12-2004

MALDI-TOF MS Data Processing Using Wavelets, Splines and Clustering Techniques.

Shuo Chen
East Tennessee State University

Recommended Citation

Chen, Shuo, "MALDI-TOF MS Data Processing Using Wavelets, Splines and Clustering Techniques." (2004). *Electronic Theses and Dissertations.* Paper 966. https://dc.etsu.edu/etd/966

MALDI-TOF MS Data Processing Using Splines, Wavelets and Clustering Techniques

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

by

Shuo Chen

December 2004

Don Hong, Ph.D., Chair

Robert Gardner, Ph.D.

Anant Godbole, Ph.D.

Tiejian Wu, Ph.D.

Keywords: mass spectrum, MALDI-TOF, splines, peak processing, binning, undecimated wavelet transform, adaptive wavelet denoising, clustering.

ABSTRACT

MALDI-TOF MS Data Processing Using Splines, Wavelets and Clustering Techniques

by

Shuo Chen

Mass Spectrometry, especially matrix assisted laser desorption/ionization (MALDI) time of flight (TOF), is emerging as a leading technique in the proteomics revolution. It can be used to find disease-related protein patterns in mixtures of proteins derived from easily obtained samples. In this paper, a novel algorithm for MALDI-TOF MS data processing is developed. The software design includes the application of splines for data smoothing and baseline correction, wavelets for adaptive denoising, multivariable statistics techniques such as clustering analysis, and signal processing techniques to evaluate the complicated biological signals. A MatLab implementation shows the processing steps consecutively including step-interval unification, adaptive wavelet denoising, baseline correction, normalization, and peak detection and alignment for biomarker discovery.

DEDICATION

I dedicate this thesis to my parents.

ACKNOWLEDGMENTS

I would like to thank Dr. Hong for all of his precious advice and support during the preparation of this thesis. I would like to thank Jonathan Xu at Mass Spectrometry Research Center, Vanderbilt University for providing the data set 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) by NSF IGMS #0408086.

I would also like to thank the rest of my committee for their time and advice. Many thanks to Dr. Price for his help in Statistics, Matlab and SAS.

I would acknowledge our graduate coordinator Dr. Gardner; he has been ready to help me since my entrance to ETSU. I would like to also include a word of thanks to all the people in the Department of Mathematics who helped and taught me in these past two years.

Contents

ABSTRACT i

COPYRIGHT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

1 Introduction to MALDI-TOF Data 1

1.1 Mass Spectrum 1

1.2 TOF Mass Spectrum 2

1.3 Properties of MALDI TOF MS Data 3

2 Wavelets and Applications in MALDI TOF MS Data Analysis 6

2.1 Wavelet Analysis 6

2.2 DWT Denoising 8

2.2.1 Discrete Wavelet Transform 8

2.2.2 Wavelet Denoising 9

2.3 Advanced Wavelet Denoising Methods 10

2.3.1 UDWT Denoising 10

2.3.2 Adaptive Denoising 11

3 Clustering Analysis 13

3.1 Basic Concepts 13

3.1.1 Hierarchical Clustering 14

3.1.2 K-means Clustering 15

3.1.3 Distinct Elements 15

3.2 A Practical Example-50 Patients 16

4 Matlab mass spectrum data processing package 17

4.1 Restep 17

4.2 Denoising 19

4.3 Baseline Correction and normalization 21

4.4 Peak processing 22

5 Summary 24

BIBLIOGRAPHY 25

APPENDICES 27

A. Figures 28

B. Table 33

VITA 34

1 Introduction to MALDI-TOF Data

Matrix-assisted laser desorption/ionization, time of flight (MALDI-TOF) mass spectrometry (MS) is a technology on the cutting edge and plays an important role in the proteomics revolution. MALDI-TOF MS can directly measure the complex mixture of proteins obtained from biological fluids such as serum, urine, or nipple aspirate fluids. By comparing healthy and ill tissue, the disease-related proteomic patterns can be found[6].

1.1 Mass Spectrum

The mass spectrum technique is an attractive analytical tool used in the study of molecular biology. It converts neutral molecules into gaseous ions and separates those ions according to their mass weights. The spectrum has mass as a predictor and charge value/molecules's intensity as response.

In recent years, ionization techniques have been developed so that not only volatile samples but also liquid and solid samples can be ionized to intact molecules. The major advantage of this progress is that we can analyze compounds of higher molecular weight such as peptides, proteins and oligonucleotides. Thus, for a tiny sample of tissue, we can find the mass spectrum and determine how many proteins there are at some molecular weights. That helps us understand the protein composition of the tissues.

1.2 TOF Mass Spectrum

With increasing development of optics, laser, electronics(sensor) and signal processing technology, it is only now possible for time of flight(TOF) mass spectrometers to process the high mass range and the high sensitivity multichannel recording capabilities that were anticipated so many years ago, although the low mass resolution is still not so precise [1]. The basic principle of time of flight is that the ions are extracted from a electron impact source by a constant electrical field to final energies of approximately 500 eV, flying through the entire region to the same final kinetic energy:

$$\frac{mv^2}{2}=eV$$

resulting in velocities given by:

$$v=[\frac{2eV}{m}]^{1/2}$$

and flight times:

$$t=[\frac{m}{2eV}]^{1/2}D.$$

Thus, we can see the flight time depends on the square root of mass[1]. However, the ions which have the same molecular weights do not arrive at their destination at the same time. Actually, from the mass spectrum, we can see the time of flight of a group of ions at the same molecular weights distribute in a bell shaped pattern(normal distribution). The peak of the bell, as well as the mean, is an unbiased estimator of the flight time/mass of this kind of ion, and the ion detector refreshes (charges become zero) at a certain frequency. Thus, the spectrum we get is in the discrete data form; the x-axis is the time domain with time interval corresponding to the clock of the machine while the y-axis is the quantity of charge at each fixed point of time. By the time/mass ratio, we can figure out what the compounds contain (which molecules at which proportions).

