Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 211–220
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Wavelet-based procedures for proteomic mass spectrometry data processing

Shuo Chen[a], Don Hong[b,*], Yu Shyr[a]

[a] *Biostatistics Shared Resource, Vanderbilt Ingram Cancer Center, Vanderbilt University, Nashville, Tennessee, USA*
[b] *Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, Tennessee 37132, USA*

Available online 2 March 2007

**Abstract**

Proteomics aims at determining the structure, function and expression of proteins. High-throughput mass spectrometry (MS) is emerging as a leading technique in the proteomics revolution. Though it can be used to find disease-related protein patterns in mixtures of proteins derived from easily obtained samples, key challenges remain in the processing of proteomic MS data. Multiscale mathematical tools such as wavelets play an important role in signal processing and statistical data analysis. A wavelet-based algorithm for proteomic data processing is developed. A MATLAB implementation of the software package, called WaveSpect0, is presented including processing procedures of step-interval unification, adaptive stationary discrete wavelet denoising, baseline correction using splines, normalization, peak detection, and a newly designed peak alignment method using clustering techniques. Applications to real MS data sets for different cancer research projects in Vanderbilt Ingram Cancer Center show that the algorithm is efficient and satisfactory in MS data mining.


*Keywords:* Proteomic data processing; Biomarker discovery; Mass spectrometry; Splines; Wavelets

## 1. Introduction

Mass spectrometry (MS) is an attractive analytical tool in proteomics research. It can directly measure the complex mixture of proteins/peptides obtained from the biological tissues or cells. By comparing proteomic expressions between the biological samples from the controls and cases, one could find the disease-related proteomic patterns and biomarkers. That holds tremendous potentials for disease early diagnosis, prognosis, as well as drug target research. In practice, a biological experiment can generate an MS data set contains tens of thousands of spectra, and each spectrum contain tens of thousands sampling pairwise data points ($m/z$ location and the corresponding intensities). Analysis of these data typically relies on inferring the existence of a peptide of a particular $m/z$ from the existence of a spike in the spectrum. This process is confounded by the variability in the $m/z$ location and the shape/size of features when compared across samples. Also, the nontransparent nature of the data structure complicates the data analysis (Pedrioli et al., 2004). Therefore, the MS data processing is a very challenging task. With comparisons from another type of high-throughput

* Corresponding author. Tel.: +1 6159048339; fax: +1 6158985422.
*E-mail address:* dhong@mtsu.edu (D. Hong).

0167-9473/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.csda.2007.02.022

biomedical data genomic expression data (e.g., microarray, SNP), MS data containing the protein information peak in the continuous time-scale data frame with substantial variances such as baselines and noises and we do not have much prior information of the peaks (the peaks from what protein/peptide and the electronic charge) in the spectrum. A commonly used model (Coombes et al., 2005; Hong and Shyr, 2007; Hong et al., 2007, for example) is that each raw spectrum data can be represented in three parts: the true signal, noises, and a baseline artifact. More explicit description of these three distinct components with mathematical physics background can be found in Hong and Shyr (2007) and Hong et al. (2007). This model allows us to address signal reconstruction and subsequent biological interpretation in a mathematically principled manner. The mathematical processing of MS signals can be roughly divided into two steps. First, in the "preprocessing" step, we attempt to recover from the time of arriving data as accurately as possible, the "true" signal reflecting the mass/charge distribution of the ions originating from the sample. The preprocessing step includes registration, denoising, baseline correction, peak selection, and across samples peak alignment. In this step, these operations are performed independently of any biological information one seeks to extract from the data. The second type of processing attempts to represent the data in a form that facilitates the extraction of biological information. This step involves operations such as dimension reduction, feature selection, clustering, and pattern recognition for classification. Recently, more articles in literature address both preprocessing and processing issues (Hong et al., 2007; Tibshirani et al., 2004; Wagner et al., 2003; Yasui et al., 2003; Yu et al., 2006). In this paper, a wavelet-based procedure for more exploratory data analysis is proposed in company with a better understanding of the MS data signal's properties. The application of wavelet analysis is not only for the data denoising, but also for the feature extraction and pattern analysis. A wavelet smoothing procedure begins with a choice of wavelet family followed by a choice of thresholding method, each method involving parameters that control the amount of variation. To get visible features using wavelet smoothing, Morris et al. (2005) calculated the mean spectrum denoised by using wavelet transform and then define a peak by the range between flanking local minima and a local maximum. Here, the proposed point of view is that extracting scale-based signal content (defined by local differencing, not local averaging) and using the frequency-domain-like histograms to determine high-density regions (bins) of feature locations allow for a focus on features that distinguish themselves across all spectra.

The main contributions of this paper include analyzing MS data by wavelets method on both time domain and frequency domain, and utilizing a more adaptive wavelet method for data denoising; introducing a MS data analysis method on the wavelet domain after wavelet coefficients shrinkage which is more robust for peak selection and quantification; developing a comprehensive package, WaveSpect0, for the MS data processing. Mathematical tools and statistical techniques applied in this study involve splines for baseline correction, wavelets for adaptive denoising, and multivariate statistical techniques such as clustering analysis and signal processing techniques are combined for evaluating the complicated biological signals. A MATLAB-based software package is implemented for proteomic MS data processing, especially on the matrix assisted laser desorption/ionization (MALDI) time-of-flight (TOF) MS data. The package takes a raw proteomic MS data set, an $m \times 2n$ matrix ($n$ spectra with $m$ rows), as an input, undergoes all the above procedures, then generates an output, a $p \times 2n$ matrix ($n$ spectra with $p$ features and corresponding locations). Innovations in this package include: unifying the step interval of discrete data by spline functions on the $m/z$ domain, signal-self-oriented adaptive stationary discrete wavelet denoising, and the center alignment common feature finding algorithm. Applications to real MS data sets for different cancer research projects at the Vanderbilt Ingram Cancer Center (VICC) show that the algorithm is efficient and satisfactory in MS biomarkers discovery (Rahman et al., 2005; Xie et al., 2005; Xu et al., 2005; Yildiz et al., 2006). The idea of the algorithms design can be applied to functional data processing in general.

