BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 6 2009, pages 808–814
doi:10.1093/bioinformatics/btp060

Data and text mining

A novel comprehensive wave-form MS data processing method

Shuo Chen[1,2,3,†], Ming Li[1,2,†], Don Hong[4], Dean Billheimer[5], Huiming Li[2], Baogang J. Xu[6] and Yu Shyr[1,2,*]

[1]Division of Cancer Biostatistics, Department of Biostatistics, Vanderbilt University, [2]Cancer Biostatistics Center, Vanderbilt-Ingram Cancer Center, Nashville, TN 37232, [3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, [4]Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, [5]Department of Oncological Sciences, University of Utah, Salt Lake City, UT 84112 and [6]Department of Cancer Biology, Vanderbilt University, Nashville, TN 37232, USA

Received on October 2, 2008; revised on January 22, 2009; accepted on January 23, 2009

Advance Access publication January 28, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Mass spectrometry (MS) can generate high-throughput protein profiles for biomedical research to discover biologically related protein patterns/biomarkers. The noisy functional MS data collected by current technologies, however, require consistent, sensitive and robust data-processing techniques for successful biomedical application. Therefore, it is important to detect features precisely for each spectrum, quantify them well and assign a unique label to features from the same protein/peptide across spectra.

Results: In this article, we propose a new comprehensive MS data preprocessing package, Wave-spec, which includes several novel algorithms. It can overcome several conventional difficulties. Wave-spec can be applied to multiple types of MS data generated with different MS technologies. Results from this new package were evaluated and compared to several existing approaches based on a MALDI-TOF MS dataset.

Availability: An example of MATLAB scripts used to implement the methods described in this article, along with Supplementary Figures, can be found at http://www.vicc.org/biostatistics/supp.php.

Contact: yu.shyr@vanderbilt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

With recent developments in proteomics, especially in mass spectrometry (MS) techniques such as MALDI-TOF MS and LC-MS/MS, protein profiles of tissue, serum and urine samples have become promising for detection of biologically related protein patterns (Adam *et al*., 2002; M'Koma *et al*., 2007; Yanagisawa *et al*., 2003; Yildiz *et al*., 2007).

The complexity and high dimensionality of MS data make quantitative analysis quite challenging. In practice, a biomedical experiment can generate hundreds or thousands of spectra. Each individual spectrum can be expressed as a graph of an intensity value with continuous wave shapes in a certain m/z range (tens of thousands of sampling pairwise data points) (Chen *et al*., 2007). The raw spectra contain not only peaks that represent proteins or biomarkers of scientific interest, but also substantial background noise. (Note: throughout this article, we use the term 'raw data' to refer to the MS data we obtained from the instrumentation and software used, prior to any manipulation on our part. Due to the software used, these data are in the m/z domain. Some MS software programs supply raw time-of-flight data.) MS spectra frequently exhibit random shifts on the m/z scale, and the correction for such variation is not easy, as we normally do not have explicit protein/peptide ID information for most peaks in the spectrum. Sample preparation conditions and laser intensity add other sources of intensity measurement variation. In addition, spectra may be acquired at different times with multiple replications per sample. As a consequence of this complexity, we divide analysis efforts into three major steps: the preprocessing step (feature extraction and quantification); the spectrum quality assessment step (to address reproducibility); and finally, the statistical analysis/data mining step (to select biomarkers/features of interest).

Although each part of the analysis is important, preprocessing is a crucial step and has great impact on evaluation of MS proteomics data reproducibility as well as on accuracy of biomarker identification. All current MALDI-TOF MS preprocessing methods share the same goal: to extract and quantify peaks of interest accurately and make the result applicable for further statistical analysis. In general, a widely accepted MS data preprocessing strategy follows these steps: spectrum calibration, denoising, baseline correction, normalization, peak detection, peak quantification and peak alignment (Morris *et al*., 2005; Wagner *et al*., 2003; Yasui *et al*., 2003). Despite its importance, challenges remain in preprocessing. Indeed, subjective parameter selection is required in almost all existing preprocessing methods, and makes reproducibility a particular issue. To improve the effectiveness and reproducibility of the preprocessing procedure, we developed a new set of algorithms based on feedback concepts, which enabled us to objectively target optimal parameter settings. The algorithms are incorporated in a package called Wave-spec. We summarize the methods and the Wave-spec package as follows:

