

2014

## Enhancement of MS Signal Processing for Improved Cancer Biomarker Discovery

Qian Si

*College of William & Mary - Arts & Sciences*

Follow this and additional works at: <https://scholarworks.wm.edu/etd>



Part of the [Physics Commons](#)

---

### Recommended Citation

Si, Qian, "Enhancement of MS Signal Processing for Improved Cancer Biomarker Discovery" (2014). *Dissertations, Theses, and Masters Projects*. Paper 1539623639. <https://dx.doi.org/doi:10.21220/s2-54gz-nt12>

This Dissertation is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact [scholarworks@wm.edu](mailto:scholarworks@wm.edu).

**Enhancement of Mass Spectra Signal Processing  
For Improved Cancer Biomarker Discovery**

**Qian Si**

**Liyang, Jiangsu, China**

**Master of Science, College of William and Mary, 2008  
Bachelor of Science, Nanjing University, 2006**

**A Dissertation presented to the Graduate Faculty  
of the College of William and Mary in Candidacy for the Degree of  
Doctor of Philosophy**

**Department of Physics**

**The College of William and Mary  
January 2014**

**COPYRIGHT PAGE**

**©2013, Qian Si**

**All Rights Reserved**

## APPROVAL PAGE

This Dissertation is submitted in partial fulfillment of  
the requirements for the degree of

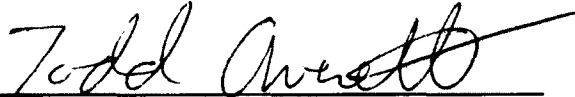
Doctor of Philosophy

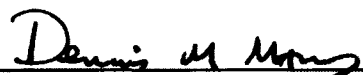
  
\_\_\_\_\_  
Qian Si

Approved by the Committee, January 2014

  
\_\_\_\_\_  
Committee Chair

Professor William Cooke, Physics  
The College of William & Mary

  
\_\_\_\_\_  
Professor Todd Averett, Physics  
The College of William & Mary

  
\_\_\_\_\_  
CSX Professor Dennis Manos, Physics and Applied Science  
The College of William & Mary

  
\_\_\_\_\_  
Chancellor Professor Eugene Tracy, Physics  
The College of William & Mary

  
\_\_\_\_\_  
Assistant Professor Daniel Vasiliu, Mathematics  
The College of William & Mary

## ABSTRACT

Technological advances in proteomics have shown great potential in detecting cancer at the earliest stages. One way is to use the time of flight mass spectroscopy to identify biomarkers, or early disease indicators related to the cancer. Pattern analysis of time of flight mass spectra data from blood and tissue samples gives great hope for the identification of potential biomarkers among the complex mixture of biological and chemical samples for the early cancer detection. One of the keys issues is the pre-processing of raw mass spectra data. A lot of challenges need to be addressed: unknown noise character associated with the large volume of data, high variability in the mass spectroscopy measurements, and poorly understood signal background and so on. This dissertation focuses on developing statistical algorithms and creating data mining tools for computationally improved signal processing for mass spectrometry data. I have introduced an advanced accurate estimate of the noise model and a half-supervised method of mass spectrum data processing which requires little knowledge about the data.

## TABLE OF CONTENTS

<b>Acknowledgements</b>	<b>ii</b>
<b>Dedications</b>	<b>iii</b>
<b>List of Tables</b>	<b>iv</b>
<b>List of Figures</b>	<b>v</b>
<b>Chapter 1. Introduction</b>	<b>1</b>
<b>Chapter 2. Peak Detection Using A Maximum Likelihood Filter</b>	<b>17</b>
<b>Chapter 3. Data Analysis</b>	<b>48</b>
<b>Chapter 4. Peak Alignment</b>	<b>75</b>
<b>Chapter 5. Conclusion</b>	<b>96</b>
<b>Bibliography</b>	<b>100</b>
<b>Vita</b>	<b>105</b>

## ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of many people.

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. William Cooke. It has been my great honor and pleasure to be his Ph.D. student for the past years. I thank him for the invaluable guidance, endless support, rich experience and great patience. I am especially grateful for his confidence and the freedom he gave me to do this work. This work would not be possible if otherwise.

Special gratitude goes to Dr. Dennis M. Manos for his helpful insights, thoughtful suggestions and excellent questions, which keep me moving forward. I would like to extend my gratitude to all of our group members for actively listening and contributing their thoughts during the group meetings.

I would also like to thank my committee members; I really appreciate their valuable time and efforts in reviewing this thesis.

It is a pleasure to thank my friends in William and Mary for their tremendous personal support throughout past years.

In addition, I am grateful to all staff members in the Physics Department for their assistance over the years. My Ph.D. journey in Small Hall would not have been this smooth without their help.

Finally, I would express my deepest gratitude to my parents for their greatest love and confidence all these years. It is their hard work and support which make me possible to go this far.

## LIST OF TABLES

**Table 1: Comparison of the similar master peaks from normal and leukemia data.**

**95**



## LIST OF FIGURES

Figure 1: Typical time of flight spectrum, showing a series of sharp peaks.	7
Figure 2: Time lag focusing for ions of the same $m/z$ but different initial velocity.	9
Figure 3: Mass spectrum of mass focusing region from 5800 to 11000 in time units.	13
Figure 4: Outside the mass focusing range, peaks have increasing widths in time units.	21
Figure 5: A simulated spectrum containing peaks with a noisy background.	29
Figure 6: An expanded view of the simulated spectrum and likelihood of Figure 5.	30
Figure 7: The cumulative distribution function ( <i>CDF</i> ) of a normal distribution, and a linear approximation (red line) to it, near the mean value.	32
Figure 8: Result from a robust estimate of background noise.	34
Figure 9: The difference between the estimated noise and the value used to generate simulated spectra as a function of the number of points in simulated spectra. The error bars show the standard deviation of these estimates for a sample of 10 spectra. Each spectrum was generated from a population with zero mean and unit standard deviation. The black dots show the expected error in the noise estimate.	35
Figure 10: Result of the automatic noise estimate for a typical spectrum.	36
Figure 11: Estimated signal to noise ratio value (lower plot) constructed from a simulated spectrum (upper plot).	38
Figure 12: Magnified view of the data in Figure 11.	39

Figure 13: The distribution of <i>SNR</i> from a simulated noise spectrum in (A), compared to the expected <i>CDF</i> in (B), plotted on a log scale in (C), and with the linear fit in (D).	41
Figure 14: Sorted <i>SNR</i> (black dots) for a simulated pure noise spectrum, with a linear fit (red) to the log ( <i>SNR</i> ) applied to the middle. The horizontal line shows the estimated <i>SNR</i> threshold.	42
Figure 15: Cumulative distribution of <i>SNR</i> from simulated spectrum.	43
Figure 16: Calculated <i>SNR</i> threshold (red line in B and C) for simulated noisy spectrum in (A).	44
Figure 17: Peak picking for a simulated spectrum.	47
Figure 18: A typical spectrum before (A) and after (B) correcting the exponentially decaying background due to the matrix clusters.	50
Figure 19: Peak picking using a floating baseline model.	52
Figure 20: If the baseline is assumed to be zero, then the estimated peak amplitude will mimic any pedestals, giving a <i>SNR</i> that stays high.	53
Figure 21: Eliminating the background.	55
Figure 22: Small peak located at 10497 time steps sits on the shoulder of nearby large peak locate at 10460 time steps.	58
Figure 23: Correcting the local background on the shoulder of a large peak.	59
Figure 24: Typical Leukemia Spectra before and after removing the local background.	60
Figure 25: Cluster detection is necessary for local background removal.	62
Figure 26: A cluster on the average of all leukemia spectra between 9840 and 9920 time steps.	63
Figure 27: Large peaks can raise the <i>SNR</i> due to their tails.	65
Figure 28: Estimating the local <i>SNR</i> threshold with a sliding window.	66

Figure 29: The local estimate of the <i>SNR</i> threshold eliminated stays above the high <i>SNR</i> values at on the tails of the large peaks.	68
Figure 30: A view of a larger region shows that the local <i>SNR</i> threshold increases where peaks are large and dense so that the <i>SNR</i> seldom exceeds the threshold due to pedestal effects.	69
Figure 31: When the region likely to contain a peak is too large, the peak picker sometimes finds a false peak at the end of the region as the likelihood returns to its high noise-only value.	70
Figure 32: Restricting the above threshold region means the likelihood will be a maximum at the true peak.	71
Figure 33: A typical spectrum showing the results of the peak picking algorithm (red X).	74
Figure 34: Constructed peak density (C) for peaks located from 9200 to 9260 (A) at time.	78
Figure 35: A start time shift makes the red spectrum approximately 3 time steps earlier than the black spectrum.	80
Figure 36: Starting time shift values for leukemia spectra.	81
Figure 37: Reported peak positions from 200 leukemia spectra in (A) and (C) compared to peak positions with starting time correction (C) (D) at same location.	82
Figure 38: Flow chart of alignment algorithm.	83
Figure 39: Window bin setting example.	86
Figure 40: Average position uncertainties vs. SD of peaks contributed to the final align peak in normal data.	88
Figure 41: Overview of alignment	90
Figure 42: Differences in reported peak position of leukemia data and normal data set at <i>m/z</i> .	91

# Chapter 1 Introduction

## 1.1. Early Cancer Detection

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal cells. It can result in death if the spread is not controlled. Cancer is the second most common cause of death in the US, accounting for nearly 1 of every 4 deaths. In 2012, estimates were that ~577,000 Americans died of cancer, more than 1,500 people a day, and ~1.6 million new cases of cancer cases were diagnosed in that year [1].

Early detection of cancer is crucial for the ultimate control and prevention. In many cases, cancer is not diagnosed and treated until cancer cells have already invaded surrounding tissues; most conventional therapies are limited once a tumor has spread beyond the tissue of origin. Detecting cancers at their earliest stages gives a higher probability of truly curing the disease by current or future treatment strategies. For example, the survival rate of prostate cancer at early stage is about 100% while the rate is only 34% when cancer is detected at advanced stage [1]. So the problem: how to find means to detect early stage cancers.

Recent technological advances in proteomics, which focuses on protein characterization, protein identification and protein function, have been introduced as a new method for early cancer detection [2, 3, 6]. All cancers involve the malfunction of genes that control cell growth and division. Proteomics may be defined simply as the large-scale characterization of proteins expressed by the genome. Even in the very early stages of disease, significant changes may arise in the type and quantity of proteins and peptides produced by the human body. The goal of proteomics is a comprehensive, quantitative description of protein expression. Unlike the study of a single protein or pathway, proteomic methods enable a systematic overview of expressed protein profiles, which ultimately could improve the diagnosis, prognosis, and management of patients by revealing the protein interactions affecting overall tumor progression. Furthermore, differential protein expression analysis can be used to indicate a range of protein markers potentially indicative of disease. Thus, one hopes that proteomics will uncover candidate markers and to indicate mechanisms that are in need of greater analysis. Although the proposition of finding protein expression differences in samples from distinct clinical groups may seem fairly straightforward, in reality it can be an extraordinary challenge. Pattern analysis of mass spectra of blood samples in particular has attracted attention as a potentially effective and efficient approach to identify potential biomarkers for early stage detection of cancer.

## **1.2. Biomarker Discovery**

A biomarker, also known as molecular marker, biological marker, or tumor marker, is a biologically derived molecule in the body that indicates the progress or status of a disease. When a biological fluid, such as blood, is measured, a protein “profile” may be developed. This leads to the potential for finding biomarkers that are over-expressed, under-expressed, or modified. Such biomarkers can then be used to differentiate pathological states (disease) from normal states or to assess and guide drug treatments. If desired, the discovered biomarker can be chemically extracted for further analysis. The concentration level or pattern of biomarkers related to a certain type of cancer can be served for early detection or diagnosis [4, 5, 6, 17, 23].

Pattern analysis of multiple biomarkers in blood samples has attracted attention as an alternative to the usage of a single biomarker for early detection of cancer [4, 5]. Proteomic biomarker patterns have proved to be more diagnostic than the use of individual biomarkers which have had limited success, and can essentially act as “fingerprints” of a disease [5, 6, 23]. The pattern differences of protein profiles between cancer and healthy samples are developed using data mining algorithms. A successful biomarker discovery program requires high-quality samples to be acquired and processed in a standardized manner due to the high degree of variability associated with protein expression in biological materials. Besides, as these pattern differences originate from the complexity of biological

fluids, which is a mixture of thousands of proteins, protein profiling requires high accuracy and high sensitivity.

### **1.3. MALDI TOF MS**

#### ***Mass Spectrometry***

Mass spectrometry (MS) is a key tool employed by proteomics for the detection, identification, and characterization of proteins [24]. It is a primary analytical platform for many aspects of proteomics, allowing unique structural and functional aspects of proteins to be characterized, and thus MS is becoming more widespread with more sensitive and higher resolution instrumentation [16, 17, 24]. By providing the sizes and relative abundances of the proteins in a complex biological/chemical sample in a rapid and precise manner, MS has the potential to make it possible to do the analysis over a large mass range simultaneously [4].

Mass spectrometry technologies such as TOF (time-of-flight) mass spectrometry analyze proteins based on their mass to charge ratio ( $m/z$ ). The essential components of a TOF mass spectrometer are the ionization source, the mass analyzer and the detector. The ionization source is used to ionize samples; the mass analyzer “analyzes” these ions by separating them based on their mass to charge ratios, the detector records these separated ions.

## **MALDI**

Significant effort has gone into the development of mass spectrometry instruments with greater resolving power and lower detection limits, which would allow the use of smaller samples [27, 28]. One important development in instrumentation is the introduction of an ionization method that can ionize large biological samples “softly” without breaking them into smaller pieces. One of the most widely used new ionization methods is matrix-assisted laser desorption ionization (MALDI), which offers very high levels of sensitivity and mass accuracy for the detection and identification of proteins [15, 16, 17]. The analyzer, in our case, TOF MS is most commonly coupled to a MALDI ion source. MALDI TOF MS has detection sensitivity in the 0.1 to 10 picomole range and charge to mass precision ranging from 0.5% to 1% [18]. Attempts at using MALDI TOF MS to analyze relatively large biomarkers were made as early as 1980s [15,16].

MALDI protein profiling, or its derivative: surface-enhanced laser desorption ionization (SELDI), uses protein mixtures containing an energy absorbing matrix (EAM) as the ion source. Our spectra used a SELDI apparatus, in which a protein mixture is first spotted onto a functionalized surface that binds proteins of interest present in the sample so the other molecules that are not of diagnostic interest can be washed away. After that, an EAM compound is applied to the surface so that it crystallizes with the desired sample peptides contained within it, in which uses a functionalized bind specific analytes from a protein mixture. MALDI and SELDI usually require fractionation of the serum to remove the most

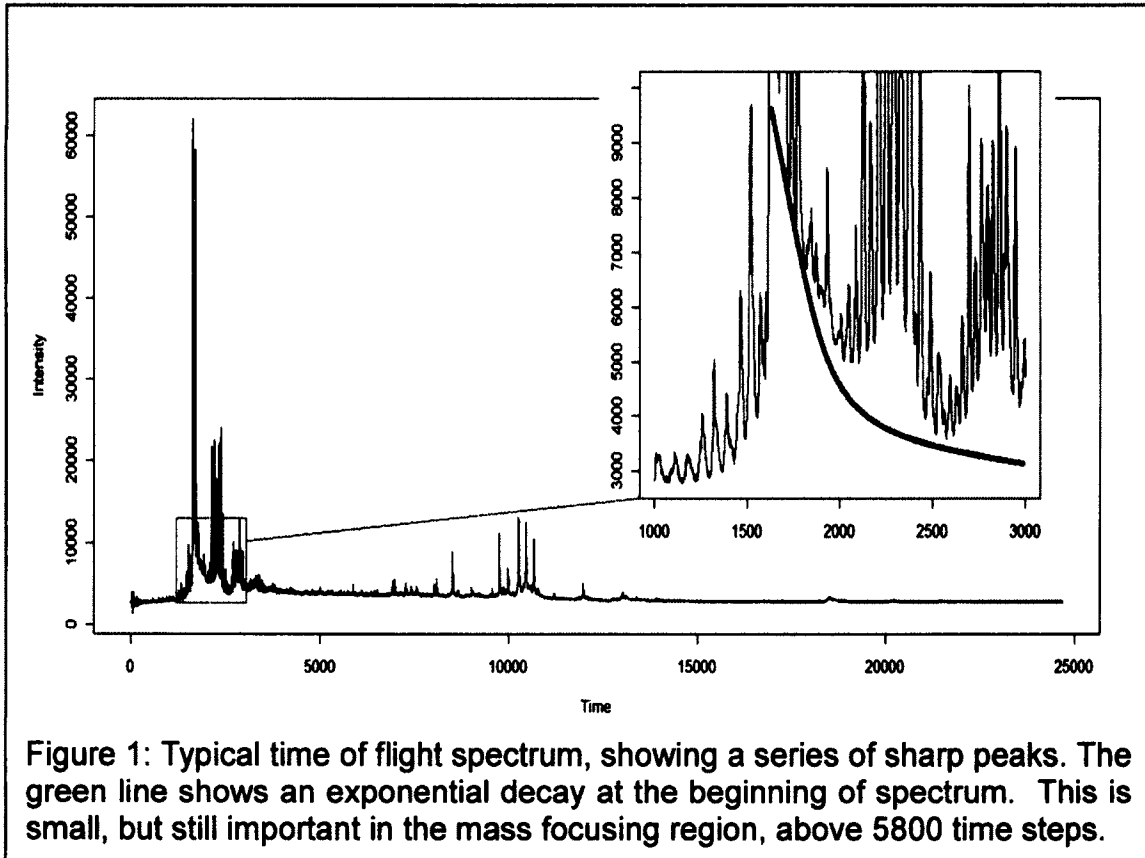


common, and uninteresting components of the serum. SELDI, thereby offering a higher concentration for the proteins of interest can provide a higher sensitivity than regular MALDI [30].