The remainder of the paper is organized as follows. In the next section, wavelet application to MS data as a denoising tool is discussed. In Section 3, the entire procedure of all MS data processing steps in WaveSpect0 is presented in detail. Examples to test the package's performance with some comparisons are shown in Section 4.

## 2. Wavelets and applications in proteomic MS data analysis

Wavelet theory is developed now into a methodology used in many disciplines. Wavelets also provide a rich source of useful tools for applications in time-scale types of problems. The attention of wavelets was attracted by statisticians when Mallat (1989) established a connection between wavelets and signal processing. Wavelet thresholding has desirable statistical optimality properties (Donoho and Johnstone, 1994, 1995). Since then, wavelets have been proved very useful in nonparametric statistics and time series analysis.

In recent years, wavelets have been applied to a large variety of biomedical signals. There is a growing interest in using wavelets in the analysis of sequence and functional genomics data (Aldoubi and Unser, 1996; Hong and Shyr, 2006; Lió, 2003). In this section, we will discuss wavelets' application in the proteomic MS data. Since the data are one dimensional and discrete, we will use 1-D discrete wavelets transforms. We carry out both data analysis and denoising process for MALDI-TOF MS data by using a stationary discrete wavelet transform (SDWT), which is shift-invariant and yields better visual and qualitative denoising, with a small added cost in computational complexity (Coombes et al., 2005).

### 2.1. Wavelets for MALDI-TOF MS Data

As a powerful tool for signal analysis, wavelets can be used to describe signals of long time intervals where we desire more precise low-frequency information and shorter regions where we want high-frequency information, by the variable-sized windows technique. In that way, by using wavelet analysis we can not only analyze the information on the frequency domain but also the information on the time domain. Nason and Silverman (1995) modified the basic discrete wavelet transform (DWT) algorithm to give a stationary wavelet transform that no longer depends on the choice of the origin. To apply stationary DWT, also called the undecimated DWT, the length of the discrete signal is required to be a multiple of $2^k$.

In MALDI data processing, using stationary wavelet transform, we adjust the data length to satisfy this requirement by padding them out by reflecting the intensities at the very end of $m/z$ region. Applying stationary discrete wavelet transform to the MALDI-TOF MS data, we get the wavelet coefficients by decomposing the MALDI-TOF MS data to the 12th level. We notice that the lower level (high-frequency) wavelet components are similar to a random process while the higher level (low-frequency) ones are not (Fig. 1).

In Fig. 2, the histograms give a better explanation by showing that the lower level coefficients would have better bell-shape with comparison to the higher level ones. The Kolmogorov–Smirnov tests were taken with the cumulative function of Cauchy and normal distribution, but the $p$-values would not differentiate the levels of coefficients distinctly. The skewness and kurtosis across levels (Fig. 3(L)) show us that the coefficients are symmetric and the thickness of tails is almost the same. The energy-frequency analysis is always a key point to understanding the signal. The frequency power density distribution is given in Fig. 3 (R). MALDI-TOF MS data sets collected from VICC show that all spectra have the relatively similar energy-frequency distribution, especially at the lower levels. Generally, for all spectra, the energy is mainly distributed at the coefficients over the 4th level. Due to the variation of across sample data sets these exploratory data analyses are important for data preprocessing.

### 2.2. Wavelet denoising strategy

In MALDI-TOF MS data processing for cancer study, the denoising process is critical since the feature selection relies only on the information from the denoised signals.

We assume that most of the noises are white Gaussian noises for all frequency levels. Some application observations of MALDI-TOF MS data include that there are more high-frequency signal components in the low $m/z$ value region, and that the higher peaks are often associated with the higher variations of intensity. A basic wavelet-based denoising procedure can be described in the following:

*Decomposition*: Select the level $N$ and type of wavelets, then determine the coefficients of the MS signal by SDWT.

For wavelet denoising, we should decide over many selections, such as the types of mother wavelet, the decomposition levels and the values of thresholds in the next step. In the decomposition of MALDI-TOF MS data from VICC projects, we apply Daubechies' wavelets of degree 6 or 8 and the decomposition level of 4 or 6.

*Thresholding*: Estimating the threshold values, based upon the analysis and empirical method.

