Chapter 16

MATHEMATICAL FRAMEWORK AND WAVELETS APPLICATIONS IN PROTEOMICS FOR CANCER STUDY

Don Hong and Yu Shyr

Cancer is a proteomic disease. Though MALDI-TOF mass spectrometry allows direct measurement of the protein signature of tissue, blood, or their biological samples, and holds tremendous potential for disease diagnosis and treatment, key challenges remain in the processing of proteomic data. In this chapter, we will introduce a wavelet based mathematical framework and computational tools for proteomic data processing, feature selection, and statistical analysis in cancer study.

Keywords: Proteomics, wavelets, mass spectrometry peak detection, peak alignment, biomarker discovery, feature selection.

1. INTRODUCTION

Proteomics, the analysis of genomic complements of proteins, has attracted more and more attention to cancer researchers due to the fact that cancer is a proteomic disease and protein arrays are a breakthrough because they allow many different proteins to be tracked simultaneously. High throughput *mass spectrometry* (MS) has been motivated greatly from recent developments in both chemistry and biology. Its technology has been extended to proteomics as a tool in rapid protein identification (Chaurand *et al*., 1999; Loo *et al*., 1999). Comparable to the exciting development of nuclear magnetic resonance methods during the past three decades, mass spectrometry entered a phase of rapid growth in the mid-eighties beginning with the introduction of soft ionization methods, such as electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI). These new techniques have allowed the use of mass spectrometry in applications involving large

molecules such as in biochemical, pharmaceutical, and medical research. Mass spectrometric methodology and examples of applications in biotechnology and cancer study can be found in Siuzdak (2003) and Roboz (2002). Some recent progress on automated peak identification for time-of-flight (TOF) mass spectra can be found in Hong and Shyr (2007).

Mass spectrometers are ion optical devices that produce a beam of gas-phase ions from samples. They sort the resulting mixture of ions according to their mass-to-charge (m/z) ratios or a derived property, and provide analog or digital output signals (peaks) from which the mass-to-charge ratio and intensity (abundance) of each detected ionic species may be determined. Masses are not measured directly. Mass spectrometers are m/z analyzers. The mass-to-charge ratio of an ion is obtained by dividing the mass of the ion (m), by the number of charges (z) that were acquired during the process of ionization. The mass of a particle is the sum of the atomic masses (in Dalton) of all the atoms of the elements of which it is composed.

The *mass spectrum* of a compound provides, in a graphical or tabular form, the intensities of all or a selected number of the acquired m/z values from the ionic species formed. Mass spectral peaks are observed in analog form (each peak with a height and a width) or digital form (each peak a simple line). The heart of any mass spectrometer is the mass selective analyzer. The concept of the linear time-of-flight analyzer was described by Stephens in 1946. The development of MALDI-TOF in 1988 (Karas and Hillenkamp) has paved the way for new applications, not only for biomolecules but also for synthetic polymers and polymer/biomolecule conjugates. Accordingly, the major areas of applications of mass spectrometry have been qualitative analysis and quantification.

Cancers secrete large and small molecules of numerous known and countless unknown structures. Enzymes that allow cancers to invade and metastasize, and surface molecules and compounds of unknown function often serve as critical parameters of cancer behavior. Discovery of trace compounds that could indicate the presence of early cancer is still theoretically possible, and still hoped for. Identification of such compounds in extremely small quantities in biologic fluids containing hundreds of other compounds is a classic undertaking for mass spectrometry. Coordinated immunologic assay, isotopic, spectroscopic, nuclear-magnetic resonance, and mass spectrometric analysis of such putative markers could advance

the diagnostic acumen so that we might recognize pre-cancerous states, or cancers so early in their course that a cure could be readily achieved (see Henschke *et al*., 1999; Srinivas *et al*., 2001; and Yanagisawa *et al*., 2003 for examples).

Mass spectrometers attempt to answer the basic questions of WHAT and HOW MUCH is present by determining ionic masses and intensities. MALDI-TOF MS is emerging as a leading technology in the proteomics revolution. Indeed, the year 2002 Nobel prizes in chemistry recognized MALDI's ability to analyze intact biological macromolecules. Though MALDI-TOF MS allows direct measurement of the protein "signature" of tissue, blood, or other biological samples, and holds tremendous potential for disease diagnosis and treatment, key challenges still remain in the processing of MALDI MS data.

The use of high-throughput mass spectrometry produces data sets comprised of spectra whose graphs are of the type shown in Figure 1. On the horizontal axis are mass/charge (m/z) values and on the vertical axis an intensity measurement that indicates a relative abundance of the particle. The analysis of such data involves inferring the existence of a peptide of a particular mass from the existence of a spike in the spectrum. The

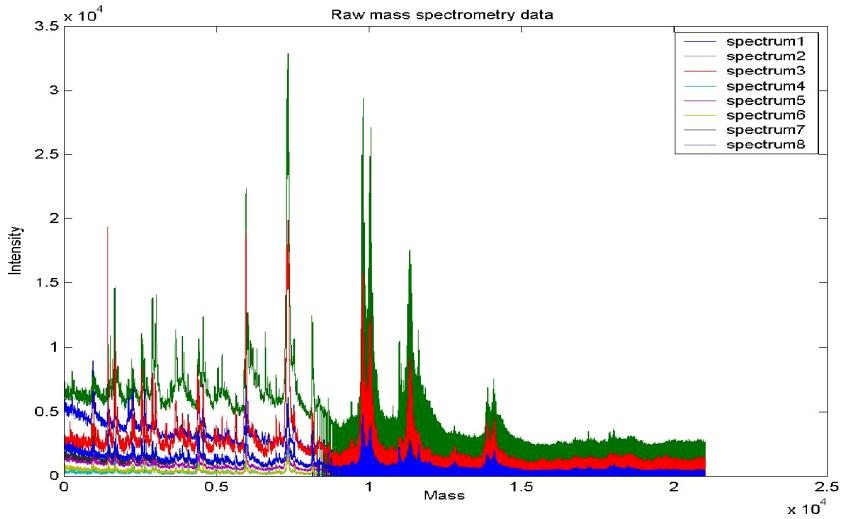

Fig. 1. Graphs of MALDI-TOF mass spectra.

data is in very high dimensional setting and there are uncertainties in peak position and as well the intensity. To identify biomarkers from these spectra, many data-analytic questions arise: What feature indicates the existence of a peptide? How does one even define a feature and subsequently extract it from a set of spectra with significant between-sample variability in intensity, background noise, and in the m/z-value at which a feature is recorded? In addition, data registration is confounded by the variability in the location and the shape/size of features when compared across samples.