For the spectrum data, mass resolution is defined as $\frac{m}{\triangle m}$. For time of flight mass spectrum, the resolution satisifies:

$$\frac{m}{\triangle m}=\frac{t}{2\triangle t}.$$

Thus, mass resolution depends on time resolution and (therefore) upon laser pulse initial kinetic energies (and velocities) [1].

1.3 Properties of MALDI TOF MS Data

In the real procedure, there are some characteristic properties of the MALDI TOF MS data (Figure 1.1):

1. The discrete data have basic shapes of several bumps, and the bumps are steep at small mass while moderate at relatively large mass.

2. The mass range of the data of interest is generally from 4000 Daltons to 50000 Daltons.

3. The intensities of proteins of smaller molecular weights are obviously greater, because of the matrix molecules that form an integral part of the technology and the breaking up of large proteins during the ionization process.

One of the main problems in MS data processing is to deal with the variation. Though it is believed that there is substantial variability in the technology, and that changes in data acquisition protocols and parameters can affect MS profiles drastically, the variability arises from the application of inadequate algorithms for the MS data processing. In this paper, we propose a novel algorithm to process MALDI-TOF data using mathematical and statistical tools and signal processing techniques. The algorithm involves a number of complicated steps. The basic goal is to identify the locations of peaks and to quantify the window sizes for accuracy. For the MS spectra sample set, one notices that the spectra are different in length of data, wave shape, amplitude, and peak position. In medical study, we could have different samples from a single tissue. Therefore, the initial key step in MS data processing is to assign common mass to each spectrum. This is also called mass alignment in literature.

One of the goals in cancer study is to find disease-related proteins in a spectrum of a certain tissue. Therefore, the signal processing of the MS data needs to be efficient and effective, and then the biomarker discovery should be based on advanced statistical analysis. In this paper, several mathematical tools are applied in MS data processing, such as splines for data smoothing and baseline correction, wavelets for adaptive denoising, multivariable statistical techniques (such as clustering analysis), and signal processing techniques are combined to evaluate the complicated biological signals. An algorithm package for MALDI-TOF MS data processing is developed, which processes the raw MALDI TOF MS data consecutively until a general form of protein distribution for a certain tissue is expressed. The algorithm package is implemented in MatLab, a commercial “Matrix Laboratory” package which operates as an interactive programming environment. Some program skills innovated in this package include: unifying the time interval of discrete data by spline functions, adaptive stationary discrete wavelet denoising, and the center alignment binning algorithm based on clustering. The software package was tested using a data set collected at Vanderbilt-Ingram Cancer Center.

The remainder of the paper is organized as follows. In the next section, we discuss wavelet applications for MALDI-TOF MS data analysis. In Section 3, clustering analysis, as a multivariable statistics technique, is applied in an initial study of biomarkers discovery in cancer study from MALDI-TOF spectra. The complete software package for MALDI-TOF MS data processing using splines, wavelets, clustering and signal processing techniques is developed in the final section. The appendix contains figures and tables.

2 Wavelets and Applications in MALDI TOF MS Data Analysis

In recent years, wavelet methodology has been developed into a powerful tool to resolve various practical problems in signal processing, image compression, numerical analysis and statistics ([8],[9],[10],[11]). Wavelets are also an important support in biomedical signal processing ([3],[4]). In this chapter, we will discuss wavelets' application in the analysis of MALDI TOF MS data. Since the data are one dimensional and discrete, we will use 1-D discrete wavelet transforms.

2.1 Wavelet Analysis

As a new tool for signal analysis, wavelets can be used for the signal of long time intervals where we want more precise low-frequency information and for shorter regions, where we want high-frequency information, by the variable-sized windows technique. By using wavelet analysis, we can not only obtain information on the frequency domain but also information on the time domain.

Comparing wavelets with sine waves, the basis of the classic Fourier analysis, we can see that sinusoids have a support of the entire time domain through the x-axis from minus to plus infinity, and are periodical and symmetric while wavelets have compact support and tend to be irregular and asymmetric. Furthermore, Fourier analysis consists of breaking up a signal into sine waves of various frequencies which convert a signal from the time domain to the frequency domain by the formula:

$$F(w) = \int f(t) e^{-jwt}\ dt$$

in the form of scalar production; a signal $f(t)$ is convoluted by the sinusoids at frequency of w. Similarly, wavelet analysis is the breaking up of a signal into shifted and scaled versions of the mother wavelet. It is important to understand that, for orthognal wavelets, any two different wavelets are orthogonal because the wavelet of each scale level has a mean value of zero, and any two wavelet lengths of two scale levels have a multiple relationship. Thus, any square integrable function can be expressed in the wavelet decomposition form. For a certain scale, the original signal is broken up into a sequence of coefficients, and each coefficient is the result of convolution with the wavelet on that time interval position. The following formula is a wavelet transform:

$$\textit{C(scale, position)} = \int \textit{f(t)}\Psi\textit{(scale, position, t)dt}$$

where Ψ(scale position) is a wavelet generated from a mother wavelet Ψ. Generally,

$$\Psi_{a,b} = a^{-1/2}\Psi(\tfrac{x-b}{a}), \text{ for } a \in R^{+} \textit{ and } b \in R$$

Thus, at different levels of scales, there is a sequence of coefficients. If scale doubles, the number of coefficients will decrease to half of the previous scale level. Clearly, wavelets have one advantage over the Fourier analysis-we can find the signal characteristics around some fixed time positions. Because of this nice local analysis performance of wavelets, aspects are revealed such as trends, breakdown points, discontinuities in higher derivatives, and self-similarity that are lost in other signal analysis techniques.

Furthermore, because it affords a new form of data with both frequency and time information, wavelet analysis can often compress or de-noise a signal without appreciable degradation.