1. Wave-spec introduces feedback, a widely used concept in engineering, into the preprocessing procedure. The parameter

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

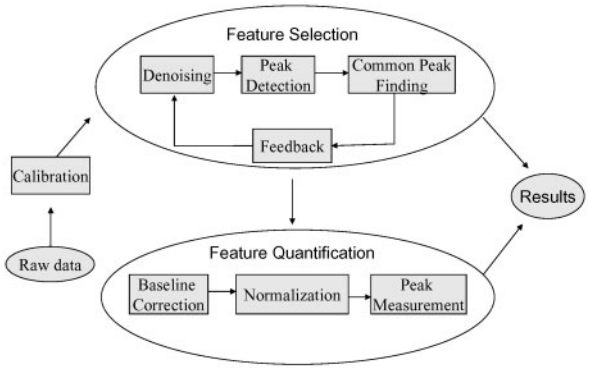

Fig. 1. Wave-spec package process chart.

settings are optimized through an objective iteration procedure; therefore, preprocessing reproducibility is greatly improved.

2. Wave-spec includes novel algorithms for finding common peaks across spectra and objective wavelet denoising parameter selection. The data-driven feedback procedures benefit from incorporating non-parametric kernel density estimation (KDE) for peak distribution across all spectra.

3. Wave-spec contains a unique feature that effectively and efficiently detects and quantifies isotopic peaks and aligns isotopic peak clusters across all spectra. Such data are typical of MALDI-TOF MS data acquired from a reflectron mass analyzer (Yu *et al*., 2006).

4. Wave-spec offers a new peak quantification method, which utilizes peak area information rather than the maximum point alone. This is a more robust measurement with less variation.

2 ALGORITHMS AND METHODS

Data Explore Software, the software component of the Voyager-Elite MALDI MS instrument (Applied Biosystems, Foster City, CA, USA), was used to obtain raw MS data. Each raw spectrum obtained using this software is composed of tens of thousands of pairwise data points (m/z versus intensity). Raw data were exported in American Standard Code for Information Interchange (ASCII) format [for a more detailed description of the data acquisition process, see Yildiz *et al*. (2007)]. Instrumentation and sample background noises are inevitably included in raw spectra, which need to be preprocessed before feature selection. The Wave-spec package consists of three major steps for preprocessing. First, it calibrates and unifies the m/z scales across all spectra. Second, it detects peaks and assigns unique IDs to the detected peaks. Finally, it quantifies peaks and provides intensities with corresponding m/z values for further analysis. The process chart for Wave-spec is illustrated in Figure 1.

2.1 Calibration

Raw spectra from MS instruments typically exhibit random shifts from their theoretical/ideal m/z positions. Such variation of peak position from spectrum to spectrum not only makes common peak finding difficult, but also makes intensity measurement inaccurate. To correct this, we carry out the first step of preprocessing: the calibration step.

The proposed calibration strategy is based on two desiderata: (i) the ideal peak should be bell shaped and (ii) the spectra should have only linear offset on the time domain. Therefore, the ideal peak shape can be approximated with a Gaussian density curve, $g(\mu,\sigma^2)$ (μ: the theoretical protein location; σ^2: estimated by the width of the known peak shape in the real data), and the spectra also can be transformed from the m/z domain to the time domain, calibrated linearly, then changed back to the m/z domain. Following are the details of the calibration process:

1. Select calibration peaks from the original spectra. The qualified candidate calibration peaks require two characteristics: (i) the ideal situation is that they show clear bell shapes (in practice, as long as there is a mode around the range of a known protein, the algorithm proposed will match that peak point to the known protein location) and (ii) their observed m/z values should be near known proteins (for which exact m/z values are known). For instance, the hemoglobin α, β with single and double charges will offer us four well-shaped calibration reference peaks in many mammalian spectra.

2. Convert the spectra from the m/z domain to the time domain t. The peak shapes in the original and ideal spectra can now be represented as $f(t)$ and $g(t)$, respectively. (Note: this step is unnecessary if time-of-flight data are directly available from the instrumentation/software used.)

3. With the spectra converted to the time domain, perform the calibration as follows. Convolve the raw spectrum peak intensities $f(t)$ with the ideal shape $g(t)$ over a finite range $[t_1,t_2]$
$$h(t)=g*f=\int_{t_1}^{t_2} g(\tau)f(t-\tau)d\tau$$
The maximum value $h(t_{max})$ of the convolution is obtained when f and g overlap the most. That is, if the peak in the original spectrum linearly shifts to position t_{max} from its $t_{original}$ position, we then make the spectrum profile as close as possible to its ideal/theoretical position. When selecting multiple known proteins, t_{max} is obtained by maximizing the sum of the convolution values of $h(t)$ on these multiple peak locations. We now have $t_{shift}=t_{original}-t_{max}$. The amount of recommended shift is $|t_{shift}|$ in the following direction:
$$|t_{shift}|\rightarrow\begin{cases}\text{shift to right,} & t_{original}<t_{max}\\ \text{shift to left,} & t_{original}>t_{max}\\ \text{keep the same,} & t_{original}=t_{max}\end{cases}$$