To date, MALDI-TOF or SELDI (surface-enhanced laser desorption ionization), a variant of MALDI technology has been applied to search cancer biomarkers, with some success. There is also preliminary evidence that we may be able to discover patterns that can reliably distinguish cancer patients from healthy individuals (Soltys *et al*., 2004; Waldsworth *et al*., 2004; Yanagisawa *et al*., 2003; Zhang *et al*., 2004 for examples). These findings should be greeted with cautious optimism. When it has been possible to identify the protein peaks, they have often turned out to be well-known acute-phase proteins. Some authors have claimed that MS is intrinsically limited in its depth of coverage, with a dynamic range that prevents it from being able to find low-abundance proteins (Diamandis, 2004). This brings us to analysis tools. There is no consensus on the best methods to analyze mass spectra from proteomic profiling experiments. Most published studies perform data preprocessing and peak detection with software from the manufacturers of MALDI/SELDI instruments. In fact, the software is extremely conservative about calling something a peak and its baseline correction algorithm introduces substantial bias into the estimates of the size of a peak. These algorithmic weaknesses can reduce the effective sensitivity of the instrument below its true capacities and can hamper its reproducibility. Many of *ad hoc* approaches have been implemented by various groups (Coombes *et al*., 2005; Morris *et al*., 2005; Yu *et al*., 2006; Chen *et al*., 2007). It is substantial to develop a comprehensive set of mathematical and computational tools for MALDI TOF MS data analysis.

Multiscale tools such as wavelets provide promising techniques for MALDI MS data analysis. The word "wavelets" means "small waves" (the sinusoids used in Fourier analysis are "big" waves), and in short, wavelet is an oscillation that decays quickly. Mathematically, *wavelets* usually are basis functions of an L^2 space that satisfy so-called multiresolution analysis

requirements (Chui, 1992; Daubechies, 1992; Hong *et al*., 2005). In recent years, wavelets have been applied to a large variety of signal processing and image compression (Mallat, 1999). Also, there is a growing interest in using wavelets in analysis of biomedical signals and functional genomics data (see Aldrobi and Unser, 1996; de Trad *et al*., 2002; Hirakawa *et al*., 1999; Lio, 2003 for examples). Wavelet theory is developed now into a methodology used in many disciplines: engineering, mathematics, physics, signal processing and image compression, numerical analysis, and statistics. Wavelets are providing a rich source of useful tools for applications in time-scale types of problems. Wavelet based methods have found applications in statistics in areas such as regression, density and function estimation, modeling and forecasting in time series analysis, and spatial analysis (Donoho and Johnston, 1995, 1998; Silverman, 1999). In particular, Donoho and Johnstone found that wavelet threshold has desirable statistical optimality properties. Since then, wavelets have proved to be very useful in nonparametric statistics and time series analysis.

In the following discussions, we will focus primarily on: (a) establishing a general mathematical framework for modeling and representing MALDI MS data that allows the recovery of, as near as possible, the "true" signal from the machine data (innovative mathematical tools to be developed include non-uniform wavelets, biological diffusion maps and geometric harmonics, and shape-preserving splines); (b) designing algorithms and developing software for performing preprocessing operations such as peak alignment and detection, baseline correction and denoising, and a statistical analysis of MALDI MS data; and (c) developing tools for feature extraction and biomarker discovery. In particular, we will focus on a wavelet based novel multiscale scheme for identifying biological signatures of MALDI-TOF MS cancer data.

2. MATHEMATICAL REPRESENTATION AND PREPROCESSING OF MALDI MS DATA

In this section we model the MS signal as being composed of three distinct components: background function, true signal, and machine noise. This model allows us to address signal reconstruction and subsequent

biological interpretation in a mathematically principle manner. We will explain how each component is created and how previous approaches heuristically address the modeling of each component separately. In the subsequent subsections, we will discuss the specific techniques we employ in order to model each component.

2.1. Mathematical Model for MALDI-TOF MS Data

The operation of a mass spectrometer can be divided into four main steps: sample introduction, ionization, mass analysis, and detection/data analysis. In the ionization stage, an ion-gas is produced from a given sample and in the mass analysis stage. The ions are then separated according to their mass-to-charge ratio (m/z) using electromagnetic fields. The first mass spectrometers were produced in the early 1900s (Thomson, 1913). Mass spectrometry has been used to investigate biological processes since the late 1930s; however, it is only recently that the advances in ionization technology permit the use of mass spectroscopy to study large molecules (up to 300,000 Da), such as proteins or peptide fragments that occur in biological samples. In particular, the MALDI spectrometer has become a central tool in modern protein research. In a MALDI-TOF spectrometer, the analyte is first embedded in a solid "matrix" that absorbs energy from a laser whose wavelength is matched to the matrix. The resulting intense heating of the matrix produces a gas-ion plume that is then accelerated through a potential difference V. An ion of mass m and charge z acquires a change in potential energy of zV which, to first order, is translated into a change of kinetic energy of $(1/2)\,mv^2$, exiting with a velocity determined by the ratio m/z. The ions then travel a length D to a detector where the density of ions is recorded as a function of the time of arrival t. The mass charge ratio m/z may then be expressed as a function of t of the form (Vestal and Juhasz, 1998):

$$m/z = A(t - t_0)^2, \tag{1}$$

where A and t_0 are constants depending on instrumental parameters such as V and D.

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 arrival 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, and deconvolution. In the preprocessing 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.

A mass spectrometer has a finite resolution power mainly due to variations in the initial position and velocity contained in the ion-plume. A "pure" sample consisting of ions of a single mass/charge ratio $y = m/z$ results in an "impulse response" $k(x, y)$ where k depends on the distribution of initial position and velocity along with the value of machine parameters (such as V and D) that are chosen to minimize the resolution $(\sigma(y)/y)$ for y in some interval of interest where $\sigma(y)$ denotes some measure of the spread of $k(x, y)$, for example, standard deviation or half-width at half-maximum. Typically, k is assumed to be of the form: $k(x, y) = \exp\{-(x - y)^2/\sigma(x)\sigma(y))$, and that $\sigma(y)$ is slowly varying over intervals. As shown in (Vestal and Juhasz, 1998), k can be explicitly calculated from the mass analyzer geometry and operating voltages and from the distributions of initial ion position and velocity. Assuming a parametric form for the initial distributions, one can find parametric representations for k.

To first order (ignoring interactions between ions), a mass spectrometer is a linear device and so the output $f(x)$ in the absence of machine noise from a sample with mass/charge distribution μ is of the form:

$$f(x) = \int k(x, y) d\mu(y).$$

Specifically we propose to consider symmetric kernels k of the form:

$$k(x, y) = \frac{\gamma((x - y)^2}{(\sigma(x)\sigma(y))},$$

where: $R_+ \rightarrow R_+$ is a decreasing function with rapid decay. For fixed y, it is usually the case that $\sigma(x)$ is approximately constant for $|x - y| = O(\sigma(y))$ and so $k(x, y)$ is essentially a small symmetric

bump. If the sample distribution is of the form $\mu_s = \sum_i \alpha_i \delta_{xi,}$, where δ_a denotes a unit mass at a, then the observed signal is a sum of bumps of the form:

$$S(x) = \sum_i \alpha_i k(x, x_i).$$

However, a real world signal differs from this idealized scenario in several ways. First, the ion-plume contains a distribution μ_m of ionized matrix molecules having a high spectral content in the low mass region, that is, the matrix produces ions of a wide variety of masses in this range. Secondly, because of collisions or because of molecular fragmentation that occurs during the time of flight, ions are spread non-locally across the mass scale. We model this non-local scattering with a second kernel, κ, with slow decay. This suggests the following "incoherent" contribution $I(x)$ to the observed signal:

$$I(x) = \int k(x, y) d\mu_m(y) + \int \kappa(x, y) d\mu(y),$$

where $\mu = \mu_s + \mu_m$. Finally, a real MS signal contains a high frequency machine electronic machine noise $\varepsilon(x)$, and so we model an observed signal by a sum of the form $f(x) = I(x) + S(x) + \varepsilon(x)$. Generally, we shall also consider a nonlinear component in I.