2.2 DWT Denoising

2.2.1 Discrete Wavelet Transform

Calculating wavelet coefficients at every possible scale is a fair amount of work, and it generates numerous data as a result. If we choose scales and positions based on powers of two, then our analysis will be much more efficient and just as accurate. We obtain such an analysis from the discrete wavelet transform (DWT). Actually, in 1988 Mallat produced a fast wavelet decomposition and reconstruction algorithm [5]. The Mallat algorithm for discrete wavelet transform (DWT) is, in fact, a classical scheme in the signal processing community, known as a two-channel subband coder (see page 1 of the book *Wavelets and Filter Banks*, by Strang and Nguyen [12]) using conjugate quadrature filters or quadrature mirror filters (QMF).

The Mallat algorithm mainly includes two steps: decomposition and reconstruction. The decomposition step begins with an original signal s, next calculates cA1 and cD1, the approximation and detail coefficients at level one, by the low-pass and high-pass filters followed by downsample, and then cA2 and cD2, and so on (see Figure 2.1).

The reconstruction step is also called the inverse discrete wavelet transform (IDWT), which starts from the approximation and detail coefficients of cAj and cDj, next inserts the odd-number signal as zeros into the coefficients, and goes through the low/high filters, then calculates the coefficients of cAj-1 by summing the signal from the two filters, and then, using cAj-1 and cDj-1, calculates those of cAj-2, and so on (see Figure 2.2).

According to the chosen wavelet, we can obtain four high/low-frequency filters of decomposition as well as the ones of reconstruction.

2.2.2 Wavelet Denoising

A noisy signal can be expressed as:

$$s(t) = f(t) + \sigma e(t)$$

where t is discrete with constant interval.

First, we assume that $e(n)$ is a Gaussian white noise N(0,1) and the noise level σ is equal to 1. Our goal is to grab the true signal f from the noisy signal s.

The basic denoising procedure can be described in this way:

1. Decomposition: Calculate the coefficients of a signal by DWT. There are two parameters we need to choose: the level of decomposition N and the type of wavelet.

2. Thresholding detail coefficients: For each level from 1 to N, select a threshold and omit the detail coefficients below the thresholds. We can choose hard or soft thresholds, and values of thresholds.

3. Reconstruction: Compute wavelet reconstruction using the original approximation coefficients of level, N, and the modified detail coefficients of levels from 1 to N by IDWT.

Comment: For hard thresholding, the thresholded coefficient x is x if $|x| > t$, and is 0 if $|x| < t$, while for soft threshold, the thresholded coefficient x is sign(x)($|x| - t$) if $|x| > t$ and is 0 if $|x| < t$. The values of the thresholds are determined by the signal itself with some algorithms, such as: Stein's Unbiased Risk Estimate (SURE) and square root of double length log. We can see a denoising example in Section 4.2.

2.3 Advanced Wavelet Denoising Methods

Recently, some advanced wavelet methods have been developed to solve many practical problems efficiently. These problems include infinite white noise processing and shift invariant signal processing.

2.3.1 UDWT Denoising

The classical DWT denoising method has a drawback: the DWT is not shift-variant. That means, even for periodic signals such as sine waves, the coefficients will change if the original signal is shifted. Considering that Fourier transforms will remain the same if there is a time shift, the DWT is dependent on the phase of the signal.

In order to restore the translation invariance, an Undecimated Discrete Wavelet Transform (UDWT) method is proposed. The UDWT algorithm is slightly different from the DWT algorithm. The DWT decomposition step generally includes two filters and two downsamplers. For downsampling, the output signal is dyadic-decimated which only records the even-ordered signal. As a result, the coefficients cAj or cDj only have half of the length of cAj-1. Comparatively, the decomposition step of UDWT does not have the downsample part. Thus, coefficients cAj or cDj are as long as cAj-1. For instance, if we decompose the original signal s of length at j level, then we will get a $j \times n$ coefficient matrix, one row of approximations, and $j-1$ rows of details. Then, by the same method, we can threshold the coefficient matrix and the hard threshold method is better [6]. Moreover, in the inverse undecimated wavelet transform, we also do not need to insert zeros as the odd indexed coefficients.

There is a restriction: we define the UDWT only for signals of length divisible by 2J, where J is the maximum decomposition level, and we use the DWT with periodic extension. There is an algorithm, called padding, that can make the signal extend to the proper length (jth power of 2).

UWDT can bring some good properties such as time-shift invariance, less loss of information (because of the redundancy of the coefficients at each level), and smoothness as well as l_2 space performance [13].

2.3.2 Adaptive Denoising

For the noisy signal model:

$$s(t) = f(t) + \sigma e(t)$$

Considering the situation that the $\sigma e(t)$ is associated with time t, it is unreasonable to set one threshold for all the coefficients. There are several different variance values on several time intervals, which are both unknown parameters. Thus, we need to find the change points or intervals. In this section, we propose an algorithm based on the wavelet toolbox of Matlab, and we focus on UWDT. The method is as follows:

1. Decomposition: We decompose the original signal at level j by UWDT to get the coefficient matrix, then we have $j-1$ rows of details.

2. Replace 2 percent of the biggest values by the mean, because the weight of these values in the detail vectors is great while the quantity is relatively small. In other words, they are outliers.

3. Use Matlab function 'wvarchg' for estimating the change points of the revised $j-1$ rows of details. For example, we have k change points for a row details, then this row can be divided into $k+1$ intervals.

4. Set different threshold values for different intervals, then determine threshold by the hard/soft method.

5. Reconstruction.

The application example of adaptive UDWT denoising can be found in Chapter 4 and the Matlab codes are in the appendix.

3 Clustering Analysis

As a multivariable statistics technique, clustering analysis is widely used in many different fields of study such as engineering, genetics, medicine, psychology, and marketing. Generally, after clustering, we get the result that the profiles of objects in the same cluster are very similar and the profiles of objects in different clusters are relatively quite different.