### ***Time-Of-Flight***

The analysis of time-of-flight spectrometry has been given in a number of places, in this brief section; we follow the treatment of Stephens, W.E. [7]. All ions from a sample are extracted from the source and accelerated by an electric field into the mass analyzer. For each ion, the potential energy of the electric field ( $U = zV$ ) is converted into the kinetic energy ( $E = \frac{1}{2}mv^2$ ), as  $zV = \frac{1}{2}mv^2$ . The accelerated ions travel through a field-free region of the analyzer for the same distance  $l$ , for a time,  $t$ , consistent with their velocity, such that,  $v = \frac{l}{t}$ . The mass to charge ratio ( $m/z$ ) of the ions is then a quadratic function of the flight time:  $\frac{m}{z} = \frac{2Vt^2}{l^2}$ . Ions of the same  $m/z$  have the same flight time and thus strike the detector simultaneously. If all ions are assumed to have the same kinetic energy, as a result of traversing the same electrostatic potential in the acceleration region, ions of smaller  $m/z$  travel faster than those of larger  $m/z$ . Therefore, as ions travel through the analyzer, they separate from each other in space, and arrive at the detector as different times. The detector measures the arrival time of the ions.

A plot of abundance versus flight time shows the distribution of the detected ions,



as in Figure 1, the amplitude of the signal corresponds to the total number of ions that struck the detector at a specified time. The flight time could be directly transferred to mass to charge ratio to identify the molecular weight of protein.

$$t = \sqrt{\frac{m l^2}{z 2V}} \quad (1)$$

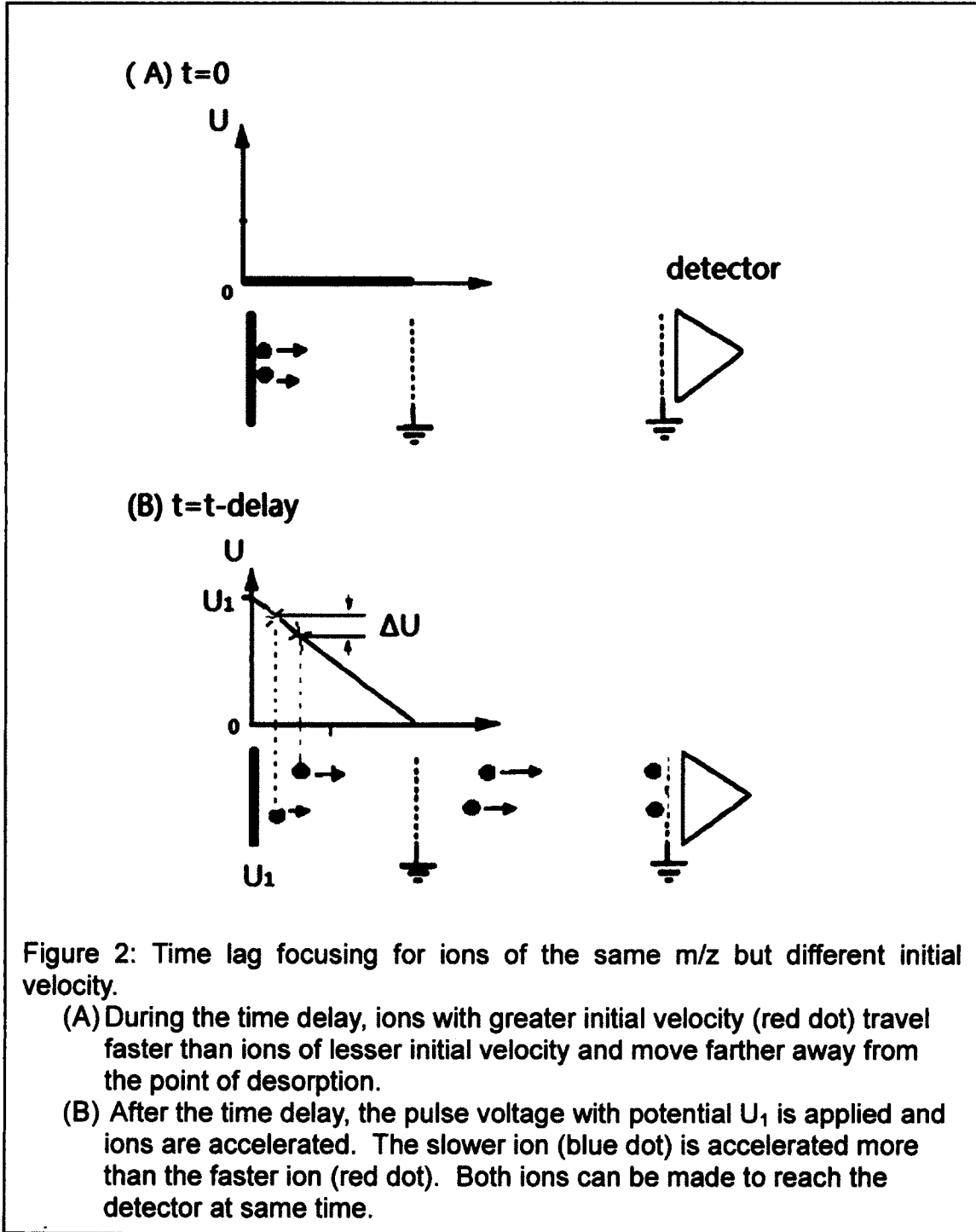
As Figure 1 shows, in the sample spectrum, at any data point in time, the measured signal is not as simple as its average value, but may be larger or smaller. In a word, there is noise fluctuation in the signal. The structure of this noise and its implications for analysis of the SELDI and MALDI spectra have been thoroughly discuss in previous writings which we draw upon below [21].

Random variations implicit in a Poisson counting process and baseline contributions from other peak tails are two major contributions to the noise. A baseline which looks like an exponential decay is usually observed at the beginning of the recording time, the cause of the baseline is believed to be from the detector overloading (saturation) at very early times due to the ionized energy-absorbing matrix molecules, which have small  $m/z$ . Peaks representing a relatively large amount of ion signal have a finite width or a shape. This largely comes from the initial velocity of ions during ionization; finite laser pulse width (4ns) and time jitter of each trial (the resulting mass spectrum is the average from 192 laser shots) also contribute to the shape of the peaks.

To form ions, the dried sample is irradiated with a pulsed laser beam; the energy absorbed by the sample results in a rapid heating and expansion. This produces collisional ionization of the proteins, usually by adding a hydrogen ion (proton) from the acidic matrix. But, this also results in an initial velocity distribution of ions so that ions of the same  $m/z$  do not all have the same initial velocity or the same initial kinetic energy. The initial kinetic energy spread then adds to the electrostatic energy to produce a spread of final kinetic energies, meaning a spread of final velocities, and thus final arrival times.

The initial velocity distribution of ions with same  $m/z$  can be compensated by delaying the acceleration voltage until a few hundreds of nanoseconds (900 ns in our case) after the ionization pulse. Then, because the ions with high kinetic energy travel farther before the acceleration voltage is turned on, those ions will

then have less potential energy. Conversely, ions that are initially slow, will travel less distance, and therefore have a greater electric potential energy when the voltage turns on. The net result is that all of the ions in a chosen mass range



end up with the same total energy, and therefore arrive at the same time at the detector. This is called time lag focusing; Figure 2 is a simple illustration of this process.

It shows the slower ion as a blue dot, and the faster ion as a red dot. When the pulse voltage with potential  $U_1$  is applied, the slower ion (blue dot) is accelerated over a longer distance than the faster ion (red dot), so both ions can be made to reach the detector at same time.

However, this is a mass dependent initial velocity compensation method. So, for any given delay pulse time, there is an optimum focused  $m/z$  range, referred to as mass focusing range, for which the peaks will be their narrowest.

As we will explain later, the peak widths remain nearly constant over a wider range of times if the spectra are plotted as a function of time, rather than  $m/z$ . A single mass spectrum usually contains many thousands, or even millions, of data points; however, information related to the biological samples is encoded in only a few hundred peaks. The standard approach to summarizing TOF spectra is to generate a list of peaks representing proteins or protein fragments that are more plentiful than the background ion signal.

## **1.4. Experimental data**

One of the first steps in the process of biomarker discovery is the collection of the biological samples that will provide the data. The data sets discussed in this thesis originated as blood samples taken from patients diagnosed with or without

a specific disease. These data were created by Semmes, and his colleagues and coworkers at the Eastern Virginia Medical School (EVMS) in Norfolk, Virginia. Special care was taken by that group to randomize the processing and to make the creation of the sample data as unbiased as possible [26, 27].

The working data set contains sample serum from 145 different patients, of which 78 were classified during the clinical portion as "normal," and 67 with various stages or forms of leukemia. The samples from the patients were processed 3 to 4 times, resulting in 425 cases for the study. Multiple TOF spectra from the same sample are called replicates.

Also, Quality Control (QC) data are taken from a single serum pool which is a mixture of a large group of nominally healthy people. Thus the QC samples are nominally identical; any variations noted must arise from the data creation process of preparation and MS measurement. We can use QC data to calibrate the measurements.

Serum samples were processed for SELDI analysis. Metal affinities (IMAC3-Cu) ProteinChip® were used during the experiment to utilize IMAC (immobilized metal ion affinity chromatography) to enrich the phosphorylated peptides in a digested protein sample. Each chip has eight sample sites that allow the processing of biological samples directly on the chip surface. Thus Serum sample complexity is reduced "on-chip" in a convenient format before laser desorption. The protein chip arrays were analyzed using the SELDI ProteinChip System (PBS-II; Ciphergen Biosystems, Fremont, CA). The protein masses

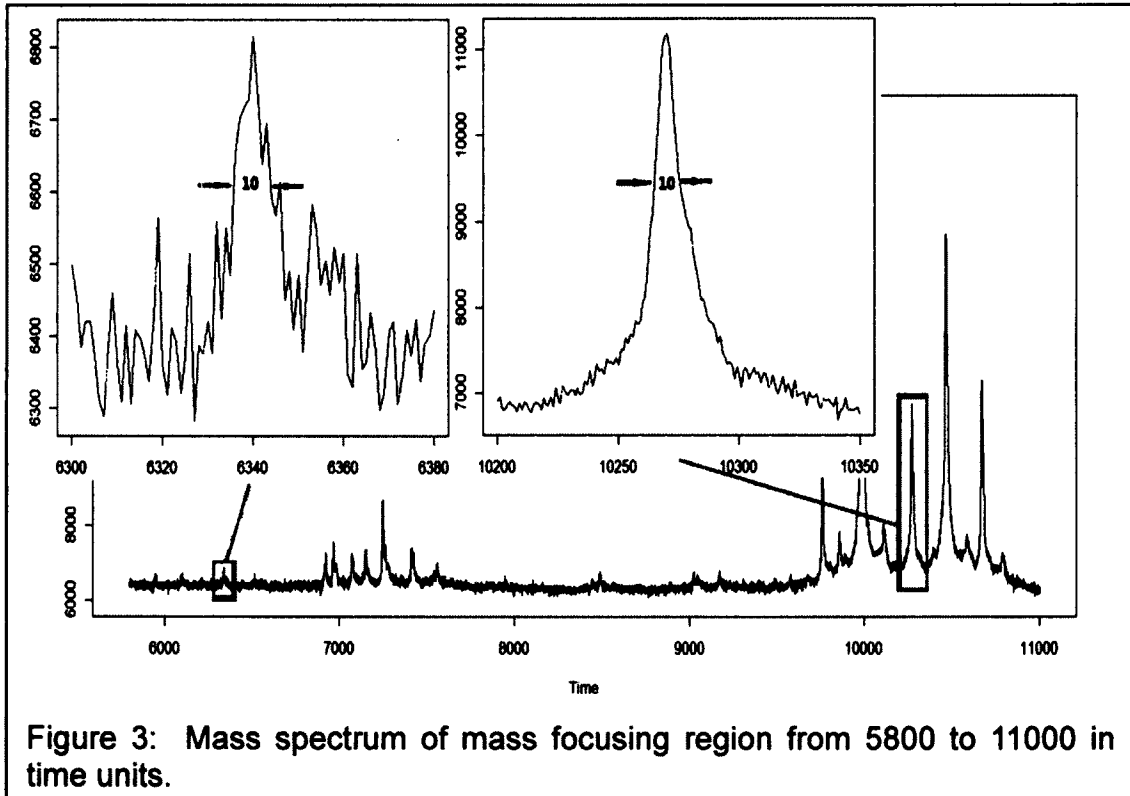
were calibrated externally using purified peptide standards (Ciphergen Biosystems, Inc.) Instrument settings were optimized using a pooled serum standard [26,27].

A nitrogen laser operating at 337 nm with a 4ns pulse width irradiates the sample, and is absorbed by the matrix crystals. Focus lag time is 900 ns. Each recorded mass spectrum was the sum of 192 laser shots taken from a rastered pattern on each sample spot. Analog (rather than counting) detectors are used here. Detector Voltage was set as 1375 Volts. Detection range is optimized from 3000 Daltons to 50000 Daltons. With the detection range optimized as this, we found that the peak width (FWHM) was nearly constant in time over a small  $m/z$  range, *e.g.*, 10 FWHM from 2700 to 10000 at  $m/z$ .

For each replicate of a sample, a spectrum was produced and associated with metadata, including the patient ID, date of collection of the sample, and the settings of SELDI TOF MS. Our data set contains 225 normal spectra and 196 Leukemia spectra.

To “pick” peaks from the raw spectra, we restrict the data analysis to the mass focusing region. The spectrum before this is dominated by the large signal due to the matrix, which has little useful biological information. Figure 3 shows a typical sample spectrum with mass focusing range from 5800 to 11000. Region 6300-6380 and 10200-10400 are expanded in inserts to show the constant peak width across the mass focusing range. In both regions, a peak is sitting on noisy background. A peak in the region of 6300-6380 time points, which has an

amplitude of 400 (arbitrary units) above the background, has the same peak width as peak with 4000 intensity in the region 10200-10400. It shows that the approximation of fixed peak shape is a reasonable assumption.



## 1.5. Data Processing Overview

This thesis presents a new method for the analysis of TOF-MS data. The purpose is to tease out key information from raw data and increase the likelihood of successful disease classification. Three major steps are discussed: first is the pre-processing of MS data, baseline subtraction is the only pre-processing step applied in this thesis; second, peaks which are interpretable features corresponding to distinct protein species are picked from each spectrum. The



first two steps are done independently for each spectrum. The third step is to identify peaks from the same group of spectra that are sufficiently close to an expected time, that we can claim they represent the same peak. This step is called peak alignment. After these steps, MS data is ready for biomarker candidates' selection and identification. Once we find a small set of peaks that can be used to computationally "predict" phenotypes (that is, the disease state of the patient) with high accuracy, these peaks will become the basis of new experiments that will then identify the underlying proteins.