Based on the above mathematical model for MALDI MS data, multi-scale deconvolution approach can be used in the statistical estimation for the MALDI MS data analysis. Mass spectrometry of proteins promises to be a very valuable tool in diagnostic applications. There are several challenges to the use of such proteomics data in classification and clustering of samples from diseased and normal patients. In particular, the number of measurements taken per sample is very large and even if considered into peak areas or peak heights, the number of potential predictors greatly exceeds the number of samples. Therefore, there has been considerable effort involved in the preprocessing of the data (see Baggerly *et al*., 2004; Coombes *et al*., 2005; Gentzel *et al*., 2003; Morris *et al*., 2005; Chen *et al*., 2007; Yu *et al*., 2006 for examples).

2.1.1. *Baseline correction and normalization*

Incoherent contribution $I(x)$ is usually approximated by a so-called *baseline* or background function. An observed MALDI MS signal $f(t)$ is often modeled as the superposition of three components:

$$f(x) = B(x) + S(x) + \varepsilon(x),$$

where, $B(x)$ is a slowly varying "baseline" that approximates the incoherent component $I(x)$, $S(x)$ is the "true" signal to be extracted, and $\varepsilon(x)$ represents a high frequency machine noise. The underlying assumption in these techniques is that B, S and ε are varying at different scales. Recently, in Coombes *et al*. (2005), Chen (2004), Chen *et al*. (2007), and Hong *et al*. (2007), wavelet-based methods was proposed in the preprocessing of SELDI and MALDI spectra data, respectively. They use baseline correction and wavelet denoising to approximate $S(x)$ and they show better peak detection than previous methods.

From experimental observations, one often includes the constraint that the baseline is non-negative and decreasing. For this purpose we are interested in the use of shape preserving splines with non-uniform knots Chen *et al*. (2007). Furthermore, we propose to use models for $I(x)$ as above to estimate its properties in order to construct better baseline approximations. In particular, we expect that $I(x)$ can be represented as $I(x) = B(x) + W(x)$, where $B(x)$ is now the projection onto a coarse space in a multiresolution and W is a component that can be represented in a wavelet basis with small coefficients. Because the bump width $\sigma(x)$ is not constant, the natural representations of $f(x)$ and $S(x)$ are in non-shift invariant spaces. We believe that using wavelets for baseline estimation and for reconstructing the true signal $S(x)$ can improve the accuracy of the appearance of the spectrum and the quality of a result from subtracting one spectrum from another.

Some simple techniques for constructing a baseline include fitting local minima with a polynomial or spline function (Chen *et al*., 2007), or using a median filter of appropriate window size (Coombes, 2005). See Figure 2 for a comparison of a MALDI mass spectrum on raw data with and without baseline correction (Chen *et al*., 2007).

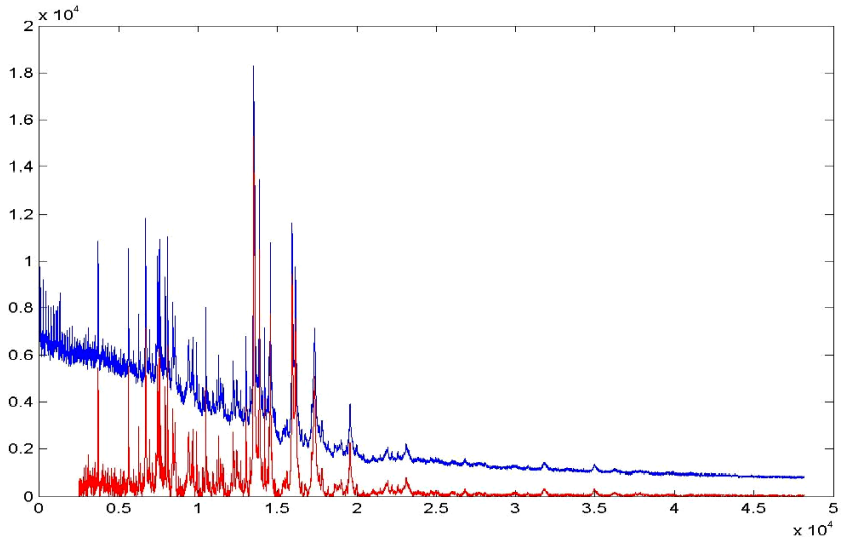

Fig. 2. MALDI MS signals with and without baseline correction.

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 Chen *et al*. (2007), we apply an l_2 averaging formula for the normalization, which is in the energy metric.

2.1.2. *Spectra registration and peak alignment*

Spectrum data registration means that aligning the time of flight data t with m/z as accurately as possible across samples so that different samples may be compared. In the simple case of (1), this means determining the constants A and t_0 for each given sample (typically done by locating known reference mass peaks in the sample). Techniques for the registration of 2D and 3D data have been well-studied and remain an active area of research in imaging science. Many of these techniques use a multiresolution approach by first registering the signals at a "coarse" resolution and then iteratively registering the signals at successively finer resolutions. This results in algorithms both computationally faster and also more "robust" (Unser *et al*., 1995).

The goal of the MALDI-TOF MS data preprocessing is to identify the locations and the intensities of peaks. The spectra, after all previous preprocessing 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. The local maximum selection, without denoising the raw data, will generate more than one peak on a true peak interval. To filter out smaller peaks, an *ad hoc* method based on the ratio of signal and noise (S/N) is proposed in (Coombes *et al*., 2005). In the next section, we will discuss in details on denoising for mass spectrometry data.

Now, let us discuss the cross sample alignment of MS data. For data samples from patients, the data first has to be preprocessed with the proper background subtracted, normalized, and the different fractions combined to obtain one integrated spectrum for each patient. The integrated spectrum is then binned or aligned so that the data for all patients in the sample is formatted in a matrix with one index representing the patients and the other index the peaks (discrete m/z's corresponding to the mean of the m/z of each bin).

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 of each column around the same m/z value maybe different from each other slightly (apart from two to three rows in the matrix). Therefore, we need to bin these peaks in order to correspond to the same m/z value. This is also called cross samples peak alignment. A so-called average spectrum is determined for the binning purpose in Morris *et al*. (2005). An efficient and effective binning method, called PSB, is developed in Hong *et al*. (2007) by projecting spectra to a function of number of peaks. See Figure 3 for PSB results. A so-called central spectrum idea by using local clustering techniques is introduced for binning in Chen (2004). Binning approach reduces the dimension. In Purohit *et al*. (2003), combining the binning procedure, a square root transform on the data is applied to help stabilize the variance and in turn, made a significant improvement in clustering results. A curve alignment method developed in Bar-Joseph *et al*. (2003), which combines spline interpolation with clustering can be employed as an idea for peak selection and binning in the preprocessing