In the 50 patients case that will be discussed in detail later in this chapter, there are many tissue mass spectra from healthy people and from several groups of sick patients who have different types of cancer. We can see that, even if we do not know the distribution in advance, we can divide the spectra into several groups that are almost the same as the real distribution through the use of clustering analysis.

Generally, we build the model in this way: the initial object can be modeled as a $p{\times}n$ matrix with n vectors of length p. According to the characteristics of the vectors, we can cluster the matrix into several groups in the form of several matrices: $p{\times}n_1$, $p{\times}n_2$, $p{\times}n_3$... for $\sum n_i = n$.

3.1 Basic Concepts

There are mainly two clustering methods: hierarchical clustering and k-means clustering. We will discuss both methods in this section.

3.1.1 Hierarchical Clustering

The hierarchical clustering method shows us a grouping structure of the data in the form of a cluster tree. The tree is not a single set of clusters, rather a multi-level hierarchy where clusters are more similar at the lower level. This allows you to decide what level or scale of clustering is most appropriate for your data.

For a $p \times n$ matrix, there are n vectors (objects), and we group them according to the their relationship (similarity). The distance between two vectors is used to measure the similarity of a pair of vectors. For two vectors x and y, both having length p, there are several types of distance between them such as

Euclidean distance:

$$d_{x,y} = [\sum (x_i \text{ - } y_i)^2]^{1/2}$$

Manhattan distance:

$$d_{x,y} = \sum |x_i \text{ - } y_i|$$

Correlation distance:

$$d_{x,y} = 1 \text{ - } \rho_{x,y}$$

and so on. Different distances may lead to production of different cluster trees.

For n vectors, we will have $n(n-1)/2$ pair distances. Then we need to link these newly formed clusters to other objects to create bigger clusters until all the objects in the original matrix are linked together in a hierarchical tree. There are several ways to create a cluster hierarchy tree such as shortest/longest distance, average distance and centroid distance. Matlab has a function to display the hierarchical tree.

Then, we should determine where to divide the hierarchical tree into clusters. We choose the proper cutoff points so that we can cut the trees into several groups.

3.1.2 K-means Clustering

Compared to the tree structure of hierarchical clustering, the k-means clustering method has set up the number of groups before clustering. All objects are then grouped into k clusters, objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. There are several member objects and a centroid, or center, in one cluster. The center for each cluster is a vector, which has the minimum sum of distances from all objects. K-means clustering uses an iterative algorithm to move objects between clusters until the sum of distances cannot be decreased further. We will apply the center concept in Chapter 4.

3.1.3 Distinct Elements

We have divided the $p\times n$ matrix into k clusters, but not all the elements in one cluster are different from the ones other clusters. Therefore, we should figure out which elements are distinct in one cluster. In other words, the characteristic elements must be determined. For any two clusters, we can do a paired t-test of the objects and find the rows of small p-values; or we can find the weighted average distance at a certain row:

$$w = d_B/(k_1 d_{w1} + k_2 d_{w2} + \epsilon) \ ,$$

where d_B is the distance between cluster centers, d_{wi} is the average (Euclidean) distance among all sample pairs in one cluster, and $k_i = n_i/(n_1 + n_2)$ [7]. Basically, if the objects in a cluster are close to each other then the distances between centers of two clusters is large. At last, if the distance is great enough or p-value is small enough, we can say this element is distinct.

3.2 A Practical Example-50 Patients

There are MALDI TOF MS data from tissue samples of 50 patients, including healthy people (normal) and cancer patients with adeno, squamous, large and other cancers. We did the clustering analysis to the 1628×50 matrix, then compared the results of clustering and the real distribution from the table, to find the distinct elements.

We mainly used hierarchical clustering, because we didn't know the exact number of groups in advance. We did cluster analysis by the Euclidean distance and correlation distance.

The hierarchical trees almost match the real distribution and the correlation method seemed to have a better performance (Figure 3.1, Figure 3.2, the real distribution Table 1). Also, we found the distinct elements of the normal and cancer cluster; rows 38, 350, 356 are significantly different and the p-value of them are 0. Moreover, rows 38 350 356 953 986 991 are relatively distinct. That means these proteins may be disease-related.

4 Matlab mass spectrum data processing package

We collected the MALDI mass spectrum data of tissues from the cancer mice and the healthy mice. In this chapter, we propose a novel method for low level processing of MALDI mass spectrum data, including following steps:

1. Restep: unifying the input discrete data by making the sample intervals constant.

2. Denoising: denoising the signal by undecimated discrete wavelet transform (UDWT) method.

3. Baseline correction and normalization: baseline correction by spline fitting and normalization of all the vectors.

4. Peak processing: trivial peak detection, binning and alignment of the data.

4.1 Restep

As we know, for most applications in discrete signal processing, the time interval between samples is kept constant (for example, sample every millisecond) unless externally clocked. However, the raw MS data's mass steps (the mass differences of two data points which are next to each other) are not uniform at different positions in a spectrum. For the signal itself, no matter whether the x-axis means the flight time or the mass weights, the intervals are not constant for the reason that the ions would not reach the detector in a constant time interval. Therefore, the inconstant property is inconvenient for discrete signal processing. For instance, when we do discrete wavelet decomposition to the vector of discrete intensities in wavelet denoising, we transform the vector from the time/mass domain to the wavelet domain by the assumption that the sample intervals are constant. It is true we could do denoising to the un-restepped data, which would also smooth the data. But, if we use wavelet denoising, the coefficients in the wavelet domain of the un-restepped data would be apart from the ones of the true mass-intensities data, and this could bring bad denoising results.

In our method, we did splines by letting x, the mass vector, and y, the intensities vector, be input first. In detail, we did interpolation splines to fit all the data and got a continuous curve. Then, we sampled the continuous curve at some frequency (every 1 Dalton or 0.5 Dalton). The resultant output is a set of discrete data with the same mass distance between any two consecutive data points. Certainly, different choices of power of the splines will produce different output: the linear and quadratic splines have smaller variation, the cubic or higher order will have more fluctuations.