### ***Pre-processing***

The primary difficulty in processing MS is to separate the true peak signal from the baseline and the noise. Traditional peak picking methods usually focus on the disentangling of these three pieces. A lot of pre-processing methods are applied to simplify the mass spectrum data before picking peaks, *e.g.*, model based baseline reduction, noise filtering, and intensity normalization. Most of these approaches take the pre-processing as an independent step from the peak picking part.

The only pre-processing step we use is the baseline subtraction. We treat the baseline as a smooth curve underlying the remaining spectrum after features of interest have been removed. Other people, *e.g.*, Dijkstra *et al.* [30], set up a baseline model with adjustable parameters to model the baseline.

### ***Peak Picking***

We use a model-based criterion that proposes model functions to fit peaks. Other method is to move the peak finding problem to the wavelet coefficient space; then to find high coefficients that cluster together on a similar position for different scales and that thus corresponds to peaks [31,32].

### ***Alignment***

After peak picking, we perform alignment to identify the same peaks in different spectra, even if they are slightly mis-aligned. First we allow spectra to shift a few time-steps to the left or to the right to correct the starting time shift, and then we get a master peak list summarized from peaks which correspond to the same feature based on the shifted spectra. Wong *et al.* [33] developed a strategy that aligns spectra using selected local features (peaks from the average spectrum) as anchors, spectra are locally shifted by inserting or deleting some points on the  $m/z$  axis. Antoniadis *et al.* [32] proposed an alignment based on landmarks in the wavelet framework. More recently, Kong *et al.* [34] proposed a Bayesian approach that uses the expectation–maximization algorithm to find the posterior mode of the set of alignment functions and the mean spectrum for a population of patient, and Feng *et al.* [35] modeling the  $m/z$  by an integrated Markov chain shifting (IMS) method.

We will go through MS data processing steps in great detail in the following chapters and we will show that our method is very useful in narrowing the search for protein biomarker candidates.

## **1.6. Dissertation Outline**

This dissertation is divided into five chapters; Chapter 1 gives background information about how this dissertation is related to biomarker research and a brief introduction of mass spectrum data from MALDI. Chapter 2 explains the peak picking method, including algorithms to eliminate background artifacts based on a signal model and a local noise calculation. Chapter 3 focuses on the analysis of real data, as opposed to model data, and mentioned a local background elimination method, which models a local background by using the set of spectra where there were no peaks in that region. Chapter 4 describes how to align mass peaks from the multiple spectra present in a dataset using non-parametric inferential statistical methods. Chapter 5 summarizes our conclusions.

Overall, we have successfully created an automatic algorithm for optimizing the data processing; this transforms the raw data into a more useful representation, including only confirmed peaks, which will lead to an improved understanding of the biological information contained in the data.

# Chapter 2 Peak Detection Using A

## Maximum Likelihood Filter

A mass spectrum provides rapid and precise measurements of the mass-to-charge ratios and relative abundances of the proteins present in a complex biological/chemical mixture. Mass spectrometry (MS) signals from a time of flight (TOF) measurement are the sums of ions arriving at a specified time, and the time series output usually contain thousands or even millions of data points. However, only a few hundred peaks contain the biologically important information. The process of reducing an entire TOF spectrum to a small list of peaks with their amplitudes and related uncertainties is called peak picking.

In this chapter, we will begin with an overview of the method of fitting spectra with a maximum likelihood approach, and then discuss the specific details of the method, such as an automatic noise level determination, automatic signal to noise threshold, demonstrating the use this technique on simulated data.

### 2.1. Overview of the Method

Our peak identification algorithm locates a peak by finding those times where a data segment has the highest likelihood of describing a peak centered in that segment. This combines the accuracy of parameter fitting with the speed of local

filtering. The algorithm automatically gives precise estimates of the peak position and intensity with no supervision required. The major goals of our research have been to provide a reasonable set of conditions that simultaneously maximizes the number of true detected peaks while minimizing the number of false detected peaks. A false detected peak occurs when noise fluctuations in a series of nearby points happen to mimic the shape of a peak. Eventually, we will separate the true peaks from the false peaks by requiring that true peaks must occur at the same location in several spectra from the same data set, but the details of this will be discussed in Chapter 4.

There are two steps in the strategy of our peak picker algorithm. First, we calculate the likelihood of a data segment having a peak centered in it as the data segment, or window, slides along the entire region of interest. We use data windows as wide as the Full Width at Half Maximum (FWHM) of the expected peaks for this calculation. That calculation also estimates the amplitude of the peak simultaneously. Next, we use the peak amplitude estimation to calculate a signal to noise ratio (*SNR*) for each segment to identify the possible peak regions. This is necessary because the likelihood will be high in data windows that are well described as having a peak of nearly zero amplitude. Finally, we determine where the likelihood reaches a maximum value within each region of interest in order to identify the locations and amplitudes of the relevant peaks. This process works best when we have eliminated any slowly varying

background from the data, so we will also introduce several iterative methods for eliminating the background.

### ***Mass Spectrum Data***

In a TOF spectrum, it is usually assumed that the ions fly to the detector without any ion-ion interaction after they are created. Thus, each individual time measurement is independent of measurements at adjacent times. The mean number of ions that arrive at each time may be related to the mean number of the ions that arrive at a nearby time by a line shape function, but the individual measurements are independent. This is similar to rolling two dice, one with the sides labeled 1-6 and the other labeled with only the even numbers from 2-12. The average value of the second die will be twice as large as the average value of the first die, but on any two rolls, the actual values will be uncorrelated. So in a mass spectrum, the average values from one data point to another data point are related but their measured values are not, even within a mass peak.

A mass spectrum signal consists of three major components: peaks, background and noise. A peak is the relative abundance of ions of a specific charge to mass ratio originating from a specific protein in the sample, the background is a relatively feature-less signal caused by other sources unrelated to the protein that creates the peak plus a constant offset of the analog to digital converter (ADC). The noise primarily comes from statistical fluctuations in the ion count and variations in the gain of the detector. Electrical pickup noise and the  $\pm\frac{1}{2}$  bit

digital conversion noise are usually small. These three parts lead to our model of the mass spectrum signal in the vicinity of a mass peak:

$$S_i = Ax(t_i) + B + ADC_{\text{offset}} + C\eta_i \quad (2)$$

where  $Ax(t_i)$  represents the peak, with  $A$  as the amplitude and  $x(t_i)$  as the expected peak line shape,  $B$  is the background due to the other ions,  $ADC_{\text{offset}}$  is the constant ADC offset,  $C$  is the amplitude of the noise, and  $\eta_i$  is a random variable with zero mean and unit standard deviation. The shape of most peaks is not well characterized in the wings, so we will restrict this model to the upper half of any peak, and use a Gaussian shape with a fixed width to describe  $x(t_i)$ . The constant width is a good approximation over the mass focusing range. We will assume that  $\eta_i$  follows a Gaussian distribution. We will discuss more details of the model later in this chapter, in the section Fitting Data to The Model.

As discussed in chapter 1, the success of our method largely depends on a constant peak width within the spectrum. The full width at the peaks' 50% maximum height (FWHM) is approximately constant in the mass focusing range from 5800 time points to 11000 time points, which means  $m/z$  between 2700 to 9900 Daltons. The FWHM of the peaks in this region was 10 time points. Outside the mass focusing region, the width grows slowly with mass. For example, as shown in Figure 4, the FWHM of a peak located at 7559 time steps is 10 time steps while FWHM of peak located around 18500 time steps is 10 times larger or approximately 100 time steps. We could segment the spectrum

into several separate regions with different FWHM values to extend our methods throughout the entire spectrum, but that is not discussed in this dissertation.

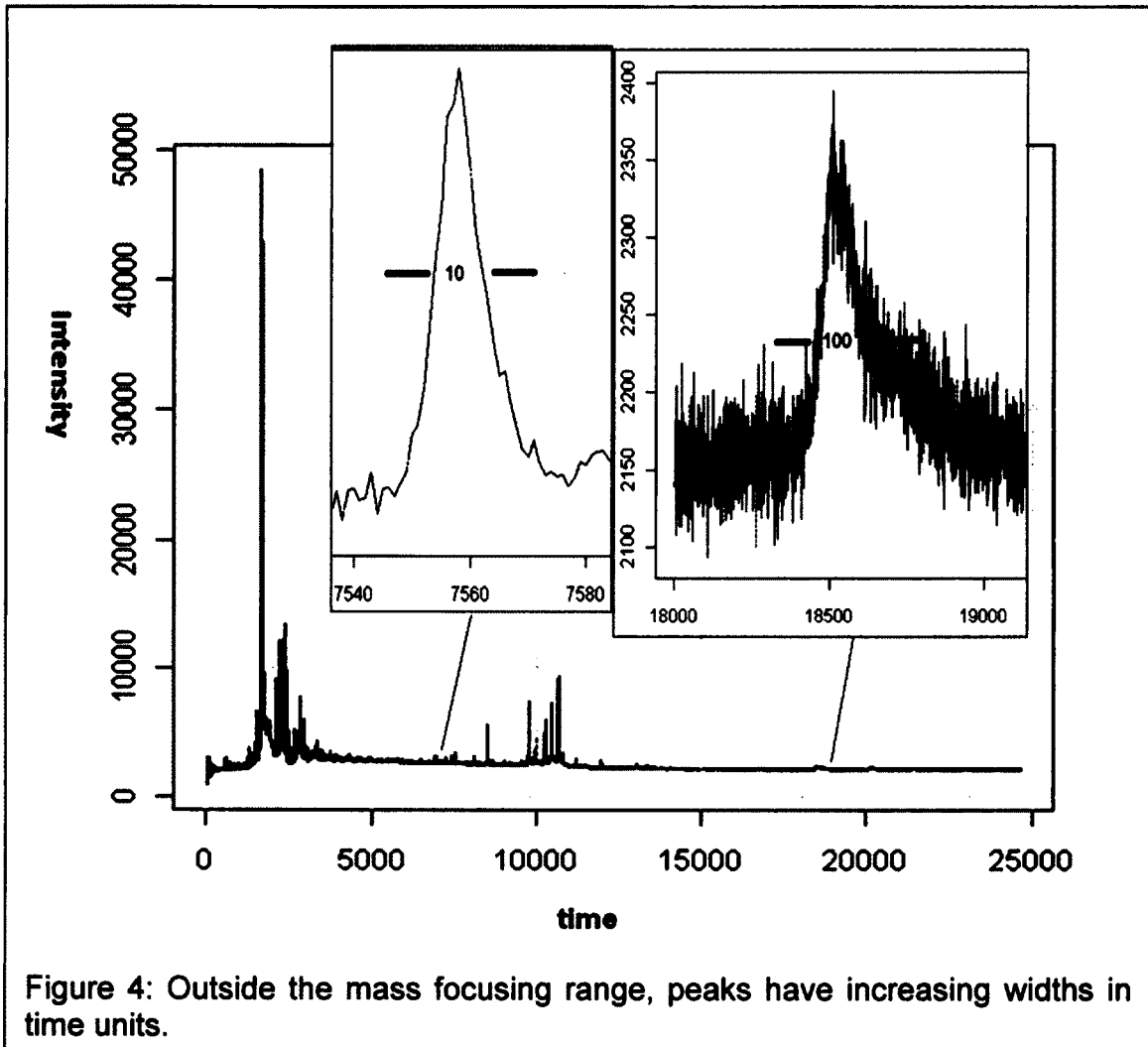


Figure 4: Outside the mass focusing range, peaks have increasing widths in time units.

### 2.1.1. Fitting Data to The Model Line Shape

The standard approach to summarizing TOF spectra is to generate a list of peaks representing proteins or protein fragments that are more plentiful than the background ion signal. i.e., a mass spectrum should be reducible to only few hundreds of features each including peak position, peak amplitude and the uncertainties of both parameters. The peak position is related to the charge to



mass ratio, while the peak amplitude indicates the amount of that protein present in the sample.

### ***Data in A Sliding Window***

We analyze the data in small segments, called windows, which are as wide as the FWHM of a typical peak. For each window, we calculate the likelihood of that window describing the data as being background plus a peak centered in the window with an amplitude of the peak that would best describe the data. We determine that a peak is, in fact, present when the likelihood is at a local maximum, *and* the maximum likelihood amplitude estimate of that peak is sufficiently large. This second condition is critical, because a region with no peak present can always be described (with high likelihood) as a region with a zero amplitude peak centered in the window. By sliding this window along the entire spectrum, we can efficiently find a large number of peaks, even without any foreknowledge of where they should occur.

We chose a window width as the FWHM to balance two effects. If the window is chosen to be too large, then often the window could have two peaks within it, and this would require a different model for an accurate likelihood calculation. Moreover, if the window were large compared to the FWHM of the peak, the likelihood calculation would be sensitive to the shape of the peak in its wings. The wings of the MALDI-MS peaks are not well characterized. If we chose the window width to be too small, then there would not be sufficient shape in the

window to identify a peak or to separate the peak amplitude from the featureless background.

### ***Signal Model***

Our model is that there is a single peak centered in the window, along with a background level. So, if one uses  $N$  adjacent data points  $(s_1, s_2 \dots s_N)$  from the time series, and a  $N$ -component vector to describe the peak line shape  $(x_1, x_2 \dots x_N)$ , then the observed data within the window should vary as :

$$S_i = Ax_i + \sqrt{\varepsilon(\mu_1 + Ax_i)}\eta_i + \mu, \quad i=1, 2 \dots N \quad (3)$$

Here  $A$  is the amplitude of any peak within the window,  $\sqrt{\varepsilon(\mu_1 + Ax_i)}\eta_i$  represents the random fluctuation in counts that are expected from a Poisson distribution with an average value of  $(\mu_1 + Ax_i)$ , and  $\mu$  is an unknown background, due to the wings of other peaks or to the large background of matrix ion clusters (the matrix is the energy absorbing material that holds and ionizes the biological proteins, described in chapter 1). Because the chances of cluster formation decrease with cluster size, the baseline diminishes monotonically as the mass-to-charge ratio increases. The slowly decaying background is large compared to the window size, so the background is almost constant within the window. The vector of random values  $\eta = (\eta_1, \eta_2 \dots \eta_N)$  is assumed to be drawn from a Gaussian distribution with zero mean and unit variance, because a Poisson distribution with

$N > 10$  ions detected is very close to a Gaussian with a width of  $\sqrt{N}$ . The factor  $\varepsilon$  is a normalization factor to convert ion counts into a measured voltage.

### **Likelihood**

The likelihood is defined as the probability of the data occurring given a model with a particular choice of parameters. We will maximize the log-likelihood for each choice of window position to determine the best fit parameters, and then choose the best window position to maximize that log-likelihood in regions where the high signal to noise ratio implies that there is a peak.