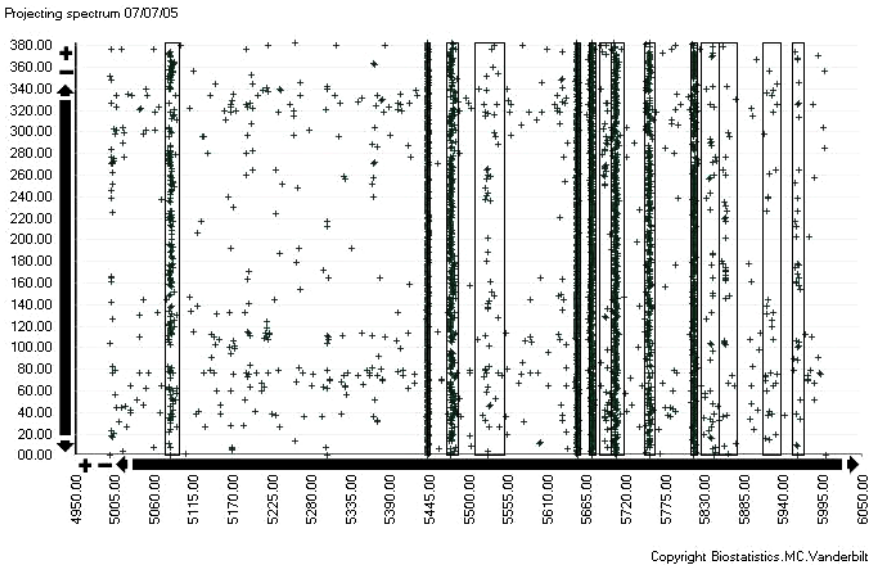

Fig. 3. A partial mass peak distribution with binned by PSB.

step. It will be interesting to compare the outcomes between the binning approach and the curve alignment method, as well as among many choices of multiscale operations.

In summary, we modeled the MALDI MS signal as a superposition of an incoherent signal created by high mass spectral content of the matrix together with non-local scattering, the true observed signal, and an electronics related noise component. We also discussed how the true observed signal is in effect a superposition of distributions of the mass values in the ion plume. In the next section, we discuss in detail the specific mathematical techniques we plan to apply to model each of these components, especially for estimation and removing the noise using wavelets.

3. MULTISCALE TOOLS

We would like to apply multiscale tools, such as wavelets in the study of MALDI MS data. In this section we first briefly introduce wavelets and then discuss how these techniques can be applied to separate the different

components of an observed MALDI MS signal because of the different time-scale characteristics of the components.

3.1. Wavelets and WaveSpec Software

Wavelets are a relatively recent development in applied mathematics. Vidakovic (1999) mentioned that the first definition of wavelets could be attributed to Morlet *et al*. (1982) (Grossmann and Morlet, 1984). Now the term wavelet is usually associated with a function $\psi \in L_2(R)$ such that its translations and dilations

$$\psi_{j,k}(x) = 2^{-j/2}\psi(t/2^j - k),$$

for integers j and k constitute an orthonormal basis of $L_2(R)$. The wavelet transform is a tool that cuts up data or functions into different frequency components, and then studies each component with a resolution matched to its scale. The wavelet transform on a finite sequence of data points provides a linear mapping to the wavelet coefficients: $\mathrm{w}_n = Wf_n$, where the matrix $W = W_{n\times n}$ is orthogonal and w_n and f_n are n-dimensional vectors. The wavelet approximation to a signal function f is built up over multiple scales and many localized positions. For the given family of scale functions and corresponding wavelet functions:

$$\phi_{J,k} = 2^{-J/2}\phi(t/2^j - k),\ \psi_{j,k} = 2^{-j/2}\psi(t/2^j - k), \quad j = 1, 2, \ldots, J.$$

The coefficients are given by the projections:

$$s_{J,k} = \int f(t)\psi_{J,k}(t)dt, \quad d_{j,k} = \int f(t)\psi_{j,k}(t)dt$$

so that

$$f(t) = \sum_k s_{J,k}\psi_{J,k}(t) + \sum_k \sum_{j=1}^{J} d_{j,k}\psi_{j,k}(t).$$

The large J refers to the relatively small number of coefficients for the low frequency, smooth variation of f, the small j refers to the high frequency detail coefficients.

When the sample size n, the number of observations, is divisible by 2, say $n = 2^J$, then the number of coefficients, n can be grouped as $n/2$

coefficients $d_{1,k}$ at the finest level, $n/4$ coefficients $d_{2,k}$ at the next finest level, $\ldots, n/2^J$ coefficients $d_{J,k}$ and $n/2^J$ coefficients $s_{J,k}$ at the coarsest level. Some wavelet applications in cancer data analysis were reviewed recently in Hong and Shyr (2006).

Multiscale analysis tools such as wavelets are providing a rich source of useful tools for applications in time-scale types of problems (Sentelle *et al*., 2002). The Fourier transform extracts details from the signal frequency, but all information about the location of a particular frequency within the signal is lost. Though window Fourier transform (WFT) can help to determine time location for nonstantionary signals, the lack of adaptivity of WFT may lead to a local under- or over-fitting. In contrast to WFTs, wavelets select widths of time slices according the local frequency in the signal. This adaptivity property of wavelets certainly can help to us to determine the location of peak difference(s) of MALDI-TOS MS protein expressions between cancerous and normal tissues in term of molecular weights.

Wavelets, as building blocks of models, are well localized in both time and scale (frequency). Signals with rapid local changes (signals with discontinuities, cusps, sharp spikes, etc.) can be precisely represented with just a few wavelet coefficients.

Wavelets can be useful in detecting patterns in DNA sequences as well. In Lio and Vannucci (2000), it was shown that wavelet variance decomposition of bacterial genome sequences can reveal the location of pathogenicity islands. The findings show that wavelet smoothing and scalogram are powerful tools to detect differences within and between genomes and to separate small (gene level) and large (putative pathogenicity islands) genomic regions that have different composition characteristics. An optimization procedure improving upon traditional Fourier analysis performance in distinguishing coding from noncoding regions in DNA sequences was introduced in Anastassiou (2000). The approach can be taken one step further by applying wavelet transforms. To find the similarities between two or more protein sequences is of great importance for protein sequence analysis. In de Trad *et al*. (2002), a comparison method based on wavelet decomposition of protein sequences and a cross-correlation study was devised that is capable of analyzing a protein sequence "hierarchically," i.e., it can examine a protein sequence at different spatial resolutions. A sequence-scale similarity vector is generated for the comparison of two sequences feasible

at different spatial resolutions (scales). The cosine Fourier series and discrete wavelet transforms are applied in Morozov (2000) for describing replacement rate variation in genes and proteins, in which the profile of relative replacement rates along the length of a given sequence is defined as a function of the site number. The new models are applicable to testing biological hypotheses such as the statistical identity of rate variation profiles among homologous protein families.

Despite advances in instrument resolution and sensitivity of MS technology, the effective resolution is limited by the distribution of naturally isotopes of common elements. This isotopic envelope of molecular weights complicates analysis of spectra when two or more species differ by only a few Daltons. The species exhibit overlapping spectral signatures, and form what is here termed a "peak cluster." The resolution of such clusters would be an important advance in biomedical research in general, and cancer research in particular.

As discussed above, an observed MALDI signal consists of a true signal $S(x)$, an incoherent signal $I(x)$ and machine noise $\varepsilon(x)$. To extract the true signal we need to remove the noise and the incoherent signal from the observed data. In the wavelet representation, the noise ε is concentrated in the fine scale wavelet coefficients and the incoherent signal can be approximated by the projection onto the coarse space spanned by the functions $\phi_{J,k}$. A variety of threshold strategies can be used to remove the machine noise from the data. A baseline can be designed using a coarse approximation and a component with small coefficients in a wavelet space.