We compared the three curves obtained: the mass-intensity original graph (the graph of the default input to the Matlab function), the intensities y graph and the splines-resampled data graph (Fig 4.1). Conspicuously, the mass-intensity graph and our splines-resampled data graph matched well with each other. They had almost the same shapes and peaks, while the intensities y graph was quite different from the others. Thus, our method could adjust the original signal into a standard discrete signal keeping the shapes and peaks of the spectra.

The advantages of the restep method are: 1. Standard discrete data is built with very little variance from the true mass-intensities data, 2. During the signal-processing step, the spectrum signal in the frequency domain can be analyzed correctly, 3. All spectrum vectors have the same length. From this graph, we can see that the restep method matched both the shape and the peaks of original signal, while the unrestepped one is shifted.

4.2 Denoising

We processed the normal discrete data by stationary discrete wavelet transform SDWT or undecimated discrete wavelet transform UDWT denoising. The coefficients of the orthogonal DWT are not redundant and efficient computationally, but they are shift-variant [6]. Thus, the denoising performance can change drastically if the starting position of the signal is shifted [6]. The UDWT is shift-invariant, and it yields better visual and qualitative denoising, with a small added cost in computational complexity ([14], [15]). For orthogonal DWT coefficients, the total number of coefficients (the approximates and details) is fixed (almost the same as the original signal) no matter at what level the signal is decomposed. The UDWT coefficients are not decimated, so that for every level decomposition, the approximate and detail coefficients will increase as much as the length of the original signal.

The UDWT also requires the length of the discrete signal to be multiple of 2^k. Since the number of our spectrum data approaches 21000, and this also is close to 5 times the 12th power of 2, (20480), we did all the processing by level 12. Thus, we got two 20480×12 matrices of approximate and detail coefficients.

Now, we could denoise the details of every level by the thresholds we set. As for the thresholds, we basically used soft or hard thresholds. When we used the hard thresholds, we set the detail coefficients below the threshold value to zero; when we used soft thresholds, after setting the detail coefficients below the threshold value to zero, we made the coefficients above threshold value shrink towards zeros. Generally with DWT, hard thresholds have better l_2 performance while soft thresholds have better smoothness. But with UDWT, since the coefficients are undecimated, hard thresholds will have both good l2 performance and smoothness properties ([13]). Additionally, the higher the level we denoise the signal, the smoother are the denoised signals. Considering the loss of the signal and the smoothness, level 12 was good for our spectrum data.

Next, we should set the threshold values. We supposed the noise essentially were white Gaussian noise. Therefore, we used multiples of the median absolute deviation (MAD)/0.67 (in RWT implementation), which gave us a robust estimate of their variability [6]. We could easily notice that the noise of our spectrum data reduced as the mass increased because we detected more ions of small molecular weights which added more noise in the small mass segments. Thus, we should set different thresholds at different mass segments by different MAD. We computed the MAD values of different mass segments, and the thresholds of different segments were different. In this way, the denoised signal would reduce the variance in the beginning part, as well as retain the useful information in the posterior part.

We can see their denoising effects from an example of a restepped discrete signal in Figure 4.2. In order to see the differences of different curves, we shifted the UDWT uniform threshold curve upward 1000 units and adaptive thresholds upward 2000 units. Clearly, the two methods kept the same peaks and shapes when removing the minor fluctuations. But, by any method, we cannot guarantee ability to remove all the noise without any loss of useful information. Our goal is to reduce the number of false peaks.

4.3 Baseline Correction and normalization

The denoised data is still different from the true proteins' distribution because of fragmentation of higher mass proteins. By no means, can we figure out how many proteins broke up. However, we can pull the curve toward the base, which makes the curve closer to the true protein distribution [6]. First, we did splines fitting to the local minimum of the denoised signal. Then, the signal minus the splined baseline equaled the estimated true signal. The estimated signal might had have very few (less than 5) points less than zero by a small value; this is generated by quadratic splines. If we used the step function as the baseline, there would be no points below zero but the baseline would not be estimated as well as the splined one. We'd prefer the splined method, by setting all the points below zero (normally very a limited number) to zero.

We often do normalization just after baseline correction, since the curve after baseline correction is closer to the true distribution of the signal. Moreover, for a certain sample, the intensity values could vary greatly from one spectrum of data to another, although they might have similar peaks and wave shapes. Therefore, we would prefer to normalize every element in one vector so that we could compare elements in different vectors. Usually, we make an element, x_i, in a vector normalized to $\dot{x}_i$ by $\dot{x}_i = \frac{x_i}{\sqrt{\frac{\sum x_i^2}{n}}}$.

4.4 Peak processing

The normalized spectra data in a group could be compared by peaks processing. First, we made all the normalized spectra of one group data as a matrix. Next, we could find the trivial local maximum as the trivial peaks of each spectrum. However, the number of these peaks for each vector seemed relatively numerous (about 1000). Therefore, we filtered the points with too low S/N ratio. The S/N ratio is defined by signal divided by noise.

Noise = $signal_{beforedenoisingandbaselinecorrection}$ - $signal_{afterdenoisingandbeforebaseline}$.

Signal = $signal_{afterbaselinecorrection}$.

We retained the signals with S/N ratio greater than 3.

Another limitation is the separation range; during the separation range (SR) (sr = 2 + $mass/1000$ Daltons), only one peak will be identified. Thus, we kept the greatest one within the sr. To retain all the significant peaks, we kept the peaks of the trivial peaks first. Then, in every SR, we only left the maximum one until the distance between every two peaks were more than or equal to the length of SR. This gave the refined peak matrix.

However, in the refined peaks matrix the positions of peaks of each column around the same mass were different from each other a little (2 or 3 rows' distance). Therefore, we should bin the peaks around some fixed mass to obtain one vector of peaks that would represent the general peaks distribution of the group. We would like to find the general case using the so-called center alignment algorithm as follows.