For each level from 1 to $N$, use the estimated threshold values and set the detail coefficients below the threshold values to zero. We can choose a soft threshold as well. For hard thresholding, the thresholded coefficient $x$ is determined by

$$x = \begin{cases} x, & |x| > t, \\ 0, & |x| < t, \end{cases}$$

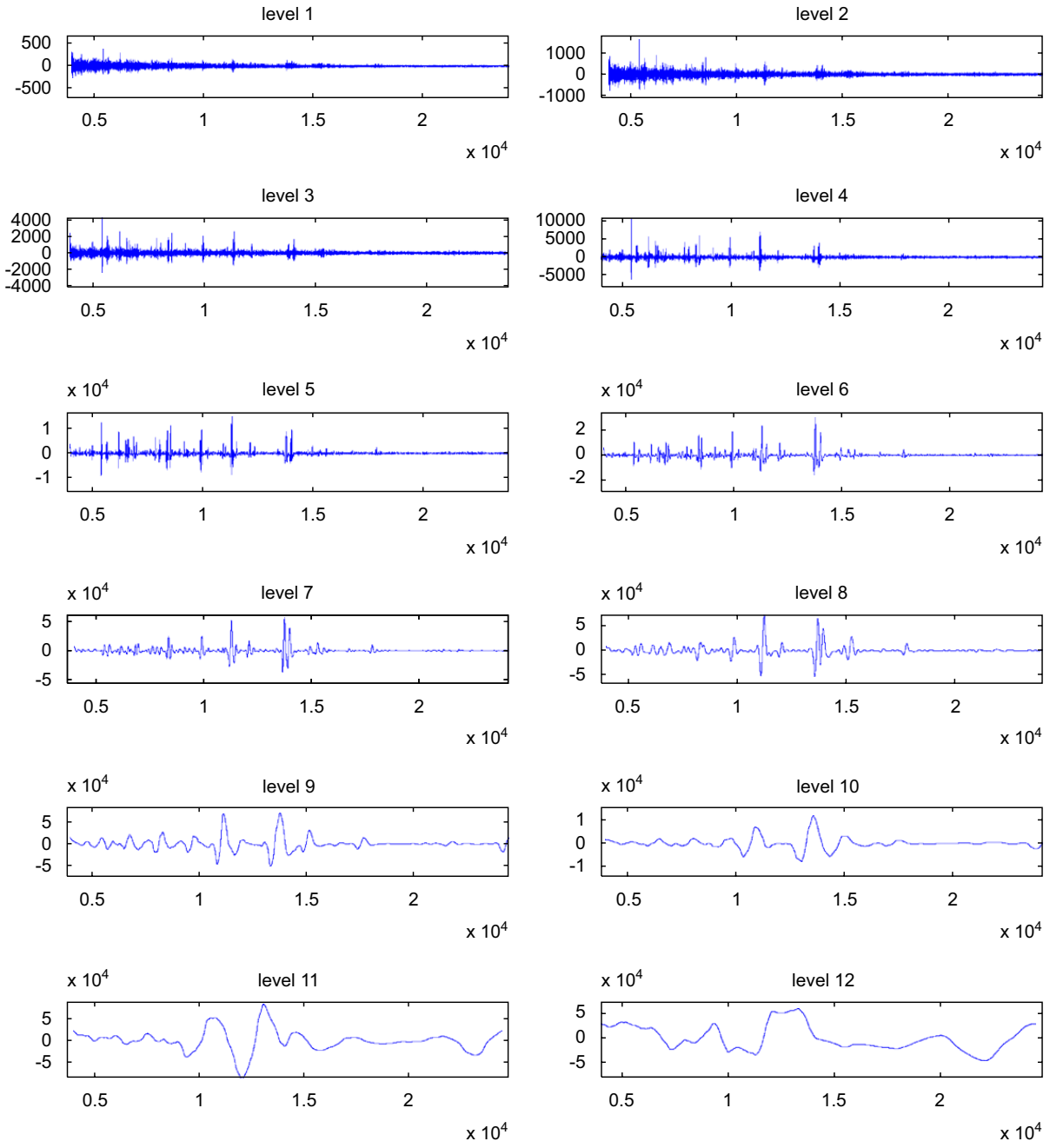


Fig. 1. Wavelets coefficients by levels.

while for the soft threshold, the thresholded coefficient $x$ is calculated by

$$x = \begin{cases} \text{sign}(x)(|x| - t), & |x| > t, \\ 0, & |x| < t. \end{cases}$$

Generally, with SDWT, hard thresholds have a better $\ell_2$ performance while soft threshold have better smoothness. However, with SDWT, since the coefficients are undecimated, hard thresholds will have both better $\ell_2$ performance and smoothness (Coombes et al., 2005; Lang et al., 1995).

Based on the knowledge of the wavelet analysis to the data set, we try to use the objective criteria to determine the threshold values. Basically, the choice of mother wavelets appears not to matter much while the values of thresholds