4. Convert the spectra from the time domain to the m/z domain.

This calibration strategy does not aim to align a single local maximum, but instead focuses on the entire peak shape. This has two advantages: (i) robustness (calibrating peaks using whole peak shape rather than peak maxima is robust to variation in maximum intensity among individual peaks) and (ii) effectiveness (using multiple known proteins' m/z positions calibrates over a wider

range, and the resulting accuracy will be greater across the m/z range).

2.2 Feature selection

As we know, the true features of each spectrum are blurred by random variation not only in their location, but also in their expression level. The major goal of the peak detection step is to identify the true features, with their m/z positions, among all the superficial peaks. To achieve this goal, we designed a feedback-based procedure. First, we detect the peaks of each spectrum by applying a time-invariant wavelet denoising technique for spectrum smoothing, and choose the local maximum as a peak location. Second, we define the common peaks across the entire collection of spectra using non-parametric KDE for peak distribution. The height of the baseline of peak distribution provides an index reflecting the denoising effects of the current wavelet denoising parameter settings; we update the denoising parameters to recalculate this index, then iterate this procedure until the index is stabilized and minimized to reach a certain noise tolerance upper limit. The optimal denoising parameter settings are those providing this final index. With a set of optimal parameters, the reproducibility of preprocessing may be improved. Following is a detailed description of this procedure.

2.2.1 Peak detection on a single spectrum Previous work has proven that time-invariant stationary discrete wavelet transform or undecimated discrete wavelet transform wavelets with a hard threshold can provide sound performance for denoising (Coombes *et al*., 2005).

Three parameters need to be set: a basis for the wavelet type, a decomposition level and wavelet coefficient thresholds. The choice of wavelet basis will not significantly affect denoising (Coombes *et al*., 2005). On the wavelet domain, we can set the decomposition level empirically based on time–frequency energy distribution, which balances smoothness and signal loss (Chen *et al*., 2007). The most important denoising parameters are wavelet coefficient thresholds, and they are the focus of our procedure; we initially set the parameters for smoothing, then update them to the optimum using a feedback index. After denoising, the local maxima (peaks) become valid surrogates for the true features of the spectrum. Furthermore, the denoising procedure reduces the false peak discovery rate as the small bumps are mostly removed.

2.2.2 Common peak finding After single spectrum peak detection, we get a peak (local maximum) list for each spectrum. However, such a list of peaks still cannot be claimed as the true set of feature locations, as the number of peaks might differ across different spectra. Further, a true feature may be represented by different m/z locations on different spectra. The purpose of common peak finding is to infer the true feature locations from the peak list of all spectra in the whole dataset. Therefore, a procedure is required to identify which peaks come from the same ions; in other words, we need to define certain boundaries to mark peaks within a boundary as the same feature across the whole dataset.

Several algorithms have been developed to solve this problem. Coombes *et al*. (2005) defined peaks within a fixed m/z or time range as one bin. The windowing approach is then repeated across the spectrum. Morris *et al*. (2005) computed a mean spectrum, and used each local maximum's two adjacent minima as the boundaries. Tibshirani *et al*. (2004) applied a clustering idea to assign the peaks with similar m/z as one bin according to their similarity in the dendrogram.

In this section, we propose a new method that utilizes the distribution of peak locations. That is, we apply a non-parametric KDE method to model peak location distribution. 'Bumps' in the peak distribution indicate the location and extent of features in the spectra. Intuitively, the higher the bump (because of a greater number of peaks at that m/z), the more likely it is to correspond to a peak common across spectra.

The details of the implementation are as follows. Let $X_1, X_2, ..., X_n$ denote a sample of the total number of n peak m/z values (locations) drawn from probability density function f. The kernel density estimate of f at the point x_i is given by:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where the kernel K satisfies $\int K(x)dx = 1$ (Sheather, 2004). This is a convolution of the sample empirical distribution $\hat{F}$ and the kernel function K_h, $(\hat{F} \star K_h)(x)$. The reason we choose a Gaussian kernel is that, with infinite support, it performs more robustly to peak m/z variation than do kernels with finite support such as Epanechikov or tri-cube kernels. For a Gaussian kernel, the bandwidth equals the standard deviation.