1. Finding the center. Using the method of multivariable statistics, we could find the center of the set of data that had the minimum sum of pairwise distances to

all the other vectors. The center vector could basically represent the general protein distribution in a mass interval of a matrix. We used correlation distance here, which is one minus the correlation of two vectors. Consider these two facts: 1. The correlation does not mean anything if the vector is too long. 2. In different mass intervals, a different spectrum data vector could represent the general case better. We would like to do the clustering on several mass intervals, and find the centers for different mass intervals. Clearly, since the refined peaks matrix has too many zeros and it is not easy to find the center, we should find which columns are the centers of the normalized matrix first. Thus we could record which vector is the center, at fixed mass intervals. Then the refined peaks of recorded center vectors are used as final centers.

2. Convergence of the close peaks. Now, we aligned the matrix by moving other vectors' peaks near the center vector's peaks to the center peaks' positions. This means we aligned the peaks in other vectors in a matrix according to the center vector. In other words, the center represented the whole matrix in some mass interval. We usually did the alignment twice, because the first time alignment would miss some true peaks if some peaks of the center were far away from the majority. When we did the alignment for the second time, we found the centers of the matrix without the center of the first matrix and then made all the existent peaks near to each other at the same positions.

3. Adjustment. Since we divided the mass axis into several intervals, the peaks could be distributed around the cut mass line. Thus, we combined the peaks of the matrix that were close to each other. Certainly, there are some nonzero rows with respect to the majority zero rows; if necessary, we could omit those nonzero rows that

had less than half of the number of columns with values nonzero. Finally, for each group of data, we had a matrix with aligned peaks. The aligned matrix had peaks at certain mass positions, which means generally speaking, that at these masses there are peaks.

The aligned matrix might have shifted from the true one. We checked the matrix by the known biomarkers, and shifted the matrix back to the known place, according to the biomarkers. We used three known biomarkers with the molecular weights around 5444, 9667 and 14041. As a result, we noticed that the first two biomarkers matched our data very well, but for the known protein of the mass weight 14041, all three groups of data peaks shifted downward about 6 Daltons. Some possible reasons: 1. The time/mass relationship should have some change in great mass. 2. There are some errors in the package. 3. Coincidence.

5 Summary

All the codes of the package are written in MatLab. The test results from eight sets of MALDI-TOF MS data from healthy mice and mice with tumors collected at Vanderbilt Ingram Cancer Center shows that this algorithm is both efficient and effective for this kind of proteomic data processing. Further study should include web-based software design for this package, increasing functions of baseline selection, data denoising using empirical mode decomposition, binning with weighted wavelet coefficients, and data analysis and biomarker identification using more advanced statistical tools.

BIBLIOGRAPHY

[1] Cotter R. J. , Time-of-Flight Mass Spectrometry Instrumentation and Applications in Biological Research, American Chemical Society, Washington, DC 1997.

[2] Hong D. and Shyr Y. *Wavelet Applications in Cancer Study*, Journal of Concrete and Applicable Mathematics, to appear. .

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

[4] Lió P. , *Wavelets in bioinformatics and computational biology: state of art and perspectives*, Bioinformatics 19 (2003), 2–9.

[5] Mallat S. (1989), "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Pattern Anal. andMachine Intell., vol. 11, no. 7, pp. 674-693.

[6] Coombes, K. R. *Improved Quantification of SELDI spectra Using Wavelet*, M.D. Anderson Biostatistics Technical Report UTMDABTR- 001-04.

[7] Goldstein, D., Ghosh, D. and Conlon, E. *Statistical Issues in the Clustering of Gene Expression Data.* Statistica Sinica, 12 (2002): 219 – 240.

[8] Chui C.K. , An Introduction to Wavelets, Academic Press, New York, NY, 1992.

[9] Daubechies I. , Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

[10] Hong D. , Wang, and R. Gardner, Real Analysis with an Introduction to Wavelets and Applications, Academic Press (Elsevier), 2004.

[11] Mallat S. , A Wavelet Tour of Signal Processing, Academic Press, 1999.

[12] Strang, G. , Nguyen T. (1996), Wavelets and filter banks, Wellesley- Cambridge Press, 1996.

[13] Lang M. Guo H, Odegard JE, Burrus CS, Wells RO Jr.(1995) Nonlinear processing of a shift invariant DWT for noise reduction. In:*Mathematical Imaging: Wavelet Applications for Dual Use*, SPIE Proceeding, vol. 2491, Orlando FL.

[14] Lang M. Guo H, Odegard JE, Burrus CS, Wells RO Jr.(1995) Noise Reduction Using an Undecimated Discrete Wavelet Transform for microchip capillary electrophoresis *Electrophoresis;* **24**, 3260-5

[15] Kamath C. Fodor JK. Gyaourova A.(2002 November). Undecimated wavelet transforms for image denoising. Lawrence Livermore National Laboratory technical report UCRL-ID-150931.