Under the assumption that the ion counts at different times are independent, the probability of observing the particular count sequence  $s = (s_1, s_2, \dots, s_N)$  is simply

$$P_N(\eta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_\eta} \exp\left(-\frac{(s_i - Ax_i - \mu)^2}{2\sigma_\eta^2}\right), \text{ where } \sigma_\eta = \sqrt{\varepsilon Ax_i} \quad (4)$$

The log-likelihood function is the natural logarithm of the above equation:

$$L^P = -\log\left(\prod_{i=1}^N \sqrt{2\pi}\sigma_\eta\right) - \sum_{i=1}^N \frac{(s_i - Ax_i - \mu)^2}{2\sigma_\eta^2}, \text{ where } \sigma_\eta = \sqrt{\varepsilon Ax_i} \quad (5)$$

That is,

$$\begin{aligned} L^P &= -\frac{1}{2} \sum_{i=1}^N \log(2\pi\varepsilon Ax_i) - \sum_{i=1}^N \frac{(s_i - Ax_i - \mu)^2}{2\varepsilon Ax_i} \\ &= -\frac{N}{2} \log(2\pi\varepsilon A) - \frac{1}{2} \sum_{i=1}^N \log(x_i) - \frac{1}{2\varepsilon} \sum_{i=1}^N \frac{(s_i - \mu)^2 - 2Ax_i(s_i - \mu) + A^2x_i^2}{Ax_i} \\ &= -\frac{N}{2} \log(2\pi\varepsilon A) - \frac{1}{2} \sum_{i=1}^N \log(x_i) - \frac{1}{2\varepsilon} \sum_{i=1}^N \left( \frac{(s_i - \mu)^2}{Ax_i} - 2(s_i - \mu) + Ax_i \right) \end{aligned} \quad (6)$$

To simplify the calculation, let:

$$\begin{aligned}\sum_{i=1}^N x_i &= N\bar{x} \\ \sum_{i=1}^N \frac{1}{x_i} &= \sum_{i=1}^N z_i = N\bar{z}\end{aligned}\tag{7}$$

So the likelihood function becomes:

$$\begin{aligned}L^P &= -\frac{N}{2} \ln(2\pi\varepsilon A) - \frac{N}{2} \ln(\overline{x_i}) - \frac{1}{2\varepsilon} \sum_{i=1}^N \left( \frac{(s_i - \mu)^2}{Ax_i} - 2(s_i - \mu) + Ax_i \right) \\ &= -\frac{N}{2} \left( \ln(2\pi\varepsilon A) + \frac{1}{\varepsilon} \left( \frac{1}{A} (\overline{zs^2} - 2\mu\overline{sz} + \bar{z}\mu^2) - 2(\bar{s} - \mu) + A\bar{x} \right) + \ln(\overline{x_i}) \right)\end{aligned}\tag{8}$$

The maximum likelihood estimation maximizes the probability of the observed outcome to get the unknown parameter values. Here the outcome means the signals within a window. The curvature is quantified by the second derivative, that is, the change in slope. In particular, if the likelihood is differentiable then its partial derivatives are zero when evaluated at any local extreme value. These points correspond to solution roots of the following conditions ( $t_0$  indicates the window location).

$$\begin{aligned}\frac{\partial L^P(A, \mu, \varepsilon, t_0)}{\partial A} &= 0, \\ \frac{\partial L^P(A, \mu, \varepsilon, t_0)}{\partial \mu} &= 0, \\ \frac{\partial L^P(A, \mu, \varepsilon, t_0)}{\partial \varepsilon} &= 0\end{aligned}\tag{9}$$

Solving the equations gives the maximum likelihood estimate of unknown parameters  $(A^*, \mu^*, \varepsilon^*)$  for a window located at  $t_0$  and for data in the window. The likelihood is sharply peak around the point  $(A^*, \mu^*, \varepsilon^*)$  in the parameter space. The second derivative is negative, as the slope changes from being positive to negative and the larger its absolute value the more sharply curved the likelihood is at its maximum. Intuitively, a sharply curved likelihood is desirable because this narrows the range over which the likelihood is close to its maximum value, that is, it narrows the range of plausible parameter values.

This method can also be considered another expression of least square estimations. In the most general terms, least squares estimation is aimed at minimizing the sum of squared deviations of the observed values for the dependent variable from those predicted by the model, which is equivalent to maximizing the probability of the observed values given the model. Essentially, the method of maximum likelihood chooses as estimates those values of the parameters that are most consistent with the sample data. To solve these equations, we set the partial derivatives with respect to  $\varepsilon, A,$  and  $\mu$  equal to zero:

$$\begin{aligned}
 \frac{\partial L^P}{\partial A} &= \frac{\bar{x}A^2 + \varepsilon A - s^2\bar{z} - \mu^2\bar{z} + 2\mu\bar{s}z}{\varepsilon A^2} = 0 \\
 \frac{\partial L^P}{\partial \mu} &= \frac{(A + \mu\bar{z} - s\bar{z})}{\varepsilon A} = 0 \\
 \frac{\partial L^P}{\partial \varepsilon} &= \frac{\bar{x}A^2 - \varepsilon A + 2A(\mu - s) + s^2\bar{z} + \mu^2\bar{z} - 2\mu\bar{s}z}{\varepsilon^2 A} = 0
 \end{aligned} \tag{10}$$

Leading to:

$$\begin{aligned}
A\varepsilon + A^2\bar{x} &= \left( z(s - \mu)^2 \right) \\
\mu &= \frac{\bar{sz} - A}{\bar{z}} \\
\bar{x}A^2 - \varepsilon A + 2A(\mu - \bar{s}) + \left( z(s - \mu)^2 \right) &= 0
\end{aligned} \tag{11}$$

They have the following solutions:

$$A = \frac{\bar{zs} - \bar{sz}}{\bar{x}\bar{z} - 1} \tag{12}$$

$$\mu = \frac{\bar{x}\bar{sz} - \bar{s}}{\bar{x}\bar{z} - 1} \tag{13}$$

$$\varepsilon = \frac{\bar{szx} - \frac{\bar{s}^2\bar{z}}{\bar{sz}}(\bar{xz} - 1)}{\bar{sz}} \tag{14}$$

The likelihood is then

$$\begin{aligned}
e^{L^f} &= \left( \prod_{i=1}^N x_i \right)^{-1} \frac{e}{(A\varepsilon)^N} \\
&= \frac{e}{\left( \prod_{i=1}^N x_i \right) \left( \frac{\bar{s}^2\bar{z} - \frac{-2\bar{s}}{\bar{x}\bar{z} - 1}}{\bar{x}\bar{z} - 1} \right)^N} \\
&= e \left( \prod_{i=1}^N \frac{\bar{x}}{x_i} \right) \left( \frac{\bar{x}\bar{z} - 1}{\bar{s}^2\bar{z}\bar{x}(\bar{x}\bar{z} - 1) - \frac{-2\bar{s}}{\bar{x}\bar{z} - 1}} \right)^N
\end{aligned} \tag{15}$$

Notice that solving the above equation only maximizes the likelihood with respect to parameters  $(A, \mu, \varepsilon)$  at the same window location  $t_0$ , i.e. for a fixed data set we obtain the parameters  $(A^*, \mu^*, \varepsilon^*)$ . Maximizing the likelihood with respect to  $t_0$  is done by computing the likelihood for each window position at  $(A^*, \mu^*, \varepsilon^*)$  and then

finding the  $t_0$  that maximizes it. However, maximizing over  $t_0$  has a different logic than the other parameters, because we are comparing different data sets as we slide the window across the peak. The justification is based upon the physical reasonableness. Because the width of the window is large compared to the final uncertainty in the position of the peak, most of the data comes from overlapping windows so that the data is nearly the same. An alternative way of looking this is that by comparing likelihood at different  $t_0$ , we are actually looking for a window in which the data best support the assumption that there is a peak in the window.

In addition to the likelihood estimation of parameters  $(A^*, \mu^*, \varepsilon^*)$ , this method automatically generates the second derivatives of  $L$  to produce estimates of the uncertainties of our parameters from the Hessian matrix at  $(A^*, \mu^*, \varepsilon^*)$ .

$$\nabla\nabla L(A^*, \mu^*, \varepsilon^*, t_0) = \begin{pmatrix} L_{AA} & L_{A\mu} & L_{A\varepsilon} \\ L_{\mu A} & L_{\mu\mu} & L_{\mu\varepsilon} \\ L_{\varepsilon A} & L_{\varepsilon\mu} & L_{\varepsilon\varepsilon} \end{pmatrix} \quad (16)$$

The result is:

$$\nabla\nabla L(A^*, \mu^*, \varepsilon^*, t_0) = \begin{pmatrix} \frac{N(\sigma_x^2 + \bar{x}^2)}{(\varepsilon^*)^2} & \frac{\sum_{i=1}^N x_i}{(\varepsilon^*)^2} & 0 \\ \frac{\sum_{i=1}^N x_i}{(\varepsilon^*)^2} & \frac{N}{(\varepsilon^*)^2} & 0 \\ 0 & 0 & \frac{2N}{(\varepsilon^*)^2} \end{pmatrix} \quad (17)$$

Where  $\bar{x}$  is the arithmetic mean of  $x$ ,  $\sigma_x^2$  is the variance.

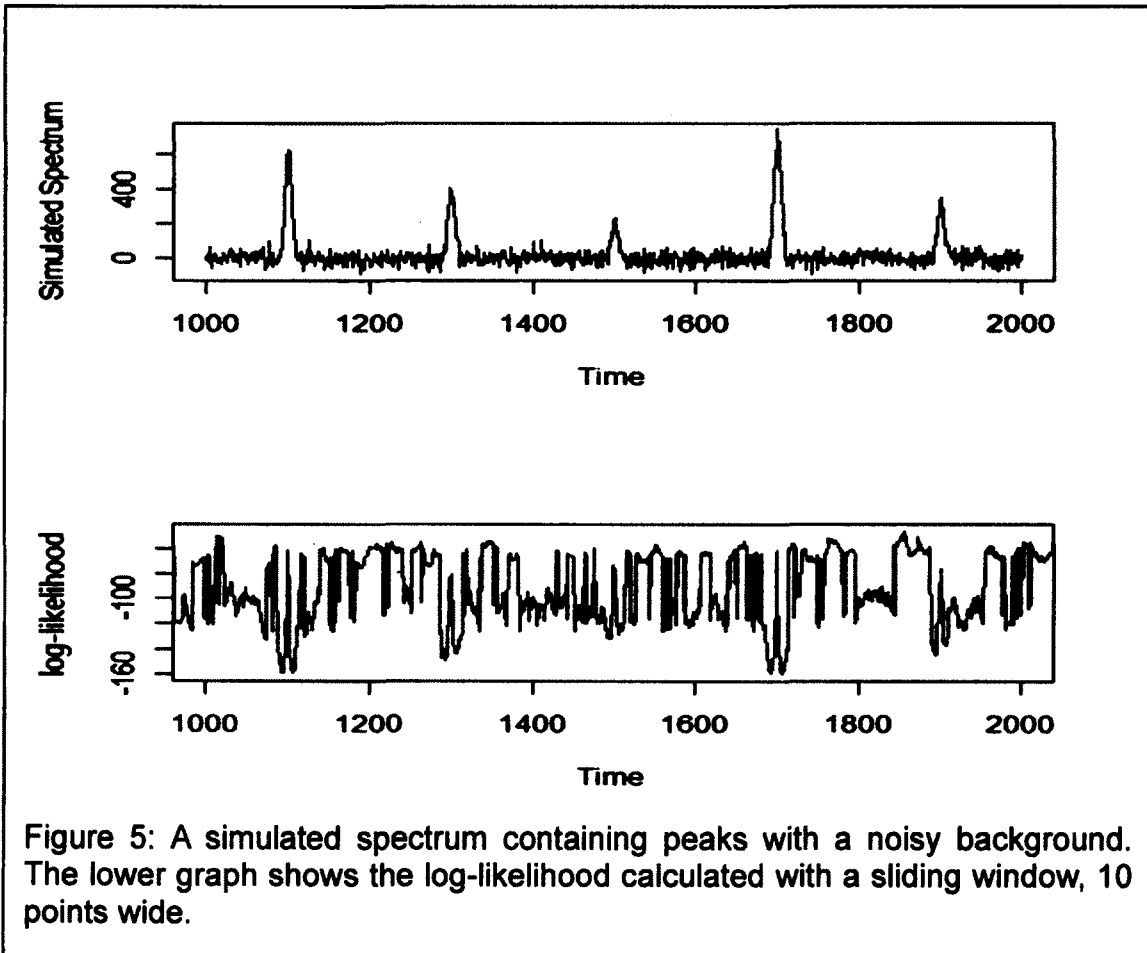
Thus the uncertainties about the best estimation of parameters are:

$$\Delta A = \frac{\varepsilon^*}{\sqrt{N(\sigma_x^2 + \bar{x}^2)}} \quad (18)$$

$$\Delta \mu = \frac{\varepsilon^*}{\sqrt{N}} \quad (19)$$

$$\Delta \varepsilon = \frac{\varepsilon^*}{\sqrt{2N}} \quad (20)$$

Figure 5 shows the calculated log-likelihood value for a simulated spectrum





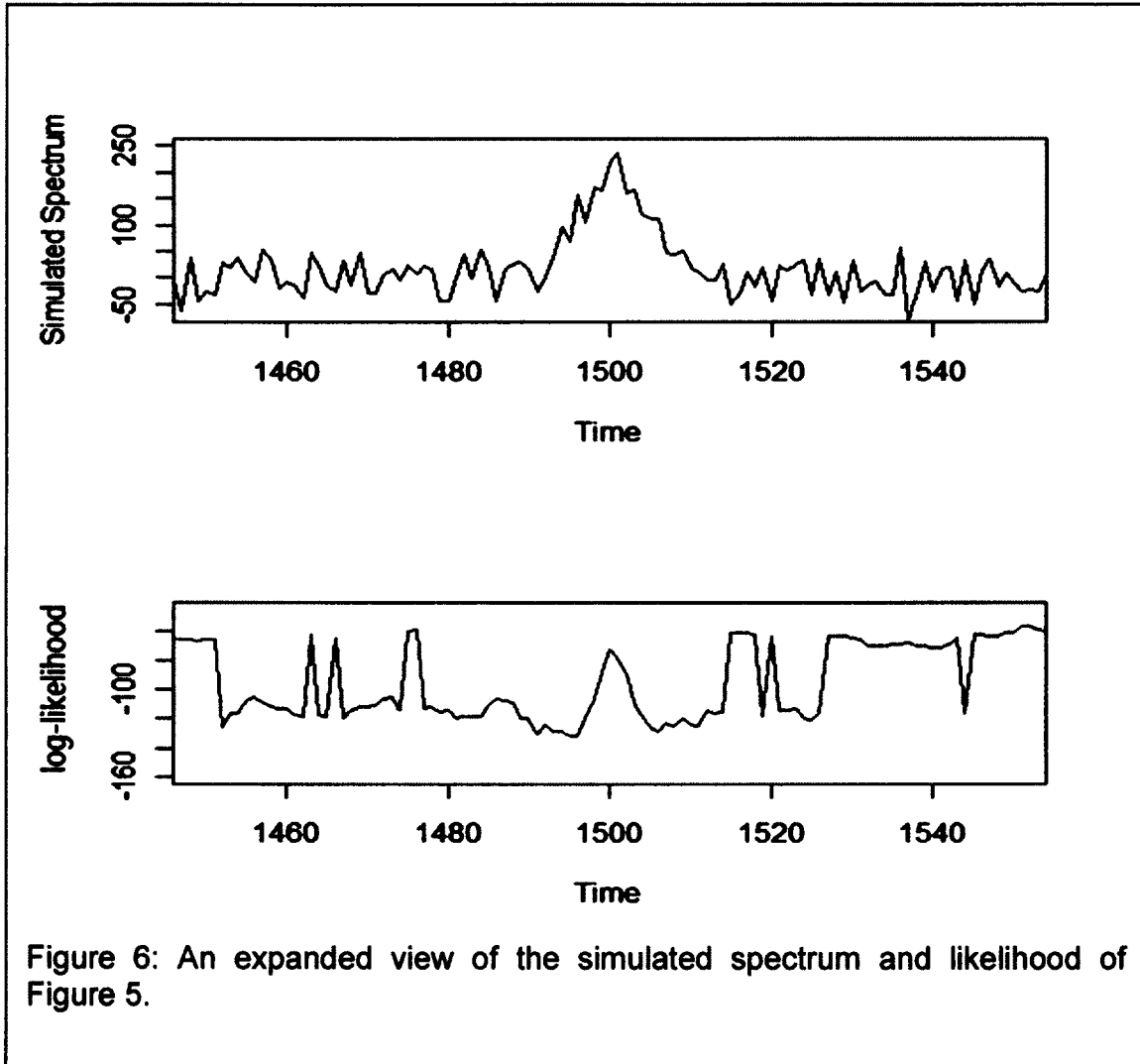


Figure 6: An expanded view of the simulated spectrum and likelihood of Figure 5.

containing 1000 time points and including 5 peaks with varying intensities. We generated the peaks using a Gaussian line shape with  $\text{FWHM}=10$  while generating the background with normally random values having a standard deviation of 30. These simulation parameters were chosen to mimic typical data.

Figure 6 is an expanded view of the region from 1450 to 1550 in Figure 5. As shown in both, the likelihood of the data fitting the model may also be high even if there is no peak in the window as the data fits a model with a zero amplitude

peak centered in the window. We need an additional step to eliminate these regions of zero amplitude peaks.

### **2.1.2. Determining Regions Likely to Contain Peaks**

Windows that are likely to contain a peak will be those where the maximum likelihood estimate of the peak amplitude is larger than an estimate of the local noise. Consequently, we will estimate the local signal to noise ratio and then introduce a threshold to identify those regions. In the following sections, we will describe how we estimate the local noise, and how we choose the appropriate threshold value to identify windows likely to contain a peak as those where the signal to noise ratio exceeds that threshold. However, in the complicated spectra considered here, we have found that a single threshold is insufficient. Accordingly, we will describe a method for iteratively resetting a local threshold in those regions where the peaks are most dense. The changing signal to noise ratio threshold will be discussed in great detail later in chapter 3.