Again, our primary interest is to detect the location and boundaries of the true features, with their m/z locations, using the KDE method. We define the local maxima of the KDE as a common feature's m/z location, and the two adjacent local minima as its boundaries. In this way, we identify the true peak locations and distinguish one from another efficiently and effectively. We could also assume the distribution pattern as a multiple normal mixture model and apply an EM algorithm to estimate μ_i, σ_i of each mode. However, we seek only the cutoff points between two adjacent bell shapes, which can be easily accessed with a non-parametric estimate, while estimating μ_i, σ_i through EM could be computationally expensive and error-laden.

2.2.3 Optimizing wavelet denoising parameters through feedback The estimation procedure of peak location distribution provides an 'index' that can be used to evaluate the performance of the denoising. If the denoising procedure is not stringent enough, more noisy peaks will be admitted to the peak list; those falsely discovered peaks are not true features, but instead randomly distributed on the m/z axis. As a result, the estimated peak location distribution curve will show an elevated baseline with a height reflecting the proportion of falsely detected peaks associated with wavelet threshold levels. Therefore, we can utilize the peak location distribution baseline information to adjust the wavelet threshold level. Supplementary Figure 1 shows how we define such a feedback index. The index is defined as $B/(A+B)$, the ratio of baseline area to total area under the distribution curve. In terms of getting rid of false peaks, we hope to see the feedback index as low as possible, as baseline decreases when the threshold level is high. On the other hand, we cannot require the feedback index to be too small, as the true features might also be removed with an overly stringent threshold. To balance the tradeoff between admitting false peaks and removing true peaks, we apply the following schema: first, we pre-specify the feedback index upper limit (e.g. 0.05). Then, we increase the wavelet threshold from relatively low levels until the feedback

index is low enough to pass the upper limit. The idea of having a feedback index upper limit is similar to the idea of setting a significance level for a statistical test, to judge the performance of the test. Although the choice of such a value is still arbitrary, it is more robust than the choices of other wavelet denoising parameters.

Through all these iterative processes, we determine the optimal wavelet threshold levels to determine true feature locations and their boundaries more accurately. This framework also can be applied to other feature-interested analyses of functional biomedical data.

2.3 Feature quantification

Feature quantification includes three steps: baseline correction, normalization and peak measurement. For baseline correction, we detect local minima using sliding windows on the denoised spectrum and then fit these local minima to a smooth curve with a spline (Chen *et al*., 2007). We apply the total ion current (TIC) method to normalize all spectra, which enforces the constraint of equal TIC for each spectrum in the dataset (Morris *et al*., 2005). Peak measurement is the last step to quantify the common features for each spectrum. Most current methods use the height of the local maximum to quantify the feature within estimated boundaries. However, point measurement may be subject to high variation from various sources. Also, height may not be a good index of the total amount of ions for a specific feature. Measuring a small region or bounded neighborhood around each peak would be more robust and informative; using small region measurement results in smaller coefficients of variation (Section 3.2).

3 APPLICATION AND EVALUATION

As described in the following three subsections, we applied the Wave-spec package to a publicly available MALDI-TOF MS dataset (Taguchi *et al*., 2007). In addition to demonstrating the capabilities of the Wave-spec package, analysis of these data also allowed us to compare the package with two other existing preprocessing packages. To complete our usage of Wave-spec, in the final subsection we show how Wave-spec can be applied to MS data with isotopic information, for instance, MALDI-TOF MS on reflectron mode or TOF-TOF MS data.

3.1 Preprocessing on MALDI-TOF MS data

In the serum MALDI-TOF dataset mentioned above, 71 spectra were obtained using a Voyager DE-STR MALDI-TOF mass spectrometer (Applied Biosystems, Foster City, CA, USA). Positive ion mass spectra were acquired in linear mode. A total of 500–525 independent spectra for each sample were averaged to generate each spectrum. We applied Wave-spec, and the results are summarized below. The results show that the spectra are calibrated to the correct m/z scales. Figure 2 shows, for example, spectra before and after calibration for the range 8740–8860 Da. Figure 2A shows all spectra in this range, suggesting two possible peaks, though they are not aligned well. Figure 2B shows two clear peak shapes, aligned well after calibration. Similar results are found in other regions. For individual peak detection and common peak finding, we applied the wavelet denoising strategy by initially setting the threshold parameter to 20-fold of the MAD/0.67 on each spectrum in this dataset (Coombes *et al*., 2005). Then, based on the peak

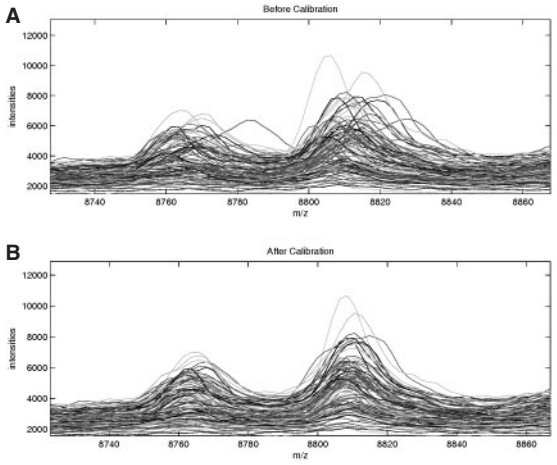

Fig. 2. Calibration effects: before (**A**) and after (**B**).