APPENDICES

A. Figures

Chapter 1

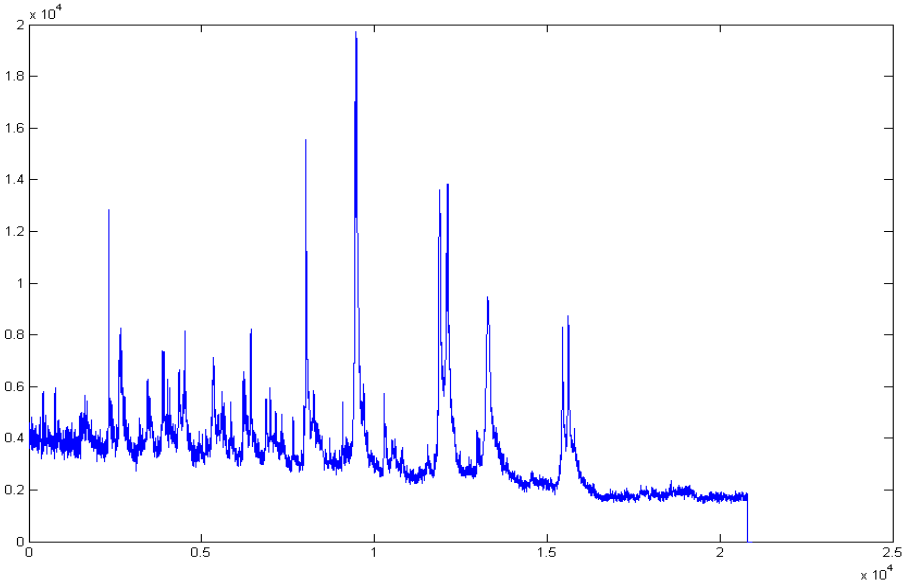

Figure 1.1 An Example of MALDI TOF MS Data

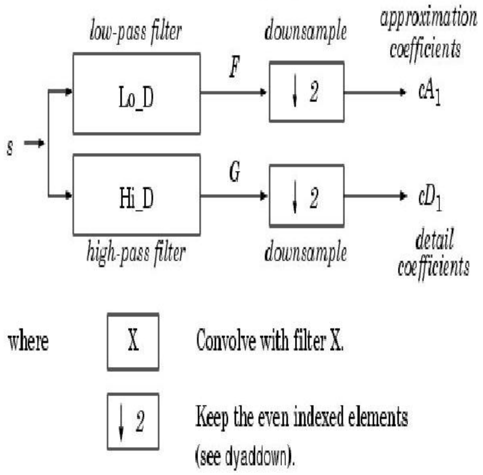

Figure 2.1 DWT Decomposition

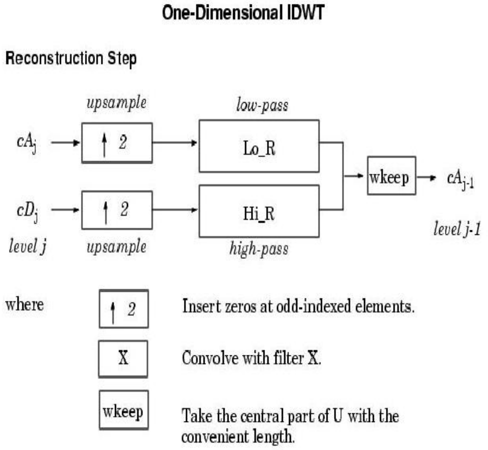

Figure 2.2 DWT Reconstruction

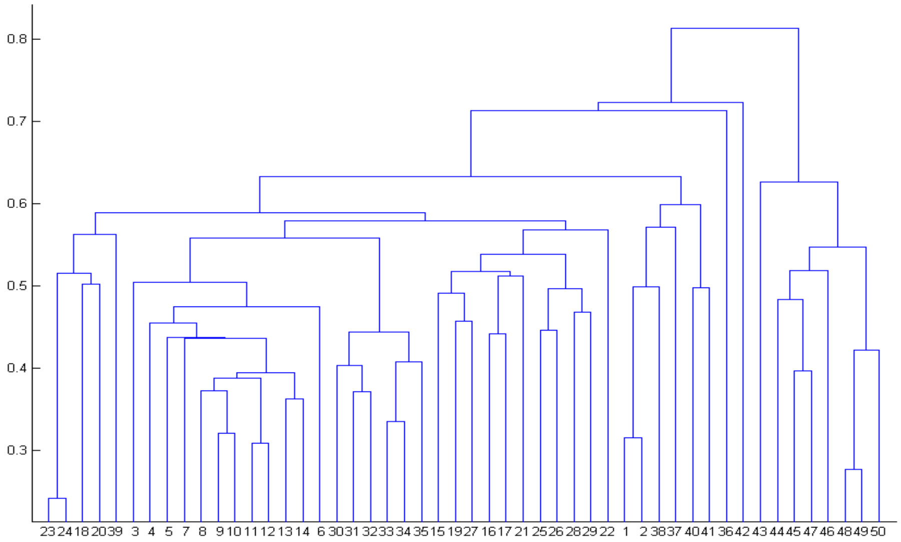

Figure 3.1 Hierarchical Trees by Euclidean Distance

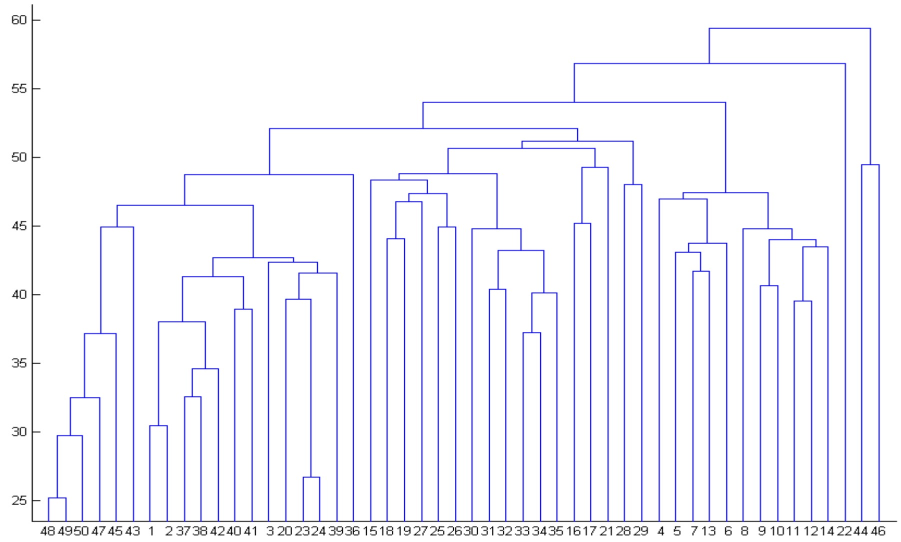

Figure 3.2 Hierarchical Trees by Correlation Distance

Chapter 4

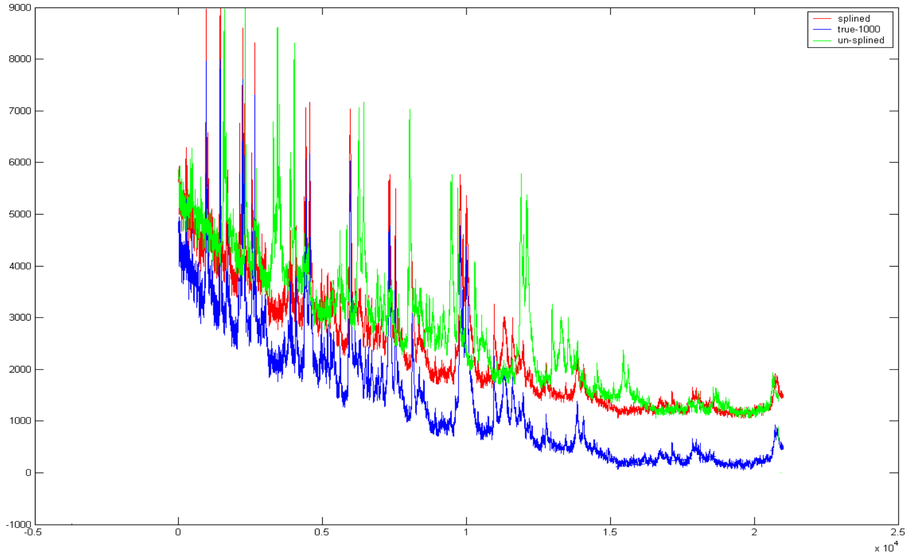

Figure 4.1 Restep

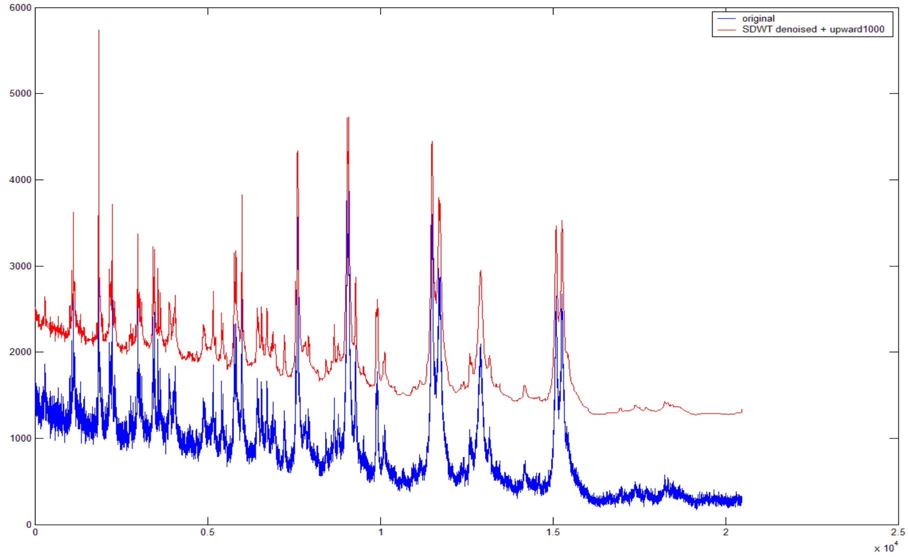

Figure 4.2 UWDT Denoising