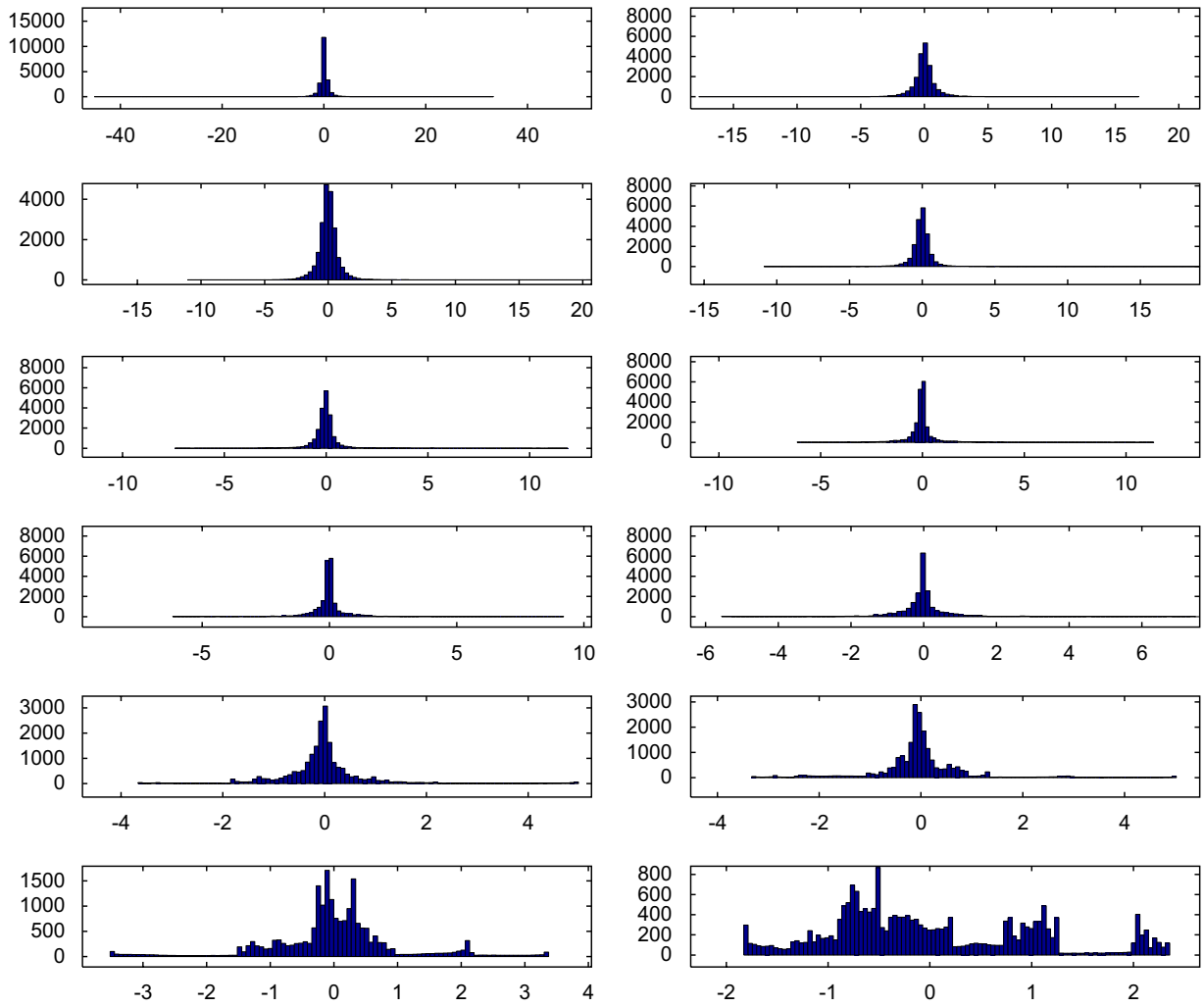


Fig. 2. Wavelet coefficients histograms by levels.

does (Coombes et al., 2005). Then, setting the values of thresholds becomes a crucial topic. According to the analysis above, we would like to set the threshold values based on the properties of data sets.

From Fig. 4, we can see that the high frequency components of spectrum data reduce in energy level as the $m/z$ value increases because the values of median absolute deviation (MAD) change quite distinctly through different $m/z$ segments. MAD/0.67 is a robust estimate of the non-normal variability. This phenomenon is caused by the fact that the machine has relatively low resolution for ions of small $m/z$ values at low $m/z$ interval. Therefore, we propose to set different thresholds at different mass segments by the changing trend of the coefficients at each level as described in Lavielle (1999). MATLAB has a built-in function, called `WVARCHG.m` for this purpose. Applying segmentally denoising strategy, the denoised signal can have reduced variance in the beginning section and retain the useful information in the posterior section.

*Reconstruction*: Reconstruct the denoised signal using the original approximation coefficients of level $N$ and the modified detail coefficients of levels from 1 to $N$ by the inverse SDWT.

A raw MALDI-TOF MS data and corresponding denoised data after baseline correction and normalization is shown in Fig. 5. In the real cancer data analysis, we find out that most wavelet coefficients of MALDI-TOF MS data at high-frequency levels, say from 1 to 4, can be almost ignored. However, we need to be very cautious when manipulating the low-frequency components for keeping as many true peaks as possible after thresholding. According to the exploratory data analysis in the beginning of this section, we select the threshold value large enough, say 40 times of MAD/0.67, to ignore most of wavelet coefficients at level 1–4, which represent the noise signals, specially in the region of large $m/z$ values. The real denoising performance shows appreciation of wavelet application and exploratory data analysis.