The discrete mass spectrum data provide information about the cancer tissue and normal tissue at particular molecular weights. The wavelet approximation to a signal function f is built up over multiple scales and many localized positions. A discrete wavelet transform (DWT) decomposes a signal into several vectors of wavelet coefficients. Different coefficient vectors contain information about the signal function at different scales. Coefficients at coarse scale capture gross and global features of the signal while coefficients at fine scale contain detailed information. Applying wavelet transform to MALDI-TOF MS data, the protein expression difference can be measured at different resolution scales based on a molecular weight-scale analysis. It may reveal more information than other conventional methods.

Following Donoho and Johnstone (1994, 1995), we can apply a variety of threshold techniques for MALDI MS data processing. The idea behind threshold is the removal of small (wavelet) coefficients, considered to be noise. This leaves large coefficients in the multiscale decomposition object that can then be used to estimate the signal after reconstruction. There are many ways to threshold. The universal threshold is computed as

$$\lambda = s\sqrt{(2\log M)},$$

where M is the number of data points (wavelet coefficients) and s is an estimate of the variation of the coefficients on the standard deviation scale. Probability threshold is selecting the pth quantile of the coefficients based on a given probability value p. Soft threshold is to modify the coefficients by the formula:

$$d_{jk}^{new} = sgn(d_{jk})(|d_{jk}| - \lambda)_+$$

for the thresholding scale λ. If the noise process is stationary, one effect of correlated noise is to yield an array of wavelet coefficients with variances that depend on the level j of the transform. This leads to level-dependent threshold, using for each coefficient a threshold that is proportional to its standard deviation (Johnstone and Silverman, 1997). The level-dependent threshold method applied in wavelet regression gives optimally adaptive behavior. Block threshold is to threshold the wavelet coefficients in groups (blocks) rather than individually to increasing estimation accuracy by utilizing information about neighboring coefficients. Since the high frequency components decrease as the mass weight increases, we used block threshold strategy for MALDI-TOF MS data denoising (Chen *et al.*, 2007).

An important development in the statistical context has been the routine use of the non-decimated wavelet transform (NDWT), also called the stationary or translation-invariant wavelet transform, see (Lang *et al.*, 1996; Nason *et al.*, 1995; Walden and Cristan, 1998) for example. Conceptually, the NDWT is obtained by modifying the Mallat DWT algorithm: at each stage, no decimation takes place but instead the filters are padded out with alternate zeros to double their length. The effect is to yield an over determined transform with n coefficients at each of $\log_2 n$ levels. The transform contains the standard DWT for every possible choice of time

origin. Johnstone and Silverman (1997) investigated the use of the NDWT in conjunction with the marginal maximum likelihood approach. The NDWT has been used for SELDI MS data analysis (Coombes, 2005). In general, in WT, hard thresholds have a better l_2 performance while soft thresholds generate better smoothness results. However, with stationary discrete wavelet transform (SDWT), since the coefficients are undecimated, hard thresholds will have both good l_2 performance and smoothness (Coombes, 2005; Lang *et al*., 1996).

When we do wavelet denoising, we are faced with many parameters to choose, such as the type of a mother wavelet, the decomposition level and the values of thresholds. 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 wavelet seems not matter much, while the value of thresholds 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 data sets' properties.

It has been observed that the high frequency components of spectrum data reduce as the mass weight increases because the values of median absolute deviation (MAD) change a lot throughout different m/z segments. MAD/0.67 is a robust estimate of the non-normal variability. This phenomenon might be caused by that the machine has relative low resolution for ions of small m/z values at low m/z interval. Therefore, we should set different thresholds at different mass segments by the changing trend of the coefficients at each level (Lavielle, 1999). In this way, the denoised signal can reduce the variance in the beginning part and retain the useful information in the posterior part.

In cancer research projects carried at Vanderbilt Ingram Cancer Center (VICC), we observed that most coefficients at levels from 1 to 4 are dumped. The reason is that they are of high frequency and low energy (the proportion of the total energy of the signal is only 10^{-12}). We also need to be very cautious when manipulating the low frequency components. We believe choosing threshold values based on the exploratory data analysis will achieve better wavelet denoising performance. This denoising method performed well in the study (Chen *et al*., 2007).

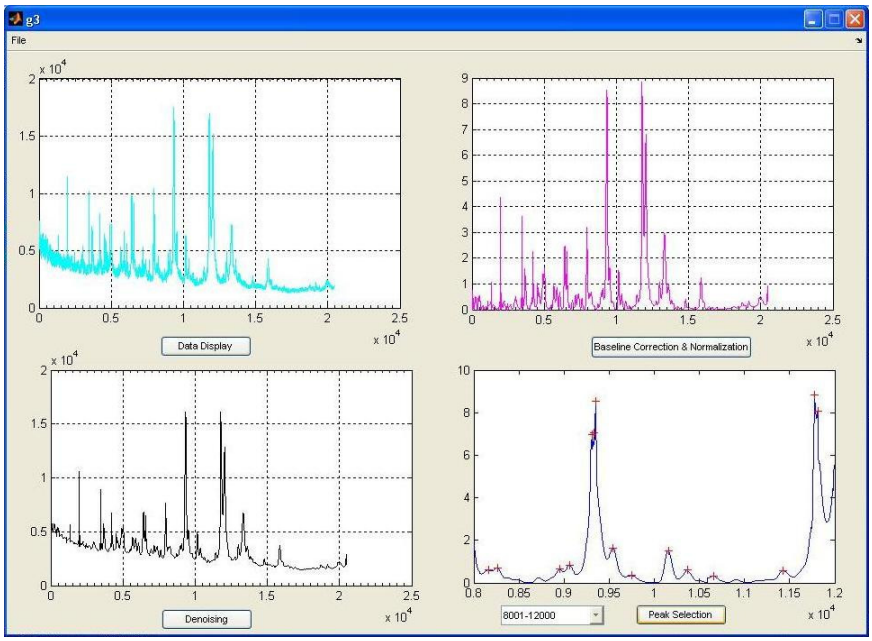

Fig. 4. Graphical user interfaces of WaveSpec software.

A software package called WaveSpec implementing the mathematical framework has been developed at Biostatistical Shared Resource of VICC. A MatLab based version of the software has been used to serve cancer research groups in VICC for MALDI-TOF MS data processing. Figure 4 shows a graphical user interface (GUI) of the software.

3.2. Diffusion Maps

Very recently, diffusion maps and geometric harmonics were introduced to understand the geometric structures of the data sets (Coifman *et al*., 2005a, 2005b). In continuous Euclidean setting, tools from harmonic analysis, such as Fourier transforms and wavelet decompositions have proven to be highly successful in image compression, signal processing, denoising, and density estimation. In statistical data analysis, it is essential to organize graphs and data sets geometrically. Geometric diffusion is a tool for structure definition of data by extending multiscale harmonic analysis to discrete graphs and subsets of R^n.

A diffusion kernel on the data set is explicitly constructed and a diffusion map is defined by employing the spectral properties, spectrum and eigenfunctions. Also, a multiscale extension scheme is defined for decomposing empirical functions into frequency bands and showing the links between the intrinsic and extrinsic geometries of the set.