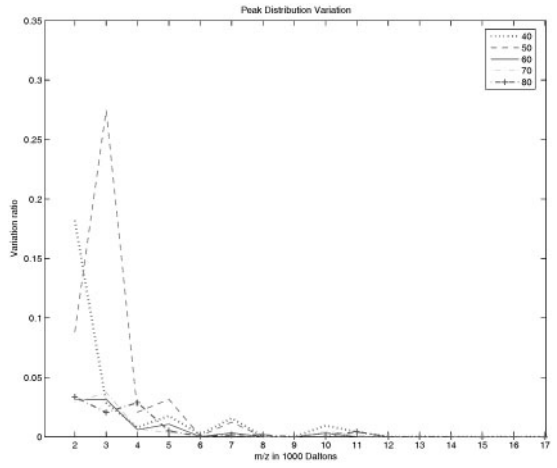

Fig. 3. Wavelet threshold selection by feedback.

list, we generated the peak density curve by KDE and estimated the baseline proportion of peaks. We repeated this procedure, gradually increasing wavelet denoising parameters from 20 to 100 by increments of five. The more peaks admitted, the more false positive peaks are included; consequently, the m/z location of the peaks shows a higher variation, which increases the baseline proportion of the peak distribution curve.

By setting an upper limit, the cutoff point of 0.05, we selected the optimum threshold parameter with respect to the number of peaks. Within the limits, we normally choose the lowest wavelet threshold level, as it will allow more common features to be detected. Applying this criterion to Figure 3, we chose a wavelet threshold of 60. In the meantime, spectrum baseline correction and normalization were carried out, and Supplementary Figure 2 shows the average curve of the baseline-corrected and normalized spectra with common peak boundaries for the region from 11 000 Da to 17 000 Da. Within the boundaries, we quantify the peak with the area under the curve (AUC). The output of the set of preprocessing steps are common peak IDs, boundaries and expression levels for each spectrum.

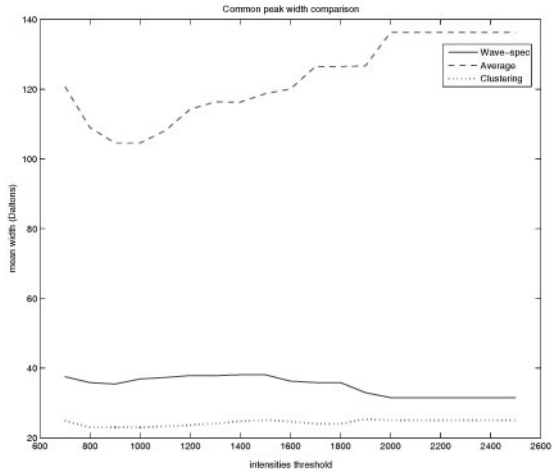

Fig. 4. Method comparison: common peak boundary width.

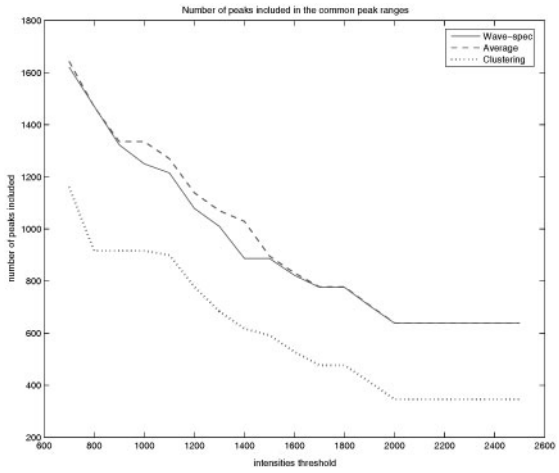

Fig. 5. Method comparison: number of peaks within boundary.