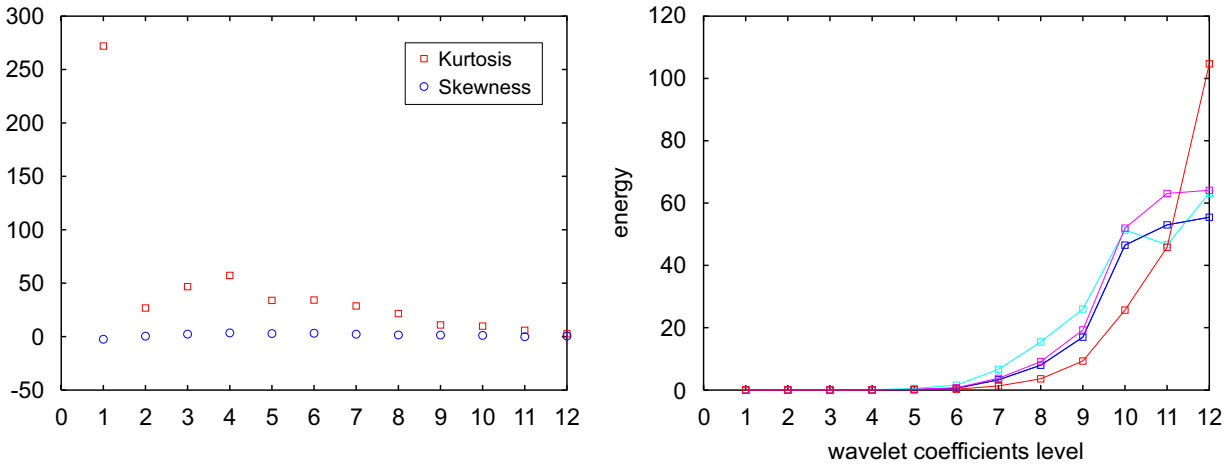


Fig. 3. (L) Wavelets coefficients' skewness and kurtosis by levels, and (R) wavelets coefficients energy through levels (4 spectra).

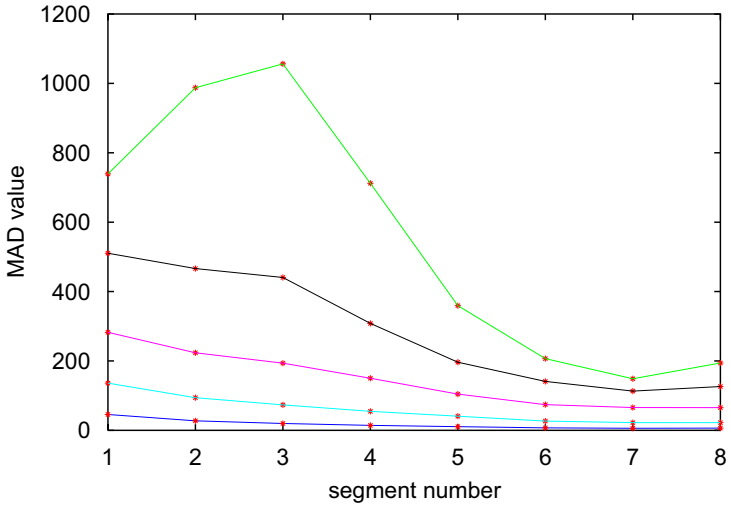


Fig. 4. Wavelets coefficients' MAD at different $m/z$ intervals of level 1–5 (from bottom to top).

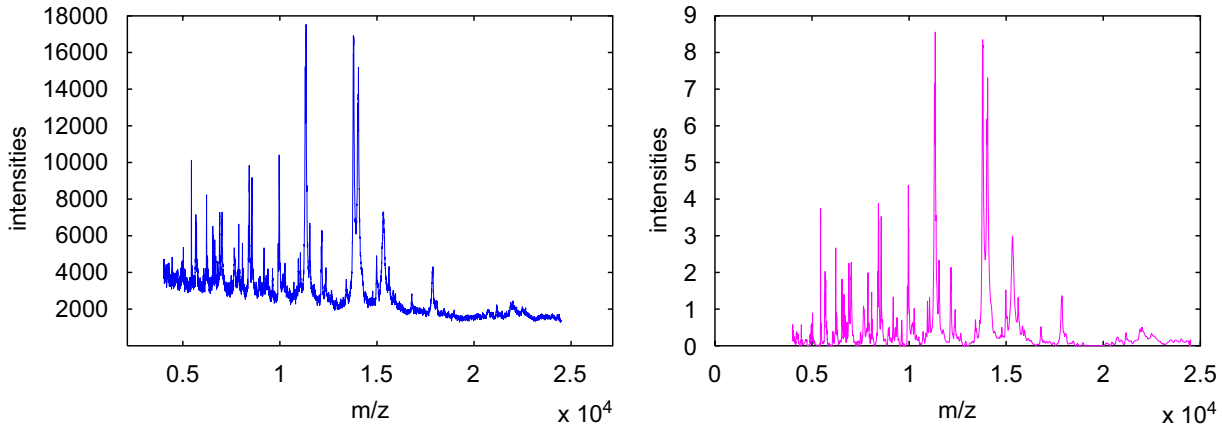


Fig. 5. (L) A raw MALDI-TOF MS data, and (R) MS data after denoising, baseline correction, and normalization.

### 2.3. Analysis on wavelet domain

To view this proteomic data analysis from another angle, rather than extracting the peaks from the spectrum after wavelets denoising, we would extract useful information on the wavelet coefficients domain. The shrinkage of wavelets coefficients in some sense, is a process of principle component extraction as well as denoising (Donoho and Johnstone, 1995). After wavelet shrinkage, the number of wavelet coefficients is only about one tenth of the length of raw spectrum and around 4 times of the number of peaks.

Feature extraction on the wavelet domain will bring us some advantages. It could overcome the difficulty in the measurement of peaks, since the use of either the height of peaks or the area under the peak curves will introduce uncertainty and variation for the peak quantification. Also, it would help to discover potential biomarkers which are usually not prominent peaks or overlapped with adjacent big peaks. Therefore, feature extraction on wavelet domain provides a promising method for the complex functional data pattern recognition.