Coifman *et al*. (2005) introduced a family of diffusion maps that allow the exploration of the geometry, the statistics, and functions of the data. Diffusion maps provide a nature low-dimensional embedding of high-dimensional data that is suited for subsequent tasks such as visualization, clustering, and regression. It will be interesting to follow diffusion map's idea by emphasizing on the biological meaning and chemical structure of the mass spectrometry data set. The kernel in the model discussed above is symmetric and non-negative and we can define an associated diffusion map for such a kernel. The formalization permits the proper identification and estimation of a wavelet spectrum. Once the characteristic frequency for a particular biological function has been determined, it is possible to identify the individual mass spectrum's "hot spots" using wavelet transform that contribute mostly to the characteristic frequency and also to the protein's biological function (de Trad *et al*., 2002). A suitable defined biological diffusion map will give potential improvements in the early detection and diagnosis of various types of cancer. It would be great to obtain a biological diffusion map for decomposing MALDI MS data into frequency bands and showing the links between the intrinsic and extrinsic biology of the data.

4. CLUSTERING AND CANCER DATA CLASSIFICATIONS

Mass spectra are intrinsically functional observations, and are well-suited to wavelet methods. We would like to apply multiscale techniques to further study preprocessed mass spectra for feature extraction. The significant difference in the findings would help the identification of protein markers.

Biomarkers are measurable molecular phenotypic parameters that characterize an organisms state of health or disease, or a response to a particular therapeutic intervention. Biomarkers are sought as instruments to help in disease risk assessment, early disease detection, and as surrogate

endpoints in clinical trials (or in some cases as surrogates for environmental and other exogenous factors such as diet). Establishing/validating biomarkers include the following steps: (a) identify candidates, (b) conducting clinical assays to diagnose known disease, (c) detection of preclinical disease (pseudo-prospectively) and establishment of screen-positive rule, (d) prospective screening, establish extend and characteristics of identified disease as well as false referral rates, and (e) quantification of overall impact on disease. The biomarker selection problems maps nicely to the problem of feature selection for classification in statistics and machine learning. Feature selection is the problem of selecting a subset of variables of minimal size that can predict, classify, or diagnose a target variable of interest as well as, or better than, the full set of available predictors. In the case of MALDI-TOF MS signals, biomarkers could be individual masses, individual mass distributions or could be expressed in terms of wavelet coefficients or principal components. Selecting a minimal set of predictors with maximum accuracy is important for treating the curse of dimensionality, for reducing the cost of observing the required variables for prediction, and for gaining insight into the domain.

There are several families of methods biomarker selection on the reconstructed MALDI-TOF MS signal. *Principal component analysis* (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. A widely used technique for the representation of sensor data is based on diagonalizing the correlation tensor of the data-set, keeping a small number of coherent structures (eigenvectors) based on principal components analysis (PCA). This approach tends to be global in character. It is possible to combine multiscale analysis and PCA to obtain proper accounting of global contributions to signal energy without loss of information on key local features. We can exploit such a combined wavelet-PCA technique in MALDI data processing.

Recently, we express MALDI MS data, after using WaveSpec preprocessing, in terms of a convex combination of dominant biological components based on principal component data for an initial investigation of

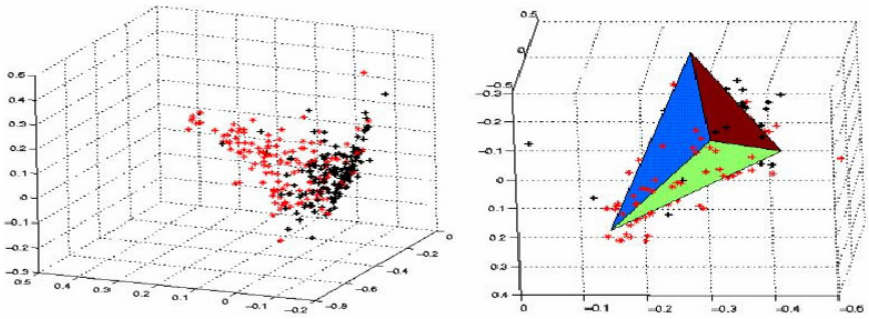

Fig. 5. MALDI-TOF MS lung cancer data plot based on selected three principal components (*left*). Simplex representation for MALDI-TOF MS lung cancer data (*right*).

feature extraction on MALDI-TOF MS data in lung cancer study (Hardin and Hong, 2006). Figure 5 shows that a convex combination (simplex) expression using super positions provides a promising tool for cancer feature extraction. The choice of analyzing wavelet basis leads to highlighting of certain features through the strengthening of a small set of coefficients, leaving the remainder at low amplitudes. It will be interesting in investigating adaptive wavelet-PCA approaches to pattern extraction from MALDI MS data, signal inversion and feature enhancement in the presence of noise.

A reliable and precise classification of tumor is essential for success in diagnosis and treatment of cancer. DNA microarrays have been used to characterize the molecular variations among tumors by monitoring gene expression profiles on a genomic scale. MALDI-TOF MS can profile proteins up to 50 kDa in size in tissue. Contributions of mass spectrometry to this infant field are largely untapped. This technology can not only directly assess peptides and proteins in sections of tumor tissue, but also can be used for high resolution image of individual biomolecules present in tissue sections. The protein profiles obtained can contain thousands of data points, necessitating sophisticated data analysis algorithms.

Clustering assigns samples to classes on the basis of their distance from objects known to be in the classes. The distance or similarity will have a large effect on the performance of the classification procedure. In orthogonal wavelet transform, the L_2 distance for the signals is equivalent to the l_2 distance for the vectors of wavelet coefficients. Therefore, we can perform clustering of MALDI MS data using wavelet coefficients with

Euclidean distance. If we view the wavelet coefficients in different scales as a microarray data set, then methods in microarray data analysis may be used for MALDI MS data analysis via wavelet coefficients as well. Shyr and Kim developed weighted flexible compound covariate methods (WFCCM) for classifying microarray data (Shyr and Kim, 2003). We can apply such methods to MALDI MS data for classification as well.

Lung cancer (Hoffman, 2002) is usually not detected early, and thus may be diagnosed at an advanced stage, where intervention or therapy is less effective. Although the incidence rate of lung cancer is lower than for breast and prostate cancer, the mortality rate of lung cancer is the highest for all cancers in both men and women. Lung cancer kills more Americans each year than the next four leading cancer killers, cancers of the colon, breast, prostate and pancreas, combined.

Precisely classifying tumors is of critical importance to cancer diagnosis and treatment. Recently, there is increasing interest in changing the basis of tumor classification from morphologic to molecular. Mass spectrometry of proteins promises to be a very valuable tool in diagnostic applications. There are several challenges to the use of such proteomics data in classification and clustering of samples from diseased and normal patients. In the following, we mention a clustering method applied to the preprocessed MALDI-TOF MS data from lung cancer patients, which was collected at VICC, using WaveSpec software (Chen, 2004).

Clustering analysis, as a multivariable statistics technique, 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 an example of 50 patients which will be discussed in detail later, there are many tissue mass spectra from healthy people and several groups of sick patients who have different types of cancer. We can see that even we do not know the distribution in advance, by clustering we can divide the spectra into several groups that are almost the same as the real distribution.

Generally, we build the model in the following: the initial object can be modeled as a $p \times n$ matrix for 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 submatrices: $p \times n_1$, $p \times n_2, \ldots,$ for $\sum n_i = n$.