3.2 Evaluation of Wave-spec compared to existing preprocessing methods

To evaluate Wave-spec, we compared the feature extraction and quantification performance of Wave-spec to that of existing algorithms, using the above 71 human serum MALDI-TOF MS samples. In our experience, feature/peak detection across all spectra is one of the most challenging parts of MS data preprocessing. The general procedure for feature detection includes individual spectrum peak detection, finding common peaks across spectra, and assigning common peaks to an individual spectrum; these processes were the focus of our evaluation of Wave-spec versus the two existing packages.

The individual spectrum peak detection procedure relies primarily on the denoising technique. The wavelet method, as a powerful signal processing tool, performs well for mass spectrum denoising (Chen *et al*., 2007; Coombes *et al*., 2005). The number of features/peaks detected is basically determined by the wavelet denoising threshold level. Based on an average spectrum selected from the 71 human serum MALDI-TOF MS samples, we applied various threshold parameters for denoising, and different numbers of peaks were detected, though some common peaks were found consistently for any threshold parameter (see Supplementary Table 1 for details). Clearly, the choice of threshold parameter affects potential features that can be detected; however, few (if any) existing software packages or methods can set wavelet denoising parameters in an objective way. Using Wave-spec's feedback concepts, we are able to provide relatively data-driven objective parameters for denoising.

Peak alignment across spectra is even more difficult, but it is an indispensable step to acquire the $n \times p$ feature-sample matrix for further statistical analysis. Recently, some researchers have tried to solve this problem using different strategies: Tibshirani *et al*. (2004) proposed an effective peak alignment algorithm based on a hierarchical clustering idea; Morris *et al*. (2005) used the mean spectrum to find common peaks, which avoids the peak alignment step. In fact, a good common peak finding algorithm should have the following properties: (i) it can define a specific common feature, say $mz1$, by setting appropriate boundaries on the m/z domain; (ii) any peak detected on a single spectrum within the boundaries should be identified as $mz1$; and (iii) the boundaries should be wide enough to include all peaks on an individual spectrum that correspond to $mz1$, but not too wide to cover two or more other features. To evaluate the performance of different algorithms, we can set up metrics to assess these properties, such as number of total common peaks, common peak window width (a precision metric) and number of peaks included within a certain window (an efficiency metric). An ideal common peak finding method should include the maximum number of true peaks from an individual spectrum in a relatively narrow range. The evaluation results are summarized in Figures 4 and 5. Figure 4 shows the differences among the three methods in terms of average common peak width. Figure 5 shows the number of peaks included in common peak ranges. In both cases, we evaluated these metrics at different intensity thresholds; by using intensity thresholds on the average spectrum, we ignore the low-intensity peaks that are more likely to be noisy peaks. The results indicate that the Wave-spec package found a shorter common feature boundary range, but included more peaks in that range.

Averaging, as a fundamental principle underlying many statistical methods (Morris *et al*., 2005), sheds some light on the general peak distribution of a dataset. Averaging is robust and easy to perform. However, the potential risk of only considering the average spectrum is that we might be penalized for ignoring intensity heterogeneity in some regions across spectra. To illustrate using this dataset, we chose regions with abundant peaks and high variation. For instance, we chose the region of m/z 11 300–11 800 Da. As we see in Figure 6, there are many peaks (fragments of serum amyloid A) in this region. Some spectra are flat; others have high intensities; and many have peaks around 11 450 Da and 11 750 Da, both of which are next to higher peaks. On the mean spectrum, however, there is no clear bump for the peaks around 11 450 and 11 750 Da (Fig. 6).

Although Wave-spec and the mean spectrum method provide a similar common peak set, some differences still exist (Figs 7 and 8). Figure 7 shows common peaks identified by the mean spectrum method (Morris *et al*., 2005), while Figure 8 shows common peaks identified by Wave-spec. The x-axis shows m/z range from 11 000 to 12 400; the y-axis shows intensity. The superimposed rectangles highlight the boundaries of common peaks detected by each method.

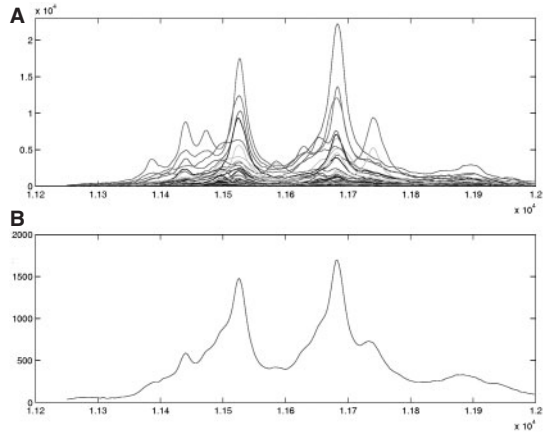

Fig. 6. A high-variation region: all spectra (**A**) and average spectrum (**B**).