At the first glance, it might not be straightforward to interpret the discovered biomarkers. But, in fact, we already have the time domain and frequency domain information of differentially expressed wavelet coefficients. Applying wavelet inverse transform, it is simple to locate the biomarkers on $m/z$ domain. In the real application, it is more helpful for the biologist to identify the biomarker regions rather than focusing only on the fixed peak locations. Based on the real data analysis in Section 4.2, we can see that the classification rate is very high, even when we set a very strict hard threshold shrinkage. In this way, the discrete wavelet coefficients can be seen as a kind of "principle components" for the differential analysis, which fulfills the dimension reduction function.

## 3. Method: processing procedures

In this section, we describe in detail a software package, WaveSpect0, for processing of proteomic MS data, which includes: step-interval unification, denoising, baseline correction, normalization, peak detection, and cross samples peak alignment.

### 3.1. Step-interval unification and denoising

First, we need to unify the input discrete data by assigning a common mass to each spectrum so that the MATLAB wavelet function can read in and represent the spectra correctly and efficiently. Here, cubic spline interpolation is applied in resampling so that the discrete data can be input with a constant sampling step-size. The advantage of such a step-interval unification is that we set up a standard discrete data read-in for MATLAB wavelet function with very little variation from the true spectrum and thus, the spectrum signal in the frequency domain can be analyzed during the signal-processing procedure with a quick correspondence to its $m/z$ values. After spline resampling, all spectrum vectors have the same $m/z$ value identifications with a constant discretized step-size. This step is necessary for the coefficient analysis on the wavelet domain, because we want to compare the coefficient on each scale from different samples sharing the same time scale. Next, we apply adaptive stationary discrete wavelet transform for denoising as we introduced in Section 2.

### 3.2. Baseline correction and normalization

The denoised data are still apart from the true proteins' distribution because there is still artificial bias, called baseline. Therefore, a baseline curve should be subtracted from the denoised spectrum. For this purpose, we first search the local minima of the denoised signal and fit them with a spline curve by using the all $m/z$ values. Then, the denoised signal minus the spline baseline curve becomes the approximation of the true signal. In our software package, we set up several choices for the baseline splines according to the desire of the smoothness of the baseline curves. More sophisticated baseline curves should be determined by statistical estimates on the spline coefficients.

To compare spectra in the same scale, the normalization step is inevitable. Since the spectrum after baseline correction is closer to the true distribution of the signal, we can normalize every element in the spectrum vector. In our package, we apply an $\ell_2$ averaging formula for the normalization, which is in the energy metric.

### 3.3. Peak detection and alignment

The final goal of the MALDI-TOF MS data processing is to identify the locations and the intensities of peaks. The spectra after all previous processing steps can be put in a matrix of column spectrum vectors of intensities. Usually, the local maxima in each column are the peaks of each spectrum. To filter out more small peaks, an ad hoc method based on the ratio of signal and noise ($S/N$) is proposed by Coombes et al. (2005).

In real application, one peak will be identified within a certain separation range ($SR$). An experimental formula for $SR$ is given by $SR = 2 + (X_i/1000)$ in Daltons, where $X_i$ is the $m/z$ location. However, in the peak matrix, the positions of peaks in each column around the same $m/z$ value may be slightly different from each other (apart from 2 to 3 rows in the matrix). Therefore, we must bin these peaks to correspond to the same $m/z$ value. This is called cross samples peak alignment. A so-called average spectrum is determined for the binning purpose by Morris et al. (2005). An new algorithm called project spectrum binning (PSB) is introduced by Hong et al. (2007). Here, we propose a "central" spectrum idea by using local clustering techniques so that the peaks of the central spectrum can be used to generate a binning. The central spectrum binning (CSB) algorithm can be described as follows.

*Step* 1 (Finding center): The center vector is defined as the spectrum of the minimum sum of pairwise correlation distances to all the other vectors. In order to align peaks more accurately, we would like to calculate the distances segmentally along the $m/z$ domain and thus, find the centers for different $m/z$ intervals. We can apply the same segment (interval) distribution in the denoising step for this segmental distance calculation.

*Step* 2 (Aligning peaks): We align the matrix by moving other vectors' peaks near the center vector's peaks (in local $SR$ binning window) to the center peaks' positions, which means that we align the peaks in other spectra according to the center spectrum's peak locations. This most likely needs more than one step to finish. After the first alignment, if there are a few peaks in other spectra appearing "too far" away from the peak locations of the center spectrum, then we may need the second time alignment using the center of the submatrix by deleting the first center spectrum in the original matrix. The algorithm can be iterated till a certain level of alignment, say 95% of peaks in all spectra have been assigned to $m/z$ locations with some central spectra. In this way, we make all "nearby" peaks to the same $m/z$ positions. There may still be a few nonzero rows with the majority of zero elements after the alignment. If necessary, we can omit those nonzero rows having, say, 1% or less nonzero elements in the row. Now we have an output matrix with $n$ spectra and $p$ features.

In the whole processing procedure, we have a consistent metric because the wavelet transform follows energy conservation. The success of wavelet denoising makes the peak selection and alignment accurate and efficient. An alignment method with a fixed bin width is used by Coombes et al. (2005). In comparison, the CSB algorithm is more efficient and accurate for cross sample peak alignment.

## 4. Results

In this section, we use a real data set from 62 healthy mice and 77 mice with tumors collected at Vanderbilt Ingram Cancer Center for evaluation of this wavelet-based MALDI-TOF MS data processing package.