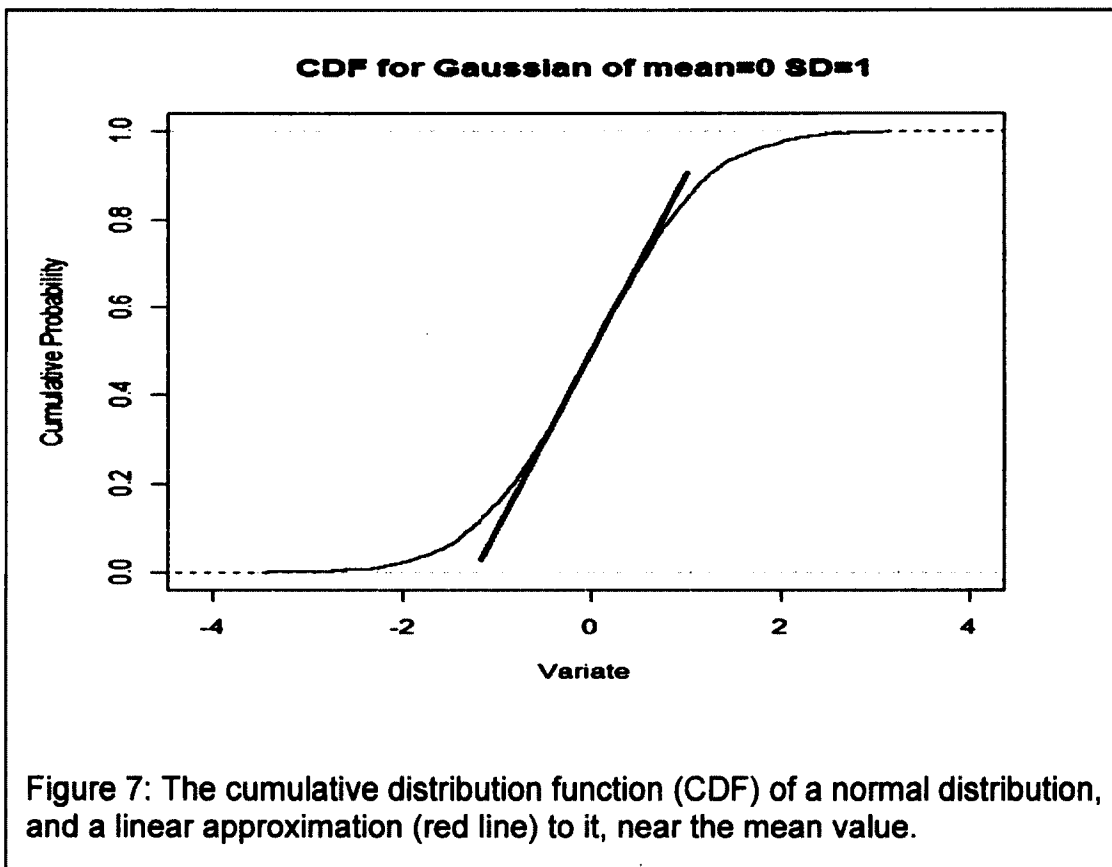
#### ***Automatic Noise Estimation***

Many factors can contribute to the noise, such as instrumentation noise from the physical and electrical components and noise from the statistical fluctuations of the discrete ion counts. Here we focus on the stationary component of the noise, which is nearly the same across the whole spectrum, and use this component in the signal to noise ratio (*SNR*) calculation.

For a normal distribution of stationary noise, the cumulative density function (CDF) is:

$$CDF : \frac{1}{2} \left( 1 + \operatorname{erf} \left( \frac{x - \mu}{\sqrt{2}\sigma} \right) \right) \quad (21)$$

where the error function is the probability of a measurement resulting in a positive value between zero and  $x$ , when the population mean is zero and the



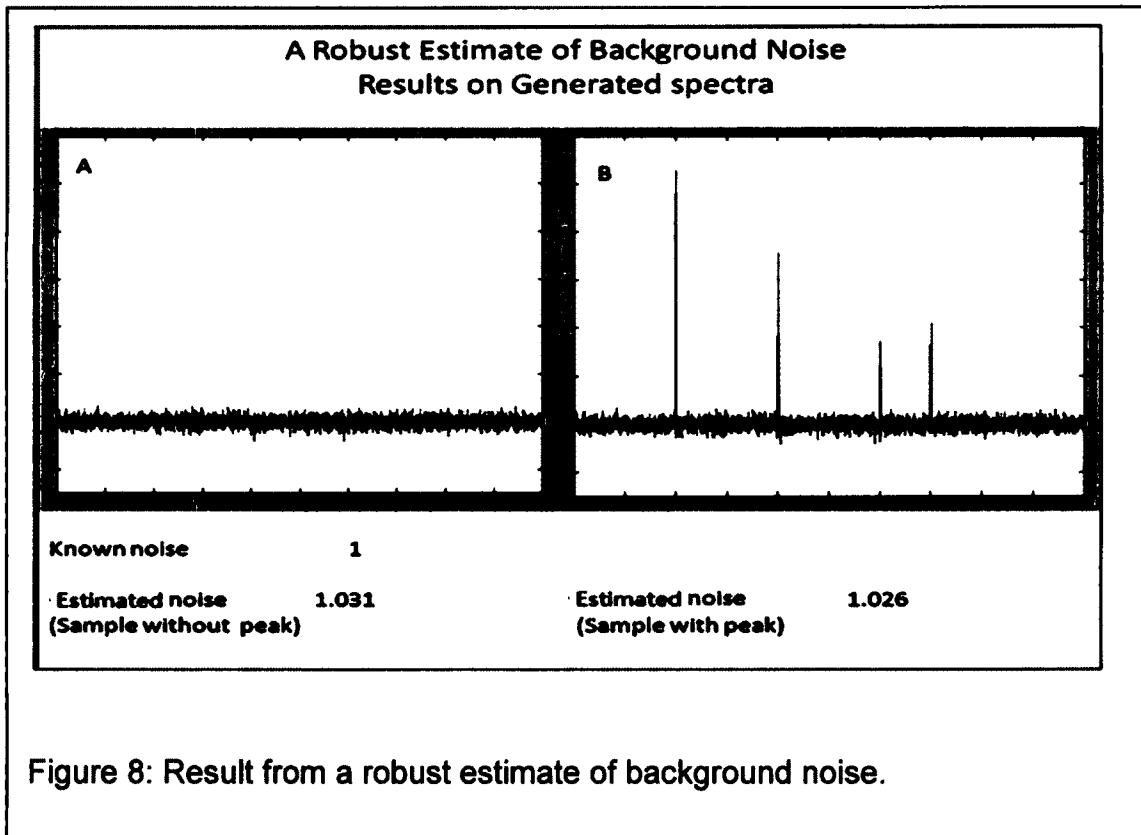
population standard deviation is  $\sigma$ . For small values of  $x$  near its mean, the

the CDF is approximately linear:

$$CDF = \frac{1}{2} + \frac{1}{2} \frac{2}{\sqrt{\pi}} \frac{x}{\sqrt{2}\sigma} + O(x^2) + \dots = \frac{1}{\sigma\sqrt{2\pi}} x + \dots \quad (22)$$

In Figure 7, we show the central part of the *CDF* of a normal distribution, and its approximation by a line. The slope of this center part of the *CDF* is inversely proportional to the standard deviation of the distribution.

Our automatic noise estimation method uses the distribution of differences between two nearby data points in a spectrum to generate a *CDF* of values that have a zero mean. We use pairs of data points that are two FWHM apart to prevent the slow variation of the line shape near its center from contributing to our estimate of the noise variation. The data at the peaks will create differences that are large (either positive or negative), and so they will not affect the slope of the *CDF* near its center. The inverse of the slope of the central part of the *CDF* then yields a representation of the stationary noise in the spectrum.



We have generated spectra with a known noise character to illustrate our noise estimation method, as shown in

Figure 8. The left hand side shows a simulated noise spectrum of 5000 normally distributed points with variance 1. The estimated noise is 1.031, only 3.1% different from the value used to generate the spectrum. The difference is approximately  $\sqrt{2/N}$  as expected for 5000 difference points. The right hand side shows a simulated spectrum including four Gaussian peaks (FWHM=10). Because the peaks primarily affect the ends of the *CDF*, the estimated noise value is still within 2.6% of the value used to generate the data.

We have repeated this for a variety of different numbers of data points and plotted the difference between the variance used to generate the data, and that estimated by our method as introduced before. As shown in

Figure 9, each blue point was generated from 10 sample spectra, and the error bars represent  $\pm$  one standard deviation of those estimations. We also plot the expected error of  $\sqrt{2/N}$  as black dots for comparison.

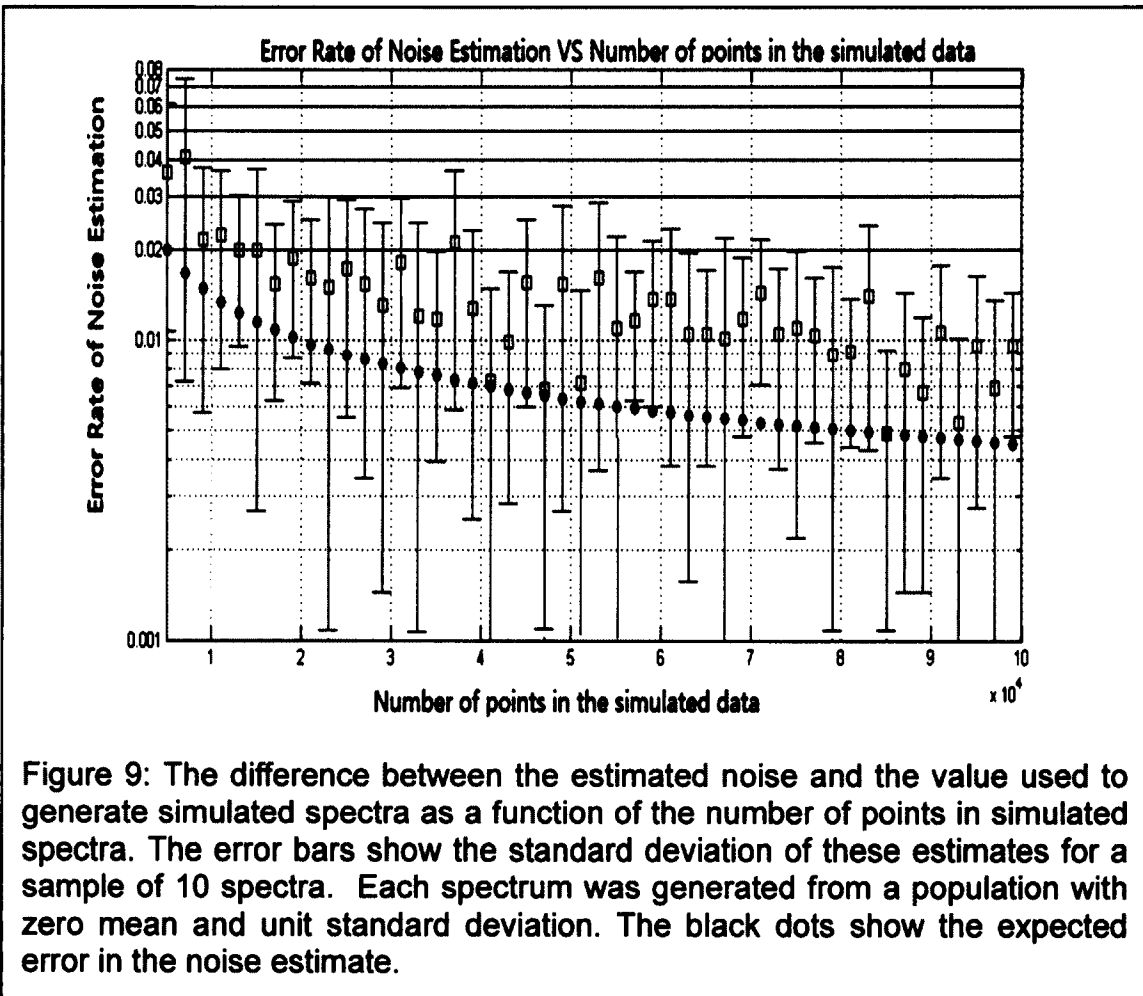
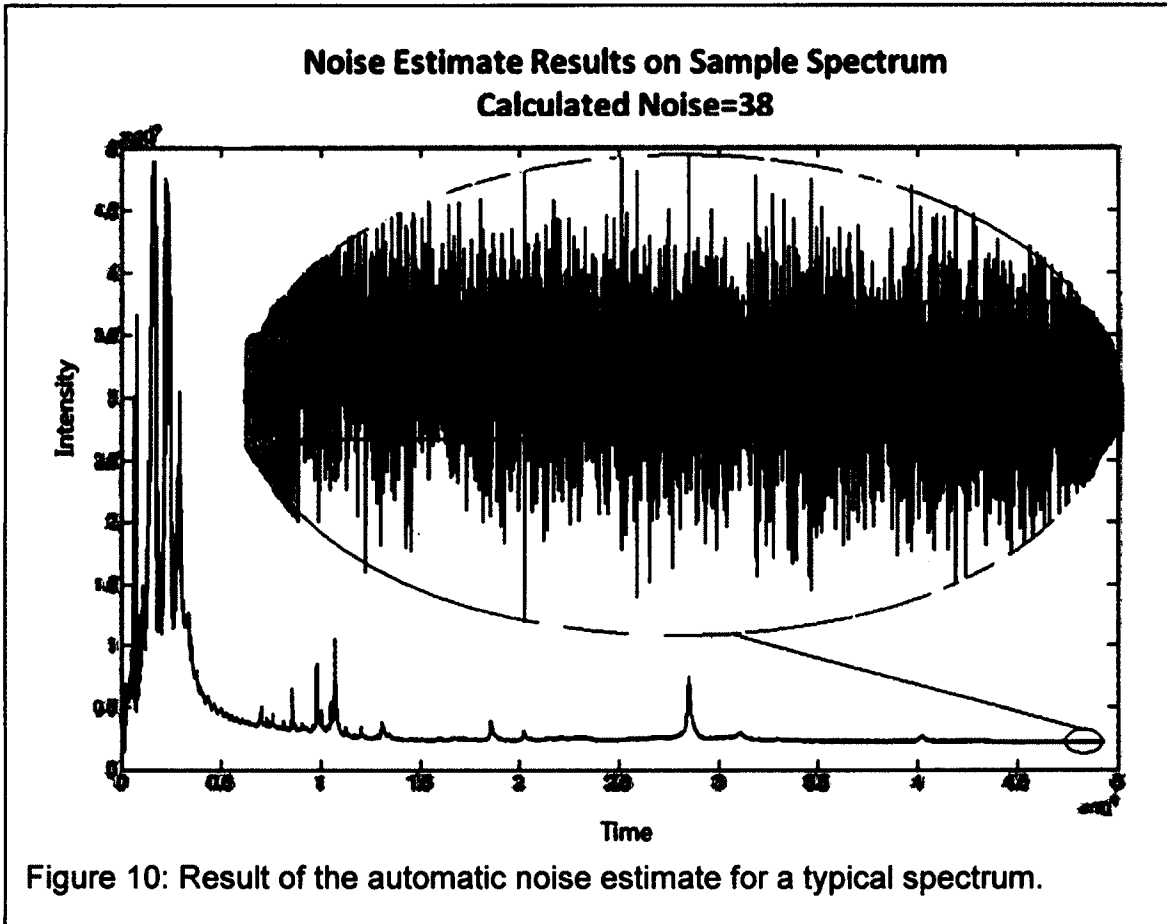


Figure 9: The difference between the estimated noise and the value used to generate simulated spectra as a function of the number of points in simulated spectra. The error bars show the standard deviation of these estimates for a sample of 10 spectra. Each spectrum was generated from a population with zero mean and unit standard deviation. The black dots show the expected error in the noise estimate.

In Figure 10, we show the results of applying this method to one of the TOF spectra used in this work. Our noise estimation method produces a noise estimate

of 38 for these 50,000 points, and the results are shown in an expanded view of the data near the end, where there appear to be no peaks.