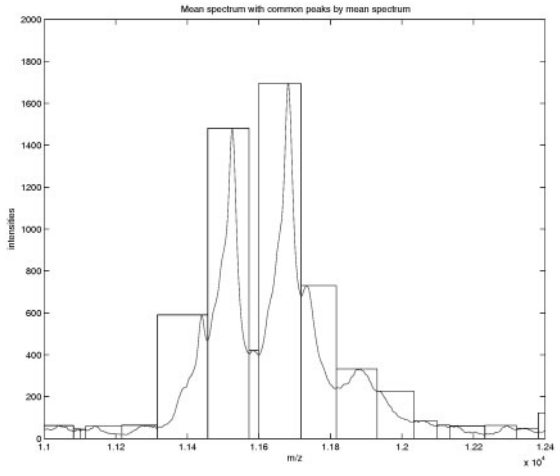

Fig. 7. Common peaks identified by mean spectrum method.

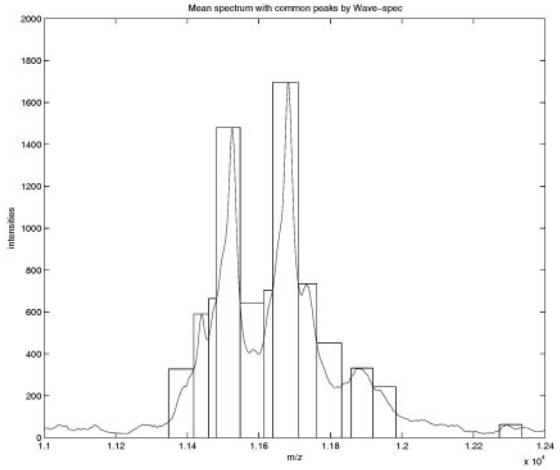

Fig. 8. Common peaks identified by Wave-spec.

From the plots, we can tell that: (i) Wave-spec detected some peaks in many spectra of a dataset that are not identified on the mean spectrum, especially for regions with high variation across spectra, and (ii) Wave-spec defined common peaks across spectra within narrower boundaries. The mean spectrum method defines the range of a common peak based on two local minima adjacent to a local maximum, and it can be too wide to include multiple different features. The clustering method of Tibshirani *et al*. (2004) splits peaks into subgroups, but it is difficult to set accurate boundaries using this method.

For feature extraction, we used the coefficient of variation (CV) to compare region-based and maximum height-based feature measuring methods. The result can be seen in Supplementary Figure 3; the paired t-test showed that the AUC method yields smaller CVs on all common features than does the single maximum point measurement method, with $P < 0.001$.

3.3 MALDI-TOF MS reflectron-mode data for frog-skin fluid samples

One unique feature of Wave-spec is that it can handle MALDI-TOF MS data acquired in reflectron mode, which has not yet been discussed in the preprocessing literature. Such data have the advantage of high resolution in detecting protein isotopes. (For more details about reflectron-mode data, see http://keck.med.yale.edu/prochem/procmald.htm.) However, preprocessing of MALDI-TOF MS reflectron-mode data is especially difficult because of the high resolution and prevalence of isotopes. Generally, such preprocessing requires advanced signal processing techniques to extract features by envelope signal detection prior to peak detection, which introduces variation and makes quantification more complex. Wave-spec, however, offers an effective and efficient way to detect all peaks and align clusters of peaks across spectra. Also, quantification by area under all isotopes is robust.

The major challenges of high-resolution data are to cluster a group of isotope peaks as one protein/peptide, to align peak clusters across spectra and then to quantify them. Denoising is not our major concern here, as high-resolution data are relatively clean. Supplementary Figure 4 shows a global view of spectra with isotope peaks, as well as a magnification of a smaller range for a closer look. In this figure, we see the clear pattern of isotope peak clusters. We first calibrated the spectra on the time domain (Supplementary Figure 5 shows data before and after calibration); next we followed standard MALDI-TOF MS data preprocessing steps: denoising, baseline correction, normalization and peak selection on individual spectra. Then, the kernel method was applied to estimate the peak distribution curve. Since KDE is also a low-pass filter, the curve shows one large peak rather than multiple bumps when there are clusters of peaks across spectra (Supplementary Fig. 6). Following the feedback procedure described in Section 2, we then detected the boundaries of different features across spectra and quantified them by measuring the region between the boundaries. In this way, we detected and quantified features across spectra acquired in reflectron mode. Given common peaks' m/z, one can obtain the expected isotopic distribution, e.g. by the use of averagine (Senko *et al*., 1995). We also were able to detect and quantify isotope information for each feature on individual spectra.