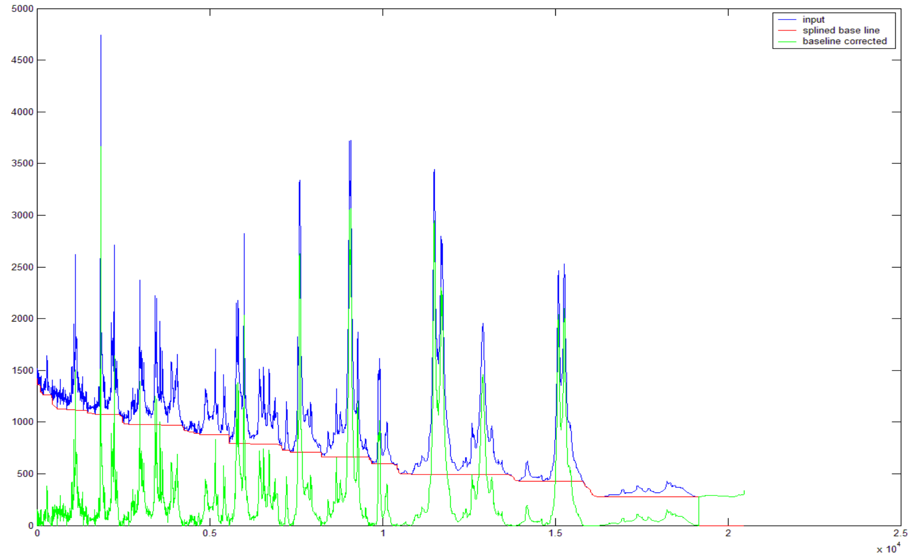

Figure 4.3 Baseline Correction

B. Table

TABLE

Table 1: The real patient distribution

Column 1	Column 2
1-14	Adeno Cancer
15-29	Squamous Cancer
30-34	Large Cancer
35-39	meta- Cancer
39-42	rec-/car Cancer
43-50	Healthy people

VITA

Shuo Chen

Education:

East Tennessee State University, Johnson City, TN

Mathematics, M.S. December, 2004.

Harbin Institute of Technology, Harbin, P. R. China

Electrical Engineering, B.S. July, 2003.

Professional Experience:

Graduate Assistant

East Tennessee State University, Johnson City, TN, 2003-2004.