***Signal to Noise Ratio calculation***

The *SNR* is the ratio of the average peak signal size to an estimate of the local noise. It is a quantitative indicator measuring how much a peak is distinguished from the background noise. The likelihood calculation produces an estimate of the peak amplitude, so the average signal size is  $A \cdot \sum_{i=1}^N x_i / N$ . The local noise is estimated as the square root of the average variance of the fitted data weighted by the amplitude of the noise estimate from the previous section plus the

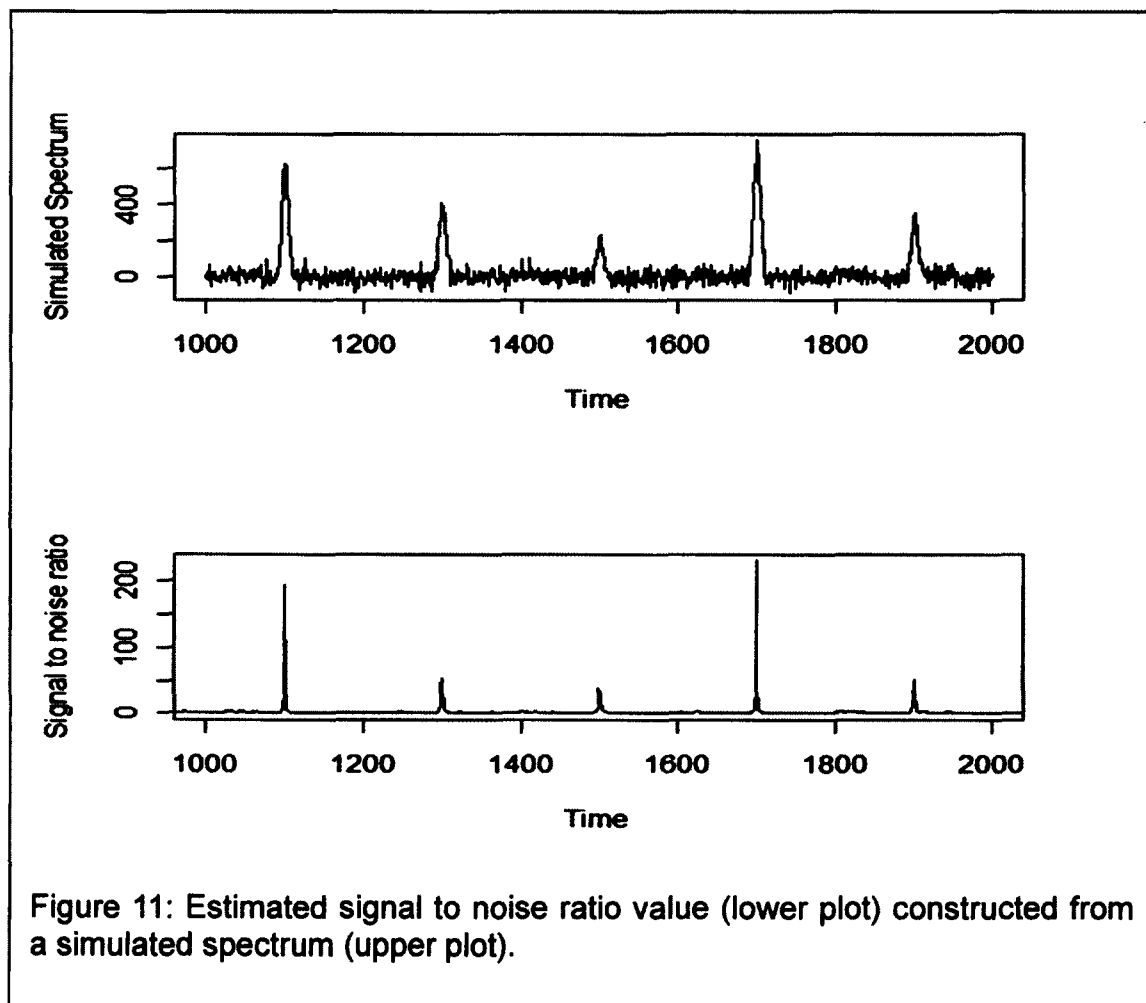
amplitude of the peak signal. This weighting accounts for the  $\sqrt{n}$  growth of the statistical fluctuations in a signal composed of  $n$  ions following a Poisson distribution. The net result is a *SNR* for a peak centered in a sliding window given by:

$$\frac{A^* \sum_{i=1}^N x_i / N}{\sum_{i=1}^N \left[ \frac{(S_i - A^* x_i)^2}{\mu_i + A^* x_i} \right]} \quad (23)$$

where  $N$  is the total time points within the window,  $A^*$  is the maximum likelihood estimate of the peak amplitude (with details in section: ***Likelihood***),  $x_i$  is the  $N$ -component vector describing the peak line shape  $(x_1, x_2 \dots x_N)$ , and  $S_i$  is the  $N$ -component vector of data points  $(s_1, s_2 \dots s_N)$  in the window of interest, and  $\mu_i$  is the local estimate of the stationary noise (with details in section: ***Automatic Noise Estimation***).



As a window slides along the spectrum, this will generate a *SNR* for each time point in the entire spectrum. Figure 11 shows the calculated *SNR* for a simulated spectrum which contains 1000 data points and 5 peaks of varying intensities embedded in a noisy background. A Gaussian shape (FWHM=10) is generated to simulate peaks, the background noise is normally distributed with standard deviation as 30. These simulation parameters were chosen to match typical data in our data set. The top plot shows the generated spectrum, while the lower plot shows the *SNR* over the same region. High values of the *SNR* indicate the regions where peaks are likely.



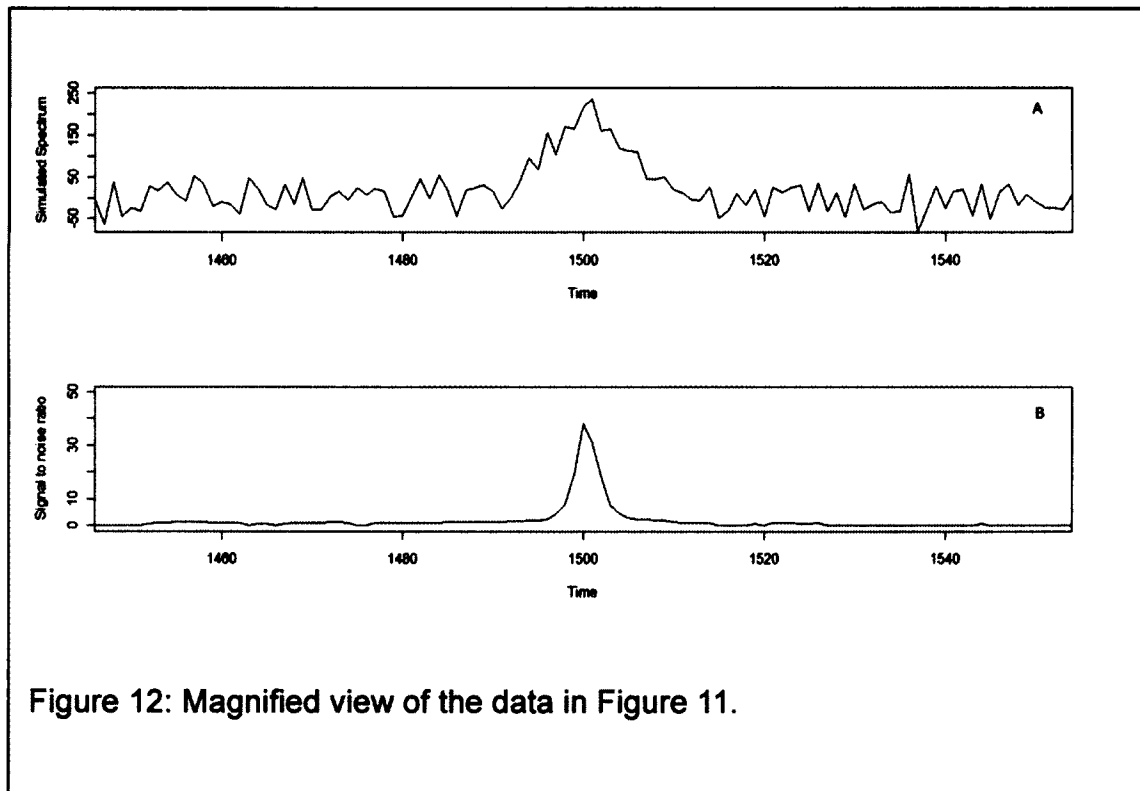


Figure 12: Magnified view of the data in Figure 11.

In Figure 12, we show a magnified view of the peak near 1500 time points. The larger the *SNR*, the more confident we are about the peak evidence. The smaller the *SNR*, the less chance there maybe peak there. A small *SNR* value will exclude the possibility that there exist a peak.

The uncertainty of reported peak position is determined as following:

$$\frac{\text{Window size (FWHM)}}{\text{SNR}}$$

(24)

### ***SNR threshold***

*SNR* values above the *SNR* threshold value indicate regions of the spectra that are likely to contain peaks. The value of the *SNR* threshold should be chosen so that small peaks with a relatively low *SNR* are detected, while the risk of

reporting false peaks remains small. As seen in the earlier figures, the *SNR* is based on the estimate of the peak amplitude  $A^*$ , which is much larger near the peaks. Consequently, if the *SNR* are sorted in increasing order, there will be a turning point, above which all the *SNR* values result from peaks. This point would be the ideal choice for a *SNR* threshold as it would separate regions with peaks from those without. Thus, to select a *SNR* threshold, we will estimate the maximum *SNR* expected from pure noise, assuming that it follows the standard *CDF* used to estimate the stationary noise.

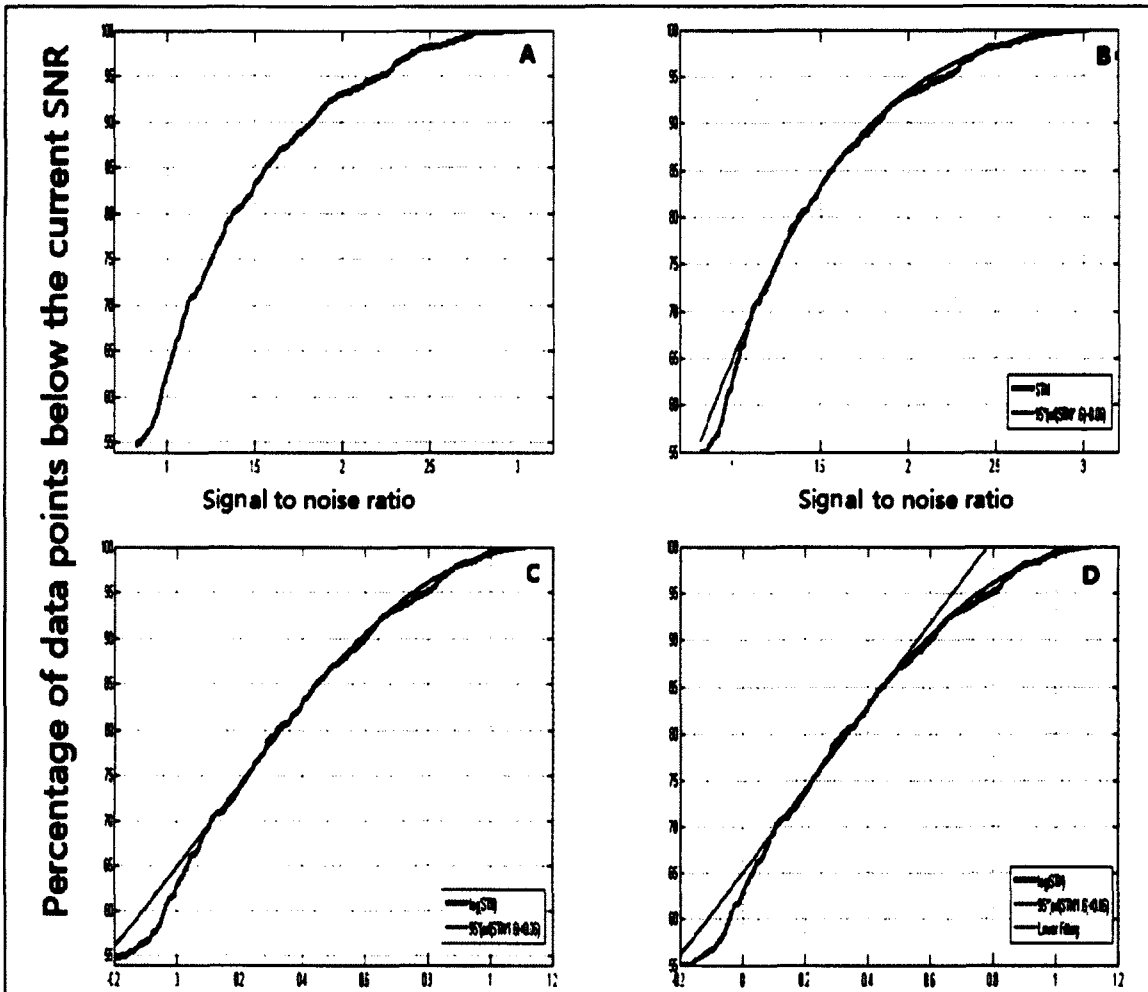


Figure 13: The distribution of SNR from a simulated noise spectrum in panel (A), compared to the expected CDF in panel (B), plotted on a log scale in (C), and with the linear fit in panel (D).

Figure 13 shows the entire process, with the sorted CDF of the SNR for a simulated pure noise spectrum of 5000 points with a standard deviation of 30 in panel (A), the expected CDF based on an error function in panel (B), the CDF on a log scale in panel (C), and the linear fit to the center region in panel (D). We choose the SNR threshold as the value where the linear estimate of the CDF reaches the point where all SNR values would be observed.

Figure 14 shows the calculated *SNR* threshold for a simulated pure noise spectrum on a logarithmic scale. Figure 14 is the same plot as panel (D) in Figure 13 with the inverted axes, Linear regression (red line) to center part of ordered log (*SNR*) gives fitting parameters, based on that, we get the maximum *SNR* as 0.78 in logarithm scale, which is  $e^{0.78}=2.18$  in a normal scale.

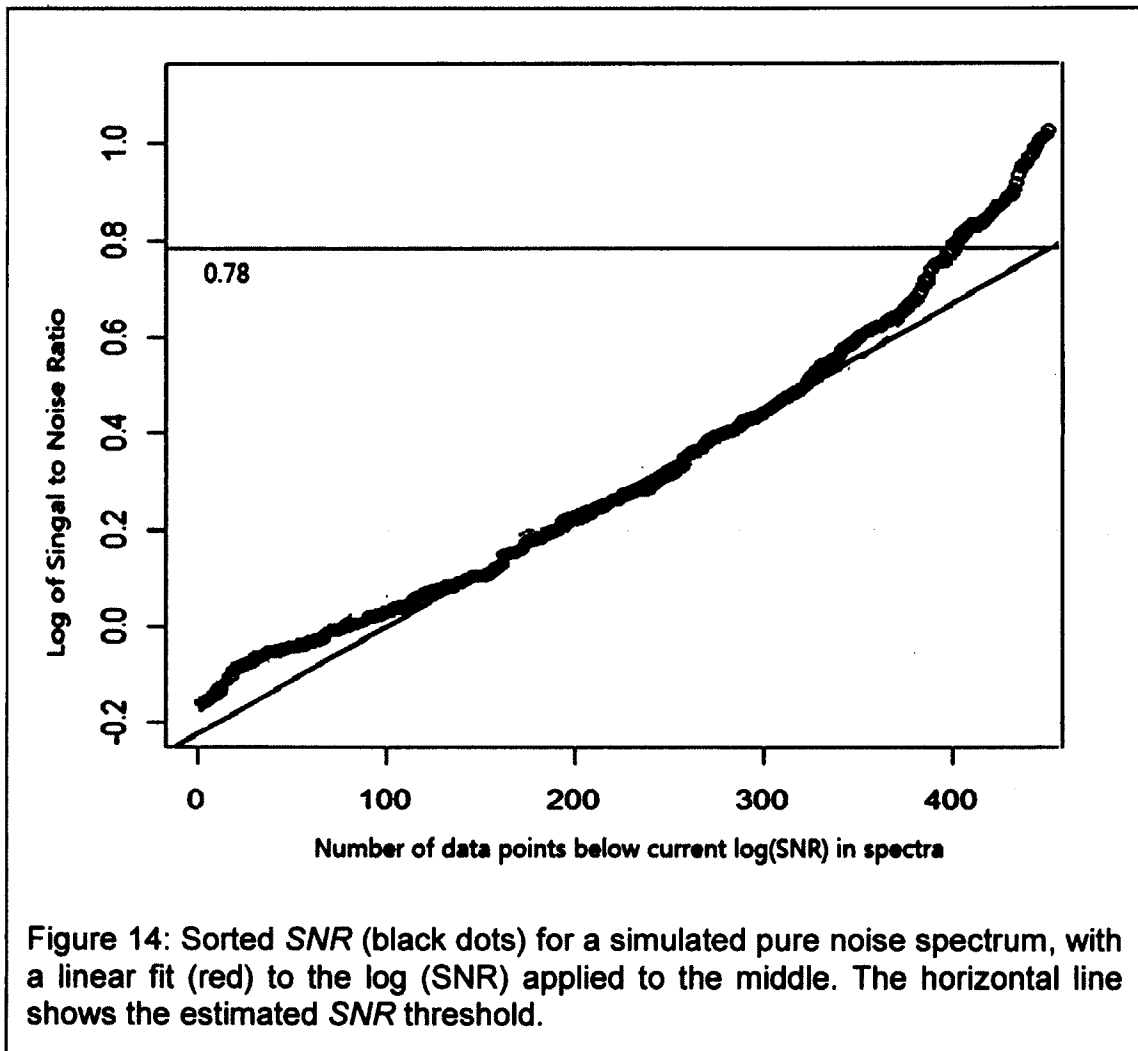


Figure 14: Sorted *SNR* (black dots) for a simulated pure noise spectrum, with a linear fit (red) to the log (*SNR*) applied to the middle. The horizontal line shows the estimated *SNR* threshold.