Hierarchical clustering and k-means clustering are two main clustering methods. 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, but rather a multi-level hierarchy, where clusters are more similar at the lower level, which allows you to decide what level or scale of clustering is most appropriate for your data.

For a $p \times n$ matrix corresponding to n vectors (objects), we use a metric (distance) to group them according to their relationship (similarity). For two vectors x and y, both having length p, some common distances are Euclidean distance: $d(x, y) = [\sum(x_i - y_i)^2]^{1/2}$, Manhattan distance: $d(x, y) = \sum |x_i - y_i|$, and correlation distance: $d(x, y) = 1 - \rho(x, y)$, where $\rho(x, y)$ is the correlation coefficient of x and y. Different distances may lead to the 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 the cluster hierarchy tree such as shortest/longest distance, average distance and centroid distance. Matlab software has a function to display the hierarchical tree. To determine where to divide the hierarchical tree into clusters, we need to choose proper cutoff points so that we can cut the trees into several groups.

Comparing with the tree structure of hierarchical clustering, the k-means clustering method has set up the number of groups before clustering. Then all objects are 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 a 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 any further.

Now, we have divided the $p \times n$ matrix into k clusters, but not all the elements in one cluster are different from the ones in other clusters. Therefore, we should figure out which elements are distinct in one cluster, in other words, the characteristic elements. For any two clusters, we can do pair t-test to the objects and find the rows of small p-values; or we can find

the weighted average distance at a certain row:

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

where d_B is the distance between cluster centers, d_{wi} is the average (Euclidean) distance among all sample pairs in one cluster, and $k_i = n_i/(n_1 + n_2)$ (Goldstein *et al.*, 2002). Basically, the objects in the same cluster are close to each other but the distances between centers of different clusters are large. At last, if the distance of two objects is great or the p-value is small enough, then we can say that these elements are distinct. A center spectrum defined for binning scheme in (Chen, 2004) can be used for clustering as well.

As an example, we consider the MALDI TOF MS data set of 50 patients collected at VICC. The tissue sample consists of normal samples and cancer samples of Adeno, squamous, large, and other cancers. We apply the clustering analysis to the 1628×50 matrix, and then compared the results of clustering with the real data distribution.

Figures 6 and 7 show the results using the hierarchical clustering with the Euclidean distance and correlation distance, respectively.

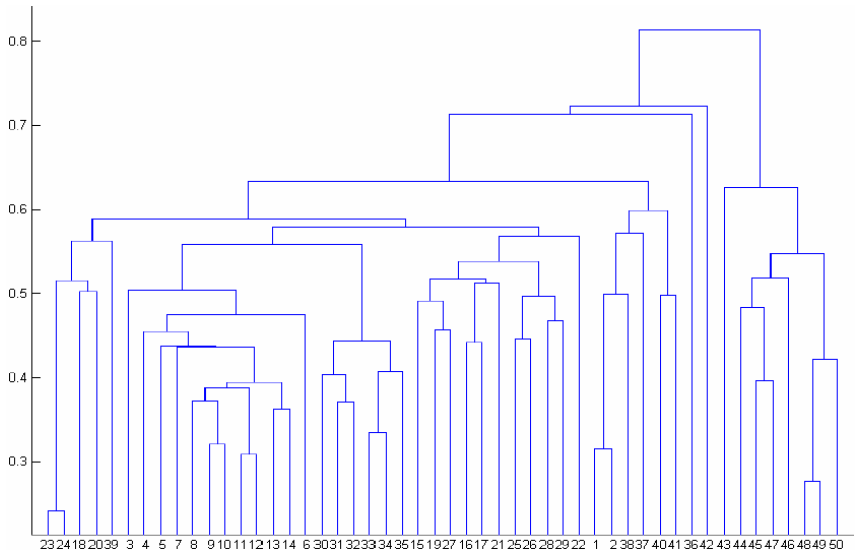

Fig. 6. Hierarchical trees by Euclidean distance.

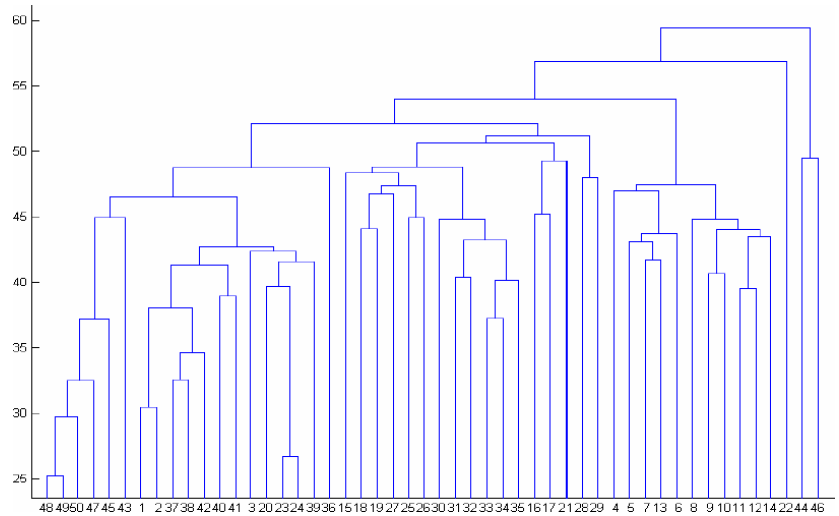

Fig. 7. Hierarchical trees by correlation distance.

Table 1. Lung cancer patients data distribution.

Labels	Cancer types
1–14	Adeno cancer
15–29	Squamous cancer
30–34	Large cancer
35–39	Meta-cancer
39–42	Other cancers
43–50	Normal

From the hierarchical trees, we can see that the clustering results match almost perfectly the real distribution data. In addition, the correlation method seems having a better performance (see Figures 6, 7, and Table 1 for comparison).

Also, when we exam the distinct elements of the normal and cancer cluster, we find that the rows: 38, 350, 356 are most significantly different and the p-values of them, which was provided by MatLab clustering program, are 0. Moreover, rows 38, 350, 356, 953, 986, and 991 are relatively distinct. That means these proteins are very likely different disease-related.

ACKNOWLEDGMENTS

The authors are grateful to Dean Billheimer, Shuo Chen, Doug Hardin, Huiming Li, Ming Li, and Jonathan B.G. Xu for their valuable discussions. This work was partially supported 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 Shyr and by National Science Foundation (IGMS 0552377), National Security Agency (#H98230-05-1-0304), and Research Enhancement Program Award from Middle Tennessee State University for Hong.