### 4.1. Peak selection based method results

The input is a $50\,000 \times 278$ matrix. We focus on the $m/z$ interval from 3001 to 23 480 Da, which contains the information of interest. As usual, the denoising procedure always has an intrinsic tradeoff between the signal's sensitivity and specificity. Therefore, it is very important to set up good denoising parameters. In this example, we use Daubechies wavelet of degree 2 for decomposition and examine the SDWT coefficients at every level. The exploratory data analysis is necessary. After decomposing the signal into level 12, we have 13 coefficient vectors, including the approximates. It can be seen that the energy of the high frequency is very low and MAD values vary more at different $m/z$ intervals. Thus, we only do thresholding to the 6th level. We set 5 $m/z$ intervals to estimate the thresholds. Through the levels, the threshold values depend on the energy and kurtosis of the coefficients.

Actually, the so-called signal self-adaptive thresholding strategy takes the wavelet analysis by analyzing the signal in both time and frequency domain. The parameters vary in different decomposition levels and as well as in $m/z$ segments based on the signal's own properties. Peak selection results will show up even more impressively in the balances of the tradeoff between the sensitivity and the specificity. To make all the spectra comparable, we apply the baseline correction and normalization to each spectrum as shown in (Fig. 5(R)).

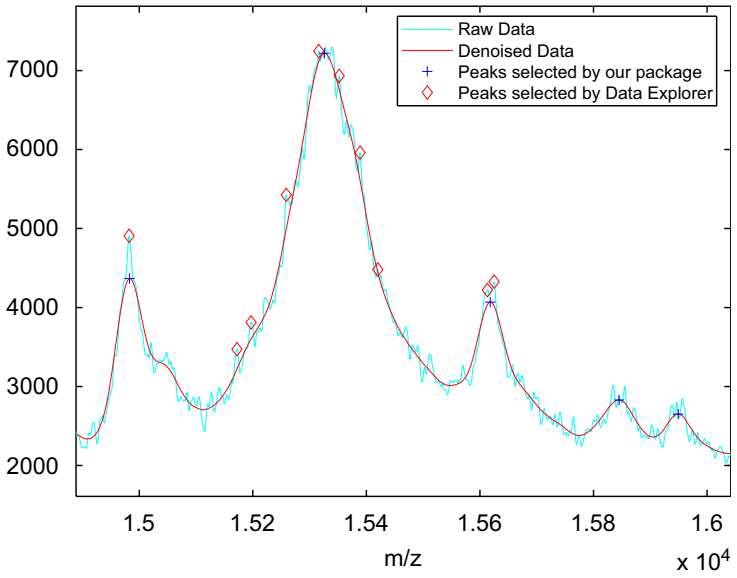


Fig. 6. Peaks selected by WaveSpect0 software (+) and Data Explorer software (◇).

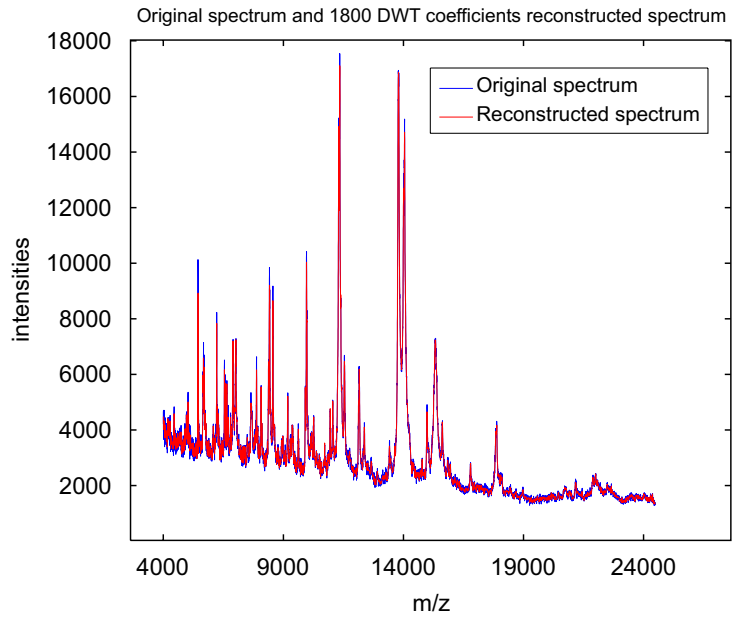


Fig. 7. The original and reconstructed spectrum.

Fig. 6 provides some comparison results in peak selection using data explorer software, which VICC normally used before we introduce WaveSpect0. Clearly, WaveSpect0 picks up more "true" peaks and eliminates many "false" peaks. In the peak alignment we use the *SR* formula mentioned in the previous section. To evaluate our feature extraction effects, an unsupervised clustering analysis is applied with correlation distance and Ward's linkage.

### *4.2. Wavelet coefficients analysis results*

As we mentioned in Section 2, the wavelet shrinkage is efficient for dimension reduction. We can apply the well-developed classification and biomarker detection methods (Hastie et al., 2001) to this wavelet coefficient data. In the data set studied in Section 4.1, we only use 1800 wavelet coefficients, which is less than one tenth of the original length of the spectrum. Fig. 7 shows the original and reconstructed spectrum using selected wavelet coefficients, which we

can barely tell the difference. To validate this method, we also run support vector machine (SVM) with linear kernel on the wavelet coefficients. The classification result is positive: the accuracy of the testing set is 99.3056, according to 10-fold cross-validation, which means that by a small number of the wavelet coefficients, we can describe the original functional data's pattern exceedingly well.