In a real spectrum, the  $SNR$  values of peaks are higher than those of the noise, so the peaks will only affect the tail of the  $CDF$  and will not influence the fitting process. As we can see from

Figure 15, which shows  $CDFs$  of  $SNR$  from spectrum with five large peaks (black line) and spectrum only contains noise (red line). The shapes of the  $CDFs$  are very close to each other at when  $SNR$  is less than 5.

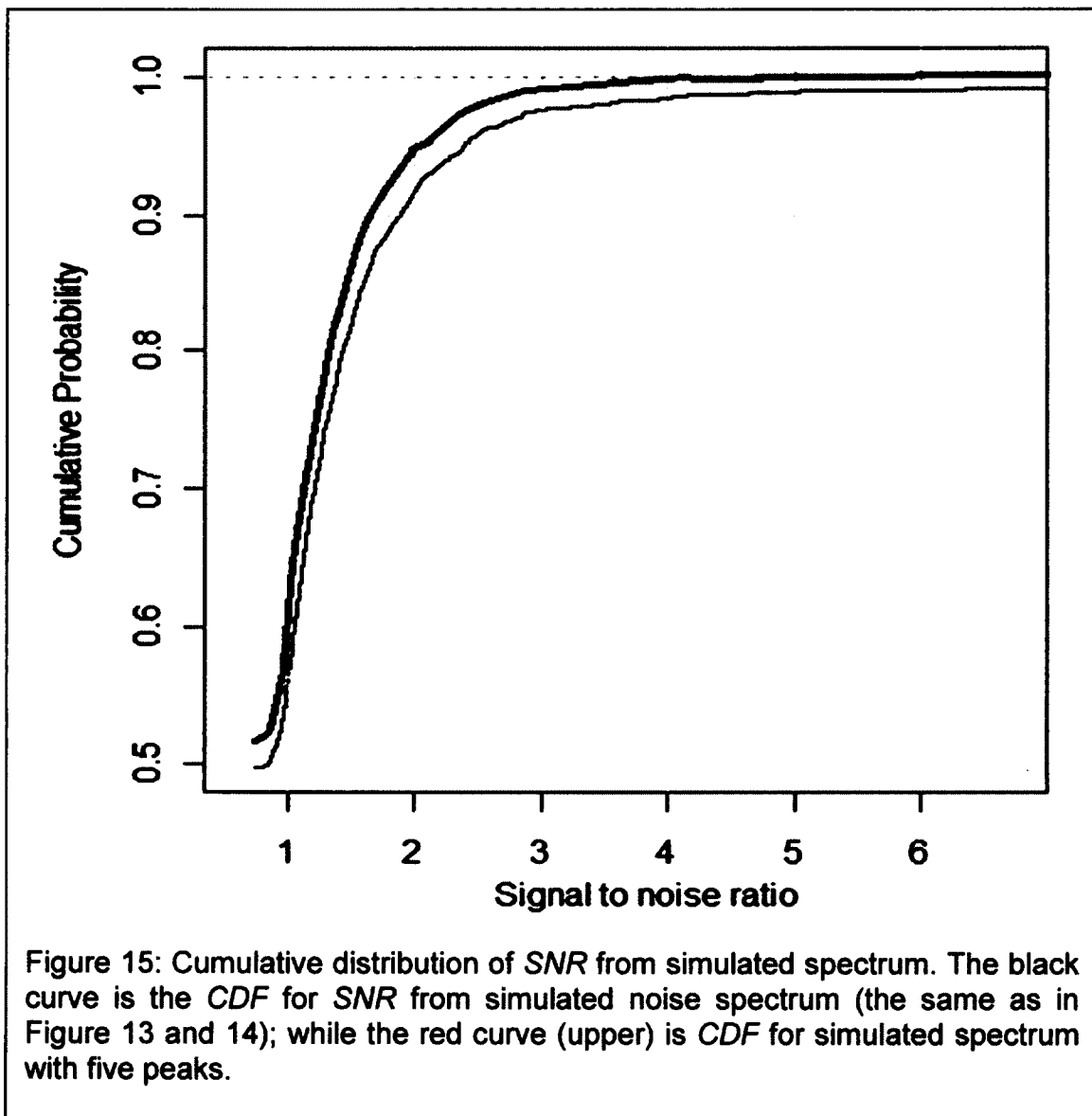
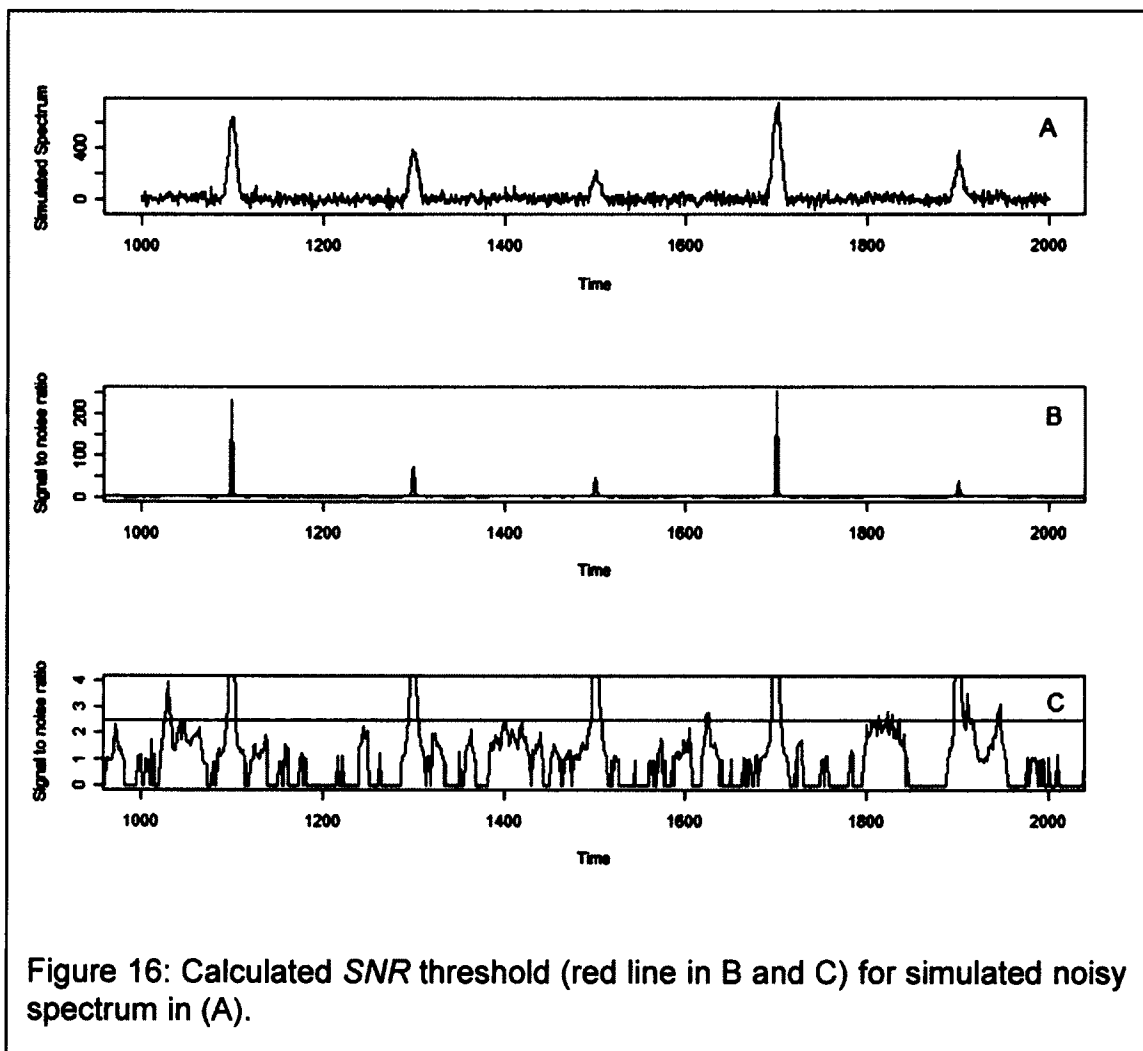


Figure 16 shows the details for the spectrum of Figure 15. Panel A shows the simulated spectrum of 1000 data points including 5 peaks (of FWHM 10) with varying intensities embedded in background noise that has a standard deviation of 30. Panel B shows the calculated *SNR* for this spectrum. Panel (C) shows an expanded view of the *SNR*, along with the estimated *SNR* threshold of 2.48. The *SNR* is based on the estimate of the peak



amplitude  $A^*$ . Sometimes the calculated  $A^*$  may lie below zero when there are

negative noise fluctuations; we will force the calculated negative amplitude to be zero.

The real spectra tend to have significant pedestals resulting from the poorly characterized tails of the mass peaks. Consequently, we have used a local SNR threshold, obtained by using this same method applied to a region local to the point in question. So, we generate an estimate of the *SNR* threshold using only those data points within  $\pm 5000$  points adjacent to each point in the spectrum. We will present the details of this process in Chapter 3. This results in a 9% false detection rate for simulated spectrum data.

## **Summary of the peak detection method**

The automatic peak picking algorithm presented in this chapter is ideal for fast handling of large data sets because the entire procedure operates unsupervised. This also adds the additional benefit that it is not sensitive to potential bias that might be introduced by investigators. Applied to the SELDI TOF-MS data, our peak picking method reduces a typical spectrum to a few hundred peaks.

A successful peak picking method, however, is only one aspect of the low-level processing. The next step is to unify many spectra by correlating the peaks that are found in one spectrum to those that are found in other spectra. This peak alignment will have the added benefit of identifying false peaks that occur when the noise fluctuations happen to mimic a peak. We will eliminate those false



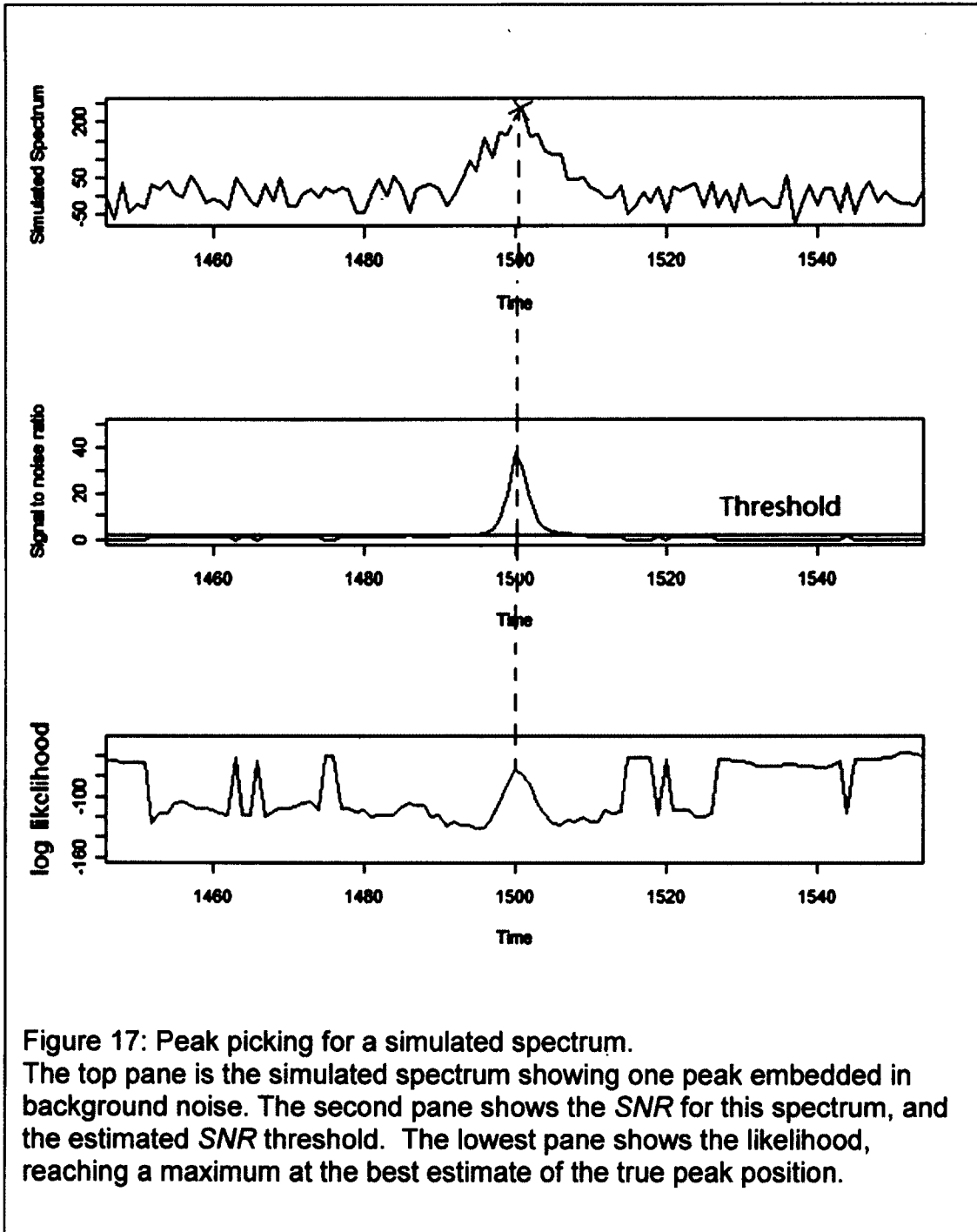
peaks because they will not occur in the same place in multiple spectra. We will discuss the alignment algorithm in detail in chapter 4.

The peak detection process described in this chapter uses six steps as follows:

1. Estimate the stationary noise level
2. Evaluate the likelihood, the peak amplitude, the baseline and the *SNR* for a sliding window
3. Determine a local *SNR* threshold based on a sliding window of 1000 points
4. Determine where the *SNR* exceeds the local *SNR* threshold.
5. Find the window position that maximizes the likelihood for each of those regions that have a *SNR* above the *SNR* threshold
6. Report maximum likelihood of the position, amplitude and uncertainties for each peak found.

Figure 17 shows the entire process applied to a simulated spectrum which includes a Gaussian peak of intensity 200, centered at the time position 1500, added to a noise level of 30. The estimated stationary noise for this spectrum is 31. The *SNR* reaches a maximum value of 36 and the estimated local *SNR* threshold is 2.48.

The likely peak region, where the  $SNR$  rises above the threshold is between 1490 and 1510 time points, and the likelihood reaches its maximum at 1500.1.