References

1. Aldroubi A and Unser M. *Wavelets in Medicine and Biology*. CRC Press, Boca Raton, FL, 1996.
2. Anastassiou D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 2000; **16**: 1073–1081.
3. Baggerly KA, Morris JS and Coombes KR. Reproducibility of SELDI mass spectrometry patterns in serum: comparing proteomic data sets from different experiments. *Bioinformatics* 2004; **20**: 777–785.
4. Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS and Simon I. Continuous representations of time-series gene expression data. *J. Comput. Biol.* 2003; **10**: 341–356.
5. Chaurand P, Stoeckli M and Caprioli RM. Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal. Chem.* 1999; **71**: 5263–5270.
6. Chen S. *MALDI-TOF MS Data Processing Using Splines, Wavelets and Clustering Techniques*. Master's Thesis in Mathematical Sciences, East Tennessee State University, Johnson City, Tennessee, 2004.
7. Chen S, Hong D and Shyr Y. Wavelet-based procedures for proteomic mass spectrometry data processing. *Comput. Stat. Data Anal.* 2007; **52**: 211–220.
8. Chen HJ, Tracy ER, Cooke WE, Semmes OJ, Sasinowski M and Manos DM. Automated peak identification in a TOF-MS spectrum. In: *Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques*. Hong D and Shyr Y (eds.) World Scientific Publication, LLC, New Jersey, 2007, pp. 113–131 .
9. Chui CK. *An Introduction to Wavelets*. Academic Press, New York, NY, 1992.

10. Coifman RR, Lafon SS, Lee AB, Maggioni M, Nadler B, Warner F and Zucker SW. Geometric diffusion as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. USA* 2005a; **102**: 7426–7431.
11. Coifman RR, Lafon SS, Lee AB, Maggioni M, Nadler B, Warner F and Zucker SW. Geometric diffusion as a tool for harmonic analysis and structure definition of data: multiscale methods. *Proc. Natl. Acad. Sci. USA* 2005b; **102**: 7432–7437.
12. Coombes KR, Tsavachidis Morris JS, Baggerly KA, Hung MC and Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* 2005; **5**: 4107–4117.
13. Daubechies I. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.
14. de Trad CH, Fang Q and Cosic I. Protein sequence comparison based on the wavelet transform approach. *Protein Eng.* 2002: **15**: 193–203.
15. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell Proteomics* 2004; **3**: 367–378.
16. Donoho DL and Johnstone IM. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* 1995; **90**: 1200–1224.
17. Donoho DL and Johnstone IM. Minimax estimation via wavelet shrinkage. *Ann. Stat.* 1998; **26**: 879–921.
18. Gentzel M, Kocher T, Ponnusamy S and Wilm M. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* 2003; **8**: 1597–610.
19. Grossmann A and Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.* 1984; **15**: 723–736.
20. Hardin D and Hong D. Convex combination expression using super positions for cancer data feature extraction. *Manuscript under preparation*, 2006.
21. Henschke CI, McCauley DI, *et al*. Early lung cancer action project: overall design and findings from baseline screening. *Lancet* 1999; **354**: 99–105.
22. Hirakawa H, Muta S and Kuhara S. The hydrophobic cores of proteins predicted by wavelet analysis. *Bioinformatics* 1999; **15**: 141–148.
23. Hoffman PC, Mauer AM and Vokes EE. Lung cancer. *Lancet* 2000; **355**: 479–485.

24. Hong D, Li HM, Li M and Shyr Y. Wavelets and projecting spectrum binning for proteomic data processing. In: *Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques*. Hong D and Shy Y (eds.) World Scientific Publication, LLC, New Jersey, 2007, pp. 159–178.
25. Hong D and Shyr Y. Wavelet applications in cancer study. *J. Concrete Appl. Math.* 2006; **4**: 505–521.
26. Hong D and Shyr Y. *Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques*. World Scientific Publication, LLC, New Jersey, 2007.
27. Hong D, Wang JZ and Gardner R. *Real Analysis with an Introduction to Wavelets*. Academic Press, New York, 2005.
28. Johnstone IM and Silverman BW. Wavelet threshold estimators for data with correlated noise. *J. R. Stat. Soc. B* 1997; **59**: 319–351.
29. Karas K and Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem*. 1988; **60**: 2299–2301.
30. Lang M, Guo H, Odegard JE, Burrus CS and Wells RO Jr. Noise reduction using an undecimated discrete wavelet transform. *Signal Process. Lett. IEEE* 1996; **3**: 10–12.
31. Lio P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* 2003; **19**: 2–9.
32. Lio P and Vannucci M. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics* 2000; **16**: 932–940.
33. Loo JA, Dejohn DE, Du P, Stevenson TI and Ogorzalek-Loo RR. Application of mass spectrometry for target identification and characterization. *Med. Res. Rev.* 1999; **19**(4): 307–319.
34. Mallat S. *A Wavelet Tour of Signal Processing*. Academic Press, New York, 1999.
35. Morlet J, Arens G, Fourgeau E and Giard D. Wave propagation and sampling theory. *Geophysics* 1982; **47**: 203–236.
36. Morozov P, Sitnikova T, Churchill G, Ayala FJ and Rzhetsky A. A new method for characterizing replacement rate variation in molecular sequences: application of the Fourier and wavelet models to Drosophila and mammalian proteins. *Genetics* 2000; **154**: 381–395.
37. Morris JS, Coombes KR, Koomen JM, Baggerly KA and Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 2005; **21**: 1764–1775.
38. Nason GP and Silverman BW. The stationary wavelet transform and some statistical applications. In: *Wavelets and Statistics*, Lecture Notes in Statistics,

No. 103. Antoniadis A and Oppenheim G. (eds.) Springer-Verlag, New York, 1995, pp. 281–300.

39. Purohit PV and Rocke DM. Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics* 2002; **3**: 1699–1703.
40. Roboz J. *Mass Spectrometry in Cancer Research*. CRC Press LLC, Boca Raton, Florida, 2002.
41. Sentelle S, Sentelle C and Sunton MA. Multiresolution-based segmentation of calcifications for the early detection of breast cancer. *Real-Time Imaging* 2002; **8**: 237–252.
42. Siuzdak G. *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, San Diego, CA, USA, 2003.
43. Silverman BW. Wavelets in statistics: beyond the standard assumptions. *Philos. Trans. R. Soc. Lond. A* 1999; **357**: 2459–2473.
44. Soltys SG, Le QT, Shi G, Tibshirani R, *et al*. The use of plasma SELDI TOF mass spectrometry proteomic patterns for detection of head and neck squamous cell cancers. *Clin. Cancer Res.* 2004; **10**: 4806–4812.
45. Srinivas PR, Srivastava S, Hanash S and Wright GL Jr. Proteomics in early detection of cancer. *Clin. Chem.* 2001; **47**: 1901–1911.
46. Thomson JJ. *Rays of Positive Electricity and Their Application to Chemical Analysis*. Longmans, Green and Co., London, 1913.
47. Vestal M and Juhasz P. Resolution and mass accuracy in matrix-assisted laser desorption ionization-time of flight. *J. Am. Soc. Mass Spectrom.* 1998; **9**: 892–911.
48. Walden AT and Cristan AC. Matching pursuit by undecimated discrete wavelet transform for non-stationary time series of arbitrary length. *Stat. Comput.* 1998; **8**: 205–219.
49. Waldsworth JT, Somers KD, Cazares LH, Malik G, *et al*. Serum protein profiles to identify head and neck cancer. *Clin. Cancer Res.* 2004; **10**: 1625–1632.
50. Yanagisawa K, Shyr Y, Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, Edgerton M, Gonzalez A, Nadaf S, Moore JH, Caprioli RM and Carbone DP. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* 2003; **362**: 433–439.
51. Yu WC, Wu BL, Lin N, Stone K, Williams K and Zhao HY. Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Comput. Biol. Chem.* 2006; **30**: 27–38.
52. Zhang Z, Bast RC, Yu Y, Li J, *et al*. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res.* 2004; **64**: 5882–5890.