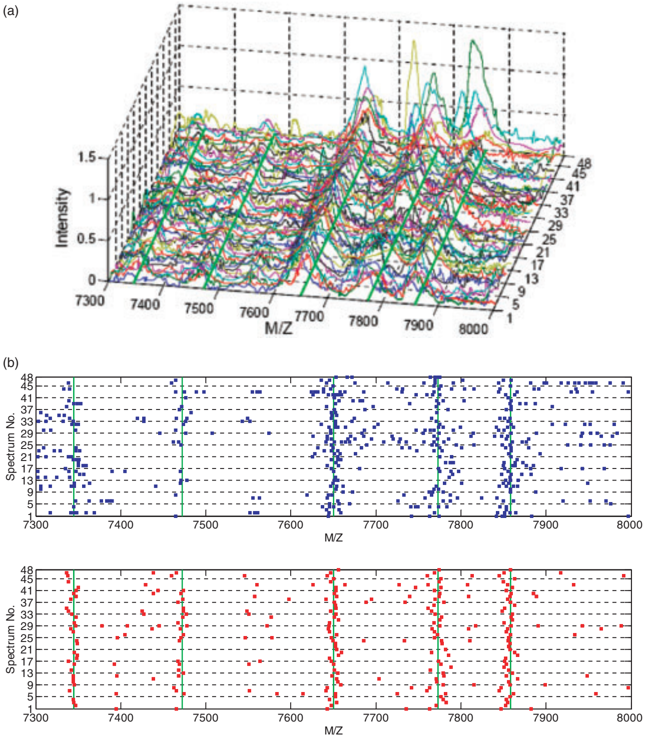

Fig. 8. (**a**) Raw data with simple baseline removal and alignment for better viewing. There are 48 spectra from disease group. Each spectrum is from a different person in the disease group. X-axis denotes m/z for each spectrum, and Z-axis denotes intensity. Y-axis denotes spectrum index. (**b**) Peak detection result by the wavelet-based method (blue) and our method (red).

risk of introducing errors while processing. Moreover, instead of requiring a manually selected denoising threshold, our method can automatically differentiate peak signal and noise signal without a preset parameter. Experiment results both on simulated spectra and stroke dataset show that our proposed peak detection method can greatly reduce FDR at the same sensitivity level.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the excellent collaboration with their collaborators in NIH.

Funding: X. Zhou would like to thank the support from the Methodist Hospital Research Institute Scholarship Award.

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature,* **422**, 198–207.

Andrieu,C. *et al*. (2001) Robust full Bayesian learning for radial basis networks. *Neural Comput.*, **13**, 2359–2407.

Baggerly,K. *et al*. (2003) A comprehensive approach to the analysis of matrix assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, **3**, 1667–1672.

Baggerly,K.A. *et al*. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments. *Bioinformatics*, **20**, 777–785.

Coombes,K.R. *et al*. (2005a) Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Inform.*, **1**, 41–52.

Coombes,K.R. *et al*. (2005b) 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*, **5**, 4107–4117.

Dijkstra,M. *et al*. (2006) Peak quantification in surface-enhanced laser desorption/ ionization by using mixture models. *Proteomics*, **6**, 5106–5116.

Dijkstra,M. *et al*. (2007) SELDI-TOF mass spectra: a view on sources of variation. *J. Chromatogr. B*, **847**, 12–23.

Du,P. *et al*. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22**, 2059–2065.

Fung,E.T. and Enderwick,C. (2002) ProteinChip clinical proteomics: computational challenges and solutions. *BioTechniques*, **81**(Suppl. 34), 40–41.

Green,P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hilario,M. *et al*. (2006) Processing and classification of protein mass spectra. *Mass Spectrom. Rev.*, **25**, 409–449.

Issaq,H.J. *et al*. (2002) The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. *Biochem. Biophys. Res. Commun.*, **292**, 587–592.

Koopmann,J. *et al*. (2004) Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. *Clin. Cancer Res.*, **10**, 860–868.

Malyarenko,D.I. *et al*. (2005) Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time series analysis techniques. *Clin. Chem.*, **51**, 65–74.

Morris,J. *et al*. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, **21**, 1764–1775.

Noy,K. and Fasulo,D. (2007) Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics*, **23**, 2528–2535.

Randolph,T.W. and Yasui,Y. (2006) Multiscale processing of mass spectrometry data. *Biometrics,* **62**, 589–597.

Sorace,J.M. and Zhan,M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinform.*, **4**, 24.

Tan,C.S. *et al*. (2006) Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics*, **22**, 1515–1523.

Vestal,M. and Juhasz,P. (1998) Resolution and mass accuracy in matrix-assisted laser desorption ionization- time-of-flight. *J. Am. Soc. Mass Spectrom.*, **9**, 892–911.

Vorderwulbecke,S. *et al*. (2005) Protein quantification by SELDI-TOF-MS-based ProteinChip system. *Nat. Methods*, **2**, 393–395.

Wang,X. *et al*. (2006) Feature extraction in the analysis of proteomic mass spectra. *Proteomics*, **6**, 2095–2100.

Yasui,Y. *et al*. (2003) An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *J. Biomed. Biotechnol.,* **4**, 242–248.