4 DISCUSSION AND CONCLUSION

The major contribution of this article is to incorporate wavelet denoising techniques and KDE into an iterative feedback procedure. This allows us to obtain optimal parameter settings through a relatively objective, data-driven process, which may increase reproducibility. The feedback concept has been widely applied to engineering in areas such as control theory and signal processing. In our comparison of feature-detection methods, Wave-spec outperformed the two alternative algorithms (Morris *et al*., 2005; Tibshirani *et al*., 2004) in terms of both precision and efficiency (with, respectively, a narrower common peak window width than the mean spectrum method, and more peaks included within a certain window than the clustering method). From another angle, the Wave-spec feedback process can be viewed as a novel wavelet shrinkage algorithm. Unlike the traditional wavelet shrinkage methods, such as SURE, developed by Donoho and Jonestone (1995), Wave-spec uses the features' variation index as the penalty item. With a preset cutoff, it can automatically approximate the optimal wavelet threshold values for an MS dataset. Wave-spec provides a framework that can deal with different types of MS data: MALDI/SELDI TOF MS data, mass spectra with isotopic peaks and full-scan LC MS data. The use of complex peptide and metabolite mixtures in LC MS requires alignment in two dimensions: the m/z dimension and the retention time dimension (Listgarten and Emili, 2005). The Wave-spec method can be applied to detect common peaks at different retention times on the m/z domain and, through the feedback mechanism, will choose optimum peak filtering parameters objectively. The framework also is applicable to 'feature-interested' analysis of functional one-dimensional or two-dimensional biomedical data, for instance, 2D gel proteomics data. Compared to alternative algorithms, the Wave-spec framework can extract and quantify features more accurately and robustly from functional biomedical data. In summary, detection and quantification of functional biomedical MS data features is the key step for data mining, as these features are the only information source for further analysis. The algorithms developed and adapted in Wave-spec ensure that the results of preprocessing possess the desired qualities of precision, efficiency, robustness and reproducibility. Adopting feedback concepts, Wave-spec has the ability to obtain optimal parameter setups, thus ensuring a more accurate common peak list across all spectra and improving biomarker identification and profile pattern recognition.

ACKNOWLEDGEMENTS

The authors are grateful to collaborators for MALDI MS projects that motivated this research. In addition, we also wish to thank Joan Zhang for her extensive work with the Wave-spec package, which has given her the expertise to help others with Wave-spec questions, and Shaun Haskins and Lynne Berry for their editorial work on this manuscript.

Funding: Lung Cancer Special Program of Research Excellence (SPORE, 2P50 CA090949-06A1); Breast Cancer SPORE (5P50 CA098131-05); GI SPORE (2P50 CA095103-06); Cancer Center Support Grant (CCSG, 5P30 CA068485-12).

Conflict of Interest: none declared.

REFERENCES

Adam,B.L. *et al*. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, **62**, 3609–3614.

Chen,S. *et al*. (2007) Wavelet-based procedures for proteomic mass spectrometry data processing. *Computat. Statist. Data Anal.*, **52**, 211–220.

Coombes,K.R. *et al*. (2005) Improved peak detection and quantification of mass spectrometry data acquired from SELDI by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, **5**, 4107–4117.

Donoho,D.L. and Johnstone,I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.*, **90**, 1200–1224.

Listgarten,J. and Emili,A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography tandem mass spectrometry. *Mol. Cell Proteomics*, **4**, 419–434.

M'Koma,A.E. *et al*. (2007) Detection of pre-neoplastic and neoplastic prostate disease by MALDI profiling of urine. *Biochem. Biophys. Res. Comm.*, **353**, 829–834.

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

Senko,M.W. *et al*. (1995) Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.*, **6**, 229–233.

Sheather,S.J. (2004) Density estimation. *Statist. Sci.*, **19**, 588–597.

Taguchi,F. *et al*. (2007) Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J. Natl. Cancer Inst.*, **99**, 838–846.

Tibshirani,R. *et al*. (2004) Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, **20**, 3034–3044.

Wagner,M. *et al*. (2003) Protocols for disease classification from mass spectrometry data. *Proteomics*, **3**, 1692–1698.

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.

Yanagisawa,K. *et al*. (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet*, **362**, 433–439.

Yildiz,P. *et al*. (2007) Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. *J. Thorac. Oncol.*, **2**, 893–901.

Yu,W. *et al*. (2006) Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Comp. Biol. Chem.*, **30**, 27–38.