Finally, we would like to mention that a demo version of the software package with major source codes can be downloaded from the website: http://www.vicc.org/biostatistics/software.php.

## Acknowledgments

The authors are grateful to professor Christophe Croux and the anonymous referees for their valuable comments and suggestions which helped to improve this work. Also, the authors would like to thank Jonathan Xu, Department of Cancer Biology, Vanderbilt University for providing data sets and many useful suggestions in this study. This research was supported in part by Lung Cancer SPORE (Special Program of Research Excellence) (P50 CA90949), Breast Cancer SPORE (1P50 CA98131-01), GI (5P50 CA95103-02), and Cancer Center Support Grant (CCSG) (P30 CA68485) for Y. Shyr, and by NSF IGMS (#0408086 and #0552377), NSA (H98230-05-1-0304), and MTSU REP for D. Hong.

## References

Aldoubi, A., Unser, M., 1996. Wavelets in Medicine and Biology. CRC Press, Boca Raton, FL.

Coombes, K.R., Kooman, J.M., Baggerly, K.A., Morris, J.S., Kobayashi, R., 2005. Improved peak detection and quantification of mass spectrometry data acquired from SELDI by denoising spectra with the undecimated discrete wavelet transform. Proteomics 5 (16), 4107–4117.

Donoho, D.L., Johnstone, I.M., 1994. Ideal spatial adaption via wavelet shrinkage. Biometrika 81, 425–455.

Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. J. Amer. Statist. Assoc. 90, 1200–1224.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Hong, D., Shyr, Y., 2006. Wavelet applications in cancer study. J. Concrete Appl. Math. 4, 505–521.

Hong, D., Shyr, Y., 2007. Mathematical framework and wavelets applications in proteomics for cancer study. In: Tan, W.-Y., Hannin, L. (Eds.), Handbook of Cancer Models with Applications to Cancer Screening, Cancer Treatment and Risk Assessment. World Scientific, Singapore, to appear.

Hong, D., Li, H., Li, M., Shyr, Y., 2007. Wavelets and projecting spectrum binning for proteomic data analysis. In: Hong, D., Shyr, Y. (Eds.), Medical Data Analysis Using Mathematical Tool and Statistical Techniques. World Scientific, Singapore, pp. 155–174.

Lang, M., Guo, H., Odegard, J.E., Burrus, C.S., Wells Jr., R.O., 1995. Noise reduction using an undecimated discrete wavelets transform. IEEE Signal Process. Lett. 3, 10–12.

Lavielle, M., 1999. Detection of multiple changes in a sequence of dependent variables. Stoch. Procress. Appl. 83 (2), 79–102.

Lió, P., 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19, 2–9.

Mallat, S., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Pattern Anal. Mach. Intelligence 11, 674–693.

Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A., Kobayashi, R., 2005. Feature extraction methodology for mass spectrometry data in biomedical applications using the mean spectrum. Bioinformatics 21, 1764–1775.

Nason, G.P., Silverman, B.W., 1995. The stationary wavelet transforms and statistical applications. In: Lecture Notes in Statistics: Wavelets and Statistics. Springer, New York, pp. 281–299.

Pedrioli, P.G.A., Eng, J.K., Hubley, R., et al., 2004. A common open representation of mass spectrometry data and its application to proteomics research. Nat. Biotechnology 22 (11), 1459–1466.

Rahman, S.M., Shyr, Y., Yildiz, P.B., Gonzalez, A.L., Li, H., Zhang, X., Chaurand, P., Yanagisawa, K., Slovis, B.S., Miller, R.F., Ninan, M., Miller, Y.E., Franklin, W.A., Caprioli, R.M., Carbone, D.P., Massion, P.P., 2005. Proteomic patterns of preinvasive bronchial lesions. Amer. J. Respir. Crit. Care. Med. 12, 1556–1562.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., Le, Q., 2004. Sample classification from protein mass spectrometry, by peak probability contrasts. Bioinformatics 20 (17), 3034–3044.

Wagner, M., Naik, D., Pothen, A., 2003. Protocols for disease classification from mass spectrometry data. Proteomics 3, 1692–1698.

Xie, L., Xu, B.J., Gorska, A.E., Shyr, Y., Schwartz, S.A., Cheng, N., Levy, S., Bierie, B., Caprioli, R.M., Moses, H.L., 2005. Genomic and protenomic analysis of mammary tumors arising in transgenic mice. J. Proteome Res. 4 (6), 2088–2098.

Xu, B.J., Shyr, Y., Liang, X., Ma, L.J., Donner, E.M., Roberts, J.D., Zhang, X., Kon, V., Brown, N.J., Caprioli, R.M., Fogo, A.B., 2005. Proteomic patterns and prediction of glomerulosclerosis and its mechanisms. J. Amer. Soc. Nephrol. 16 (10), 2967–2975.

Yasui, Y., McLerran, D., Adam, B., Winget, M., Thornquist, M., Feng, Z., 2003. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. J. Biomed. Biotechnol. 4, 242–248.

Yildiz, P., Shyr, Y., Rahman, S.M.J., et al., 2006. Approaching the diagnosis of lung cancer with serum proteomic profiling, submitted for publication.

Yu, W., Wu, B., Lin, N., Stone, K., Williams, K., Zhao, H., 2006. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. Comput. Biology Chem. 30, 27–38.