# Chapter 3 Data Analysis

In this chapter, we apply the methods from the previous chapter to spectra from the 2004 Leukemia study. Most pattern differences in mass spectra of samples such as plasma/serum are very subtle. In addition, we will demonstrate several steps taken to eliminate the featureless background signal, and show how that allows the peak detection method to operate more reliably.

## 3.1. Introduction

Mass spectrometry data consists of three components: the peak signals, a featureless background, and noise. Most other peak picking methods focus on disentangling these three pieces and apply pre-processing techniques, *e.g.* normalization, noise decomposition, and background elimination, to simplify the data before picking any peaks. Our method only uses background subtraction as the least destructive preprocessing technique to ensure that we keep as much of the information as possible. Consequently, our data analysis method requires four steps:

- (1) Eliminate the background.
- (2) Construct the likelihood for each possible window in the spectrum. This simultaneously generates maximum likelihood estimates of the model parameters and of the *SNR* for each possible window.

- (3) Determine which windows have a sufficiently large *SNR* that they are likely to contain a peak.
- (4) Determine the peak positions, amplitudes and uncertainties for each peak in those identified windows

For the best performance, we have found that the background must be removed iteratively, using the results of the peak picking procedure to better identify how to remove the baseline, and then using the results of the alignment procedure (in chapter 4) to further improve the model of the background. So, each of these four steps will generally be done repeatedly.

## **3.2. Background Elimination**

There is always a non-zero baseline in a raw spectrum because the Analog to Digital Converter (ADC) requires an offset to detect negative signal fluctuations and because very large signals tend to produce a decaying signal after the main peak. The high-signal, low mass, matrix region usually produces an exponentially decaying baseline in the earliest times we have analyzed. The details of the baseline vary with the instrument settings and with the magnitude of the matrix signal that reaches the detector. Our peak picking method allows an arbitrary baseline, as long as it is nearly constant within the fitting window. This not only corrects the ADC offset, but can also correct any background due to ions that does not vary much across a window. Our method routinely picks even small peaks but it sometimes also detects sporadic peaks when the noise fluctuations in a large

background look like peaks. To counter this, we have subtracted various models of the background, and then forced the background to be zero in the peak model. This requires a precise removal of the background from the data, consequently, there are several steps involved in background elimination.

The most common contribution to this baseline, and the easiest to correct, is the large, exponentially decaying signal associated with the matrix products [22].

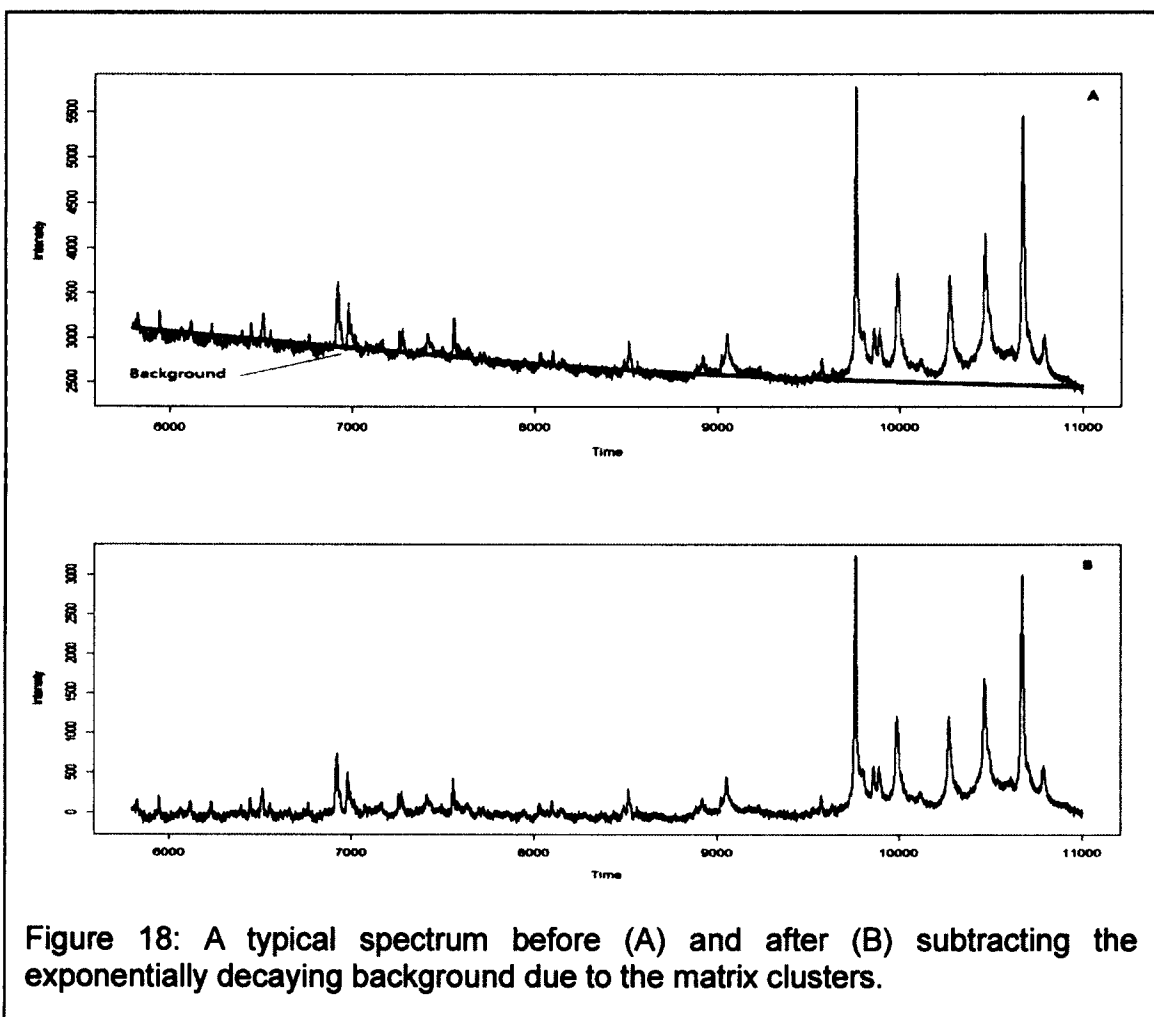
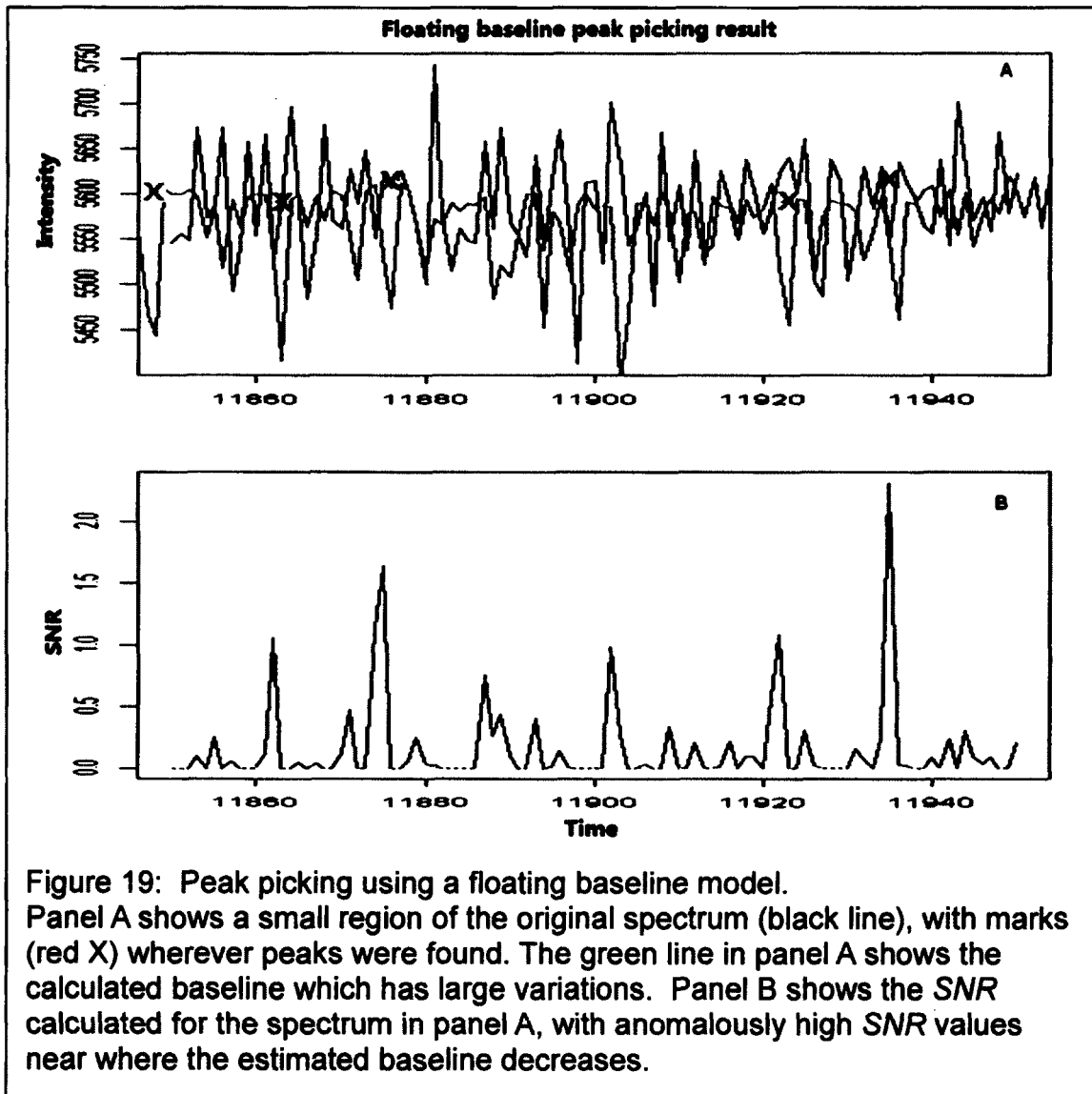


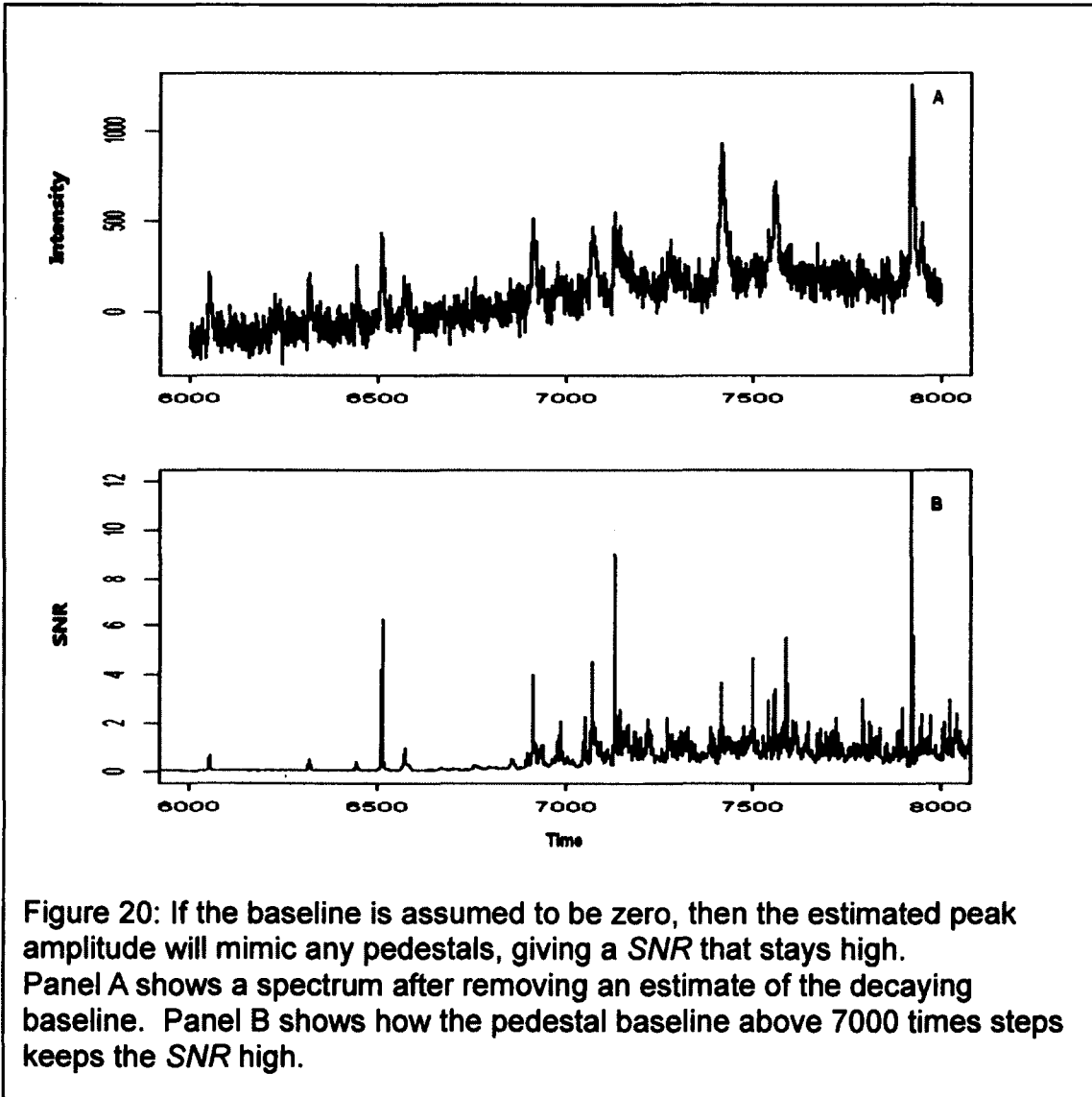
Figure 18 shows one spectrum before and after applying this background subtraction to the early part of a spectrum, where it is most noticeable.

Although this removes the largest part of the background, the accumulation of signal from the poorly understood tails of the large peaks still remains. However, there is no mathematical model for such accumulated pedestals. We therefore determine the remaining baseline by the maximum likelihood parameter estimation, *i.e.* by using a floating-baseline in our data model. However, this makes our method especially sensitive to noise fluctuations which look like peaks and introduces false peaks. Figure 19 shows an example of false peaks resulting from a large, uncorrected baseline. Moreover, since this estimated baseline (green line in panel A) has large fluctuations that are due to statistical fluctuations, rather than real variations in the true average baseline, the peak amplitudes will also be in error. For high amplitude peaks, this estimated baseline is always higher than the true average baseline.



One can avoid the errors caused by the rapidly varying baseline estimate by obtaining a better estimate of the true, average baseline, subtracting that, and then using a model that forces the baseline to be zero. Figure 20 shows that even a small residual baseline can produce a large number of false peaks. In the top panel (A), the spectrum has had an exponentially decaying background subtracted, but some baseline remains due to the combined pedestals of the

high peak density region above 7000 time steps. This small, but growing, background leads to a high SNR as shown in the lower panel (B). These high



SNR values increase the risk of reporting false peaks.

We have improved our model of the baseline based on smoothed data that is not part of a peak. Because the original floating-baseline model primarily errs by introducing false peaks, we use it to determine the regions where there are no



peaks. We use the floating-baseline peak-finding algorithm to determine the position of all peaks, including the false peaks, and then eliminate windows around those peaks to construct a baseline from the remaining data in the spectrum. Of course, some baseline points are excluded due to accidentally picked false peaks, but that tends to be a small fraction of the total data. This new background estimate eliminates most of the pedestals. We use linear interpolation to fill in the background in the regions where the data was excluded, and then a moving average to make the final baseline estimate.

The key to this baseline estimation is to size the exclusion windows and the moving average properly. For the exclusion window size, we have used a window that grows with the peak amplitude to avoid having this baseline be dominated by the rapidly changing pedestal of a particularly large peak. Accordingly, we have used an exclusion window width of:

$$W_{\text{Exclusion}} = \frac{\text{FWHM}}{2} \left( 1 + \sqrt{\frac{2A^*}{\text{FWHM}}} \right) \quad (25)$$

where  $A^*$  is the maximum likelihood estimate of the peak amplitude.

Figure 21 shows the entire process in a region between 11000 and 12200 time steps. In this region, the exponentially decaying matrix signal plays no role. Rather, the downward sloping baseline in panel A of Figure 21 is due to a slowly decreasing pedestal from all the peaks in the earlier region. The blue dots in panel B are the remaining points after the initial floating baseline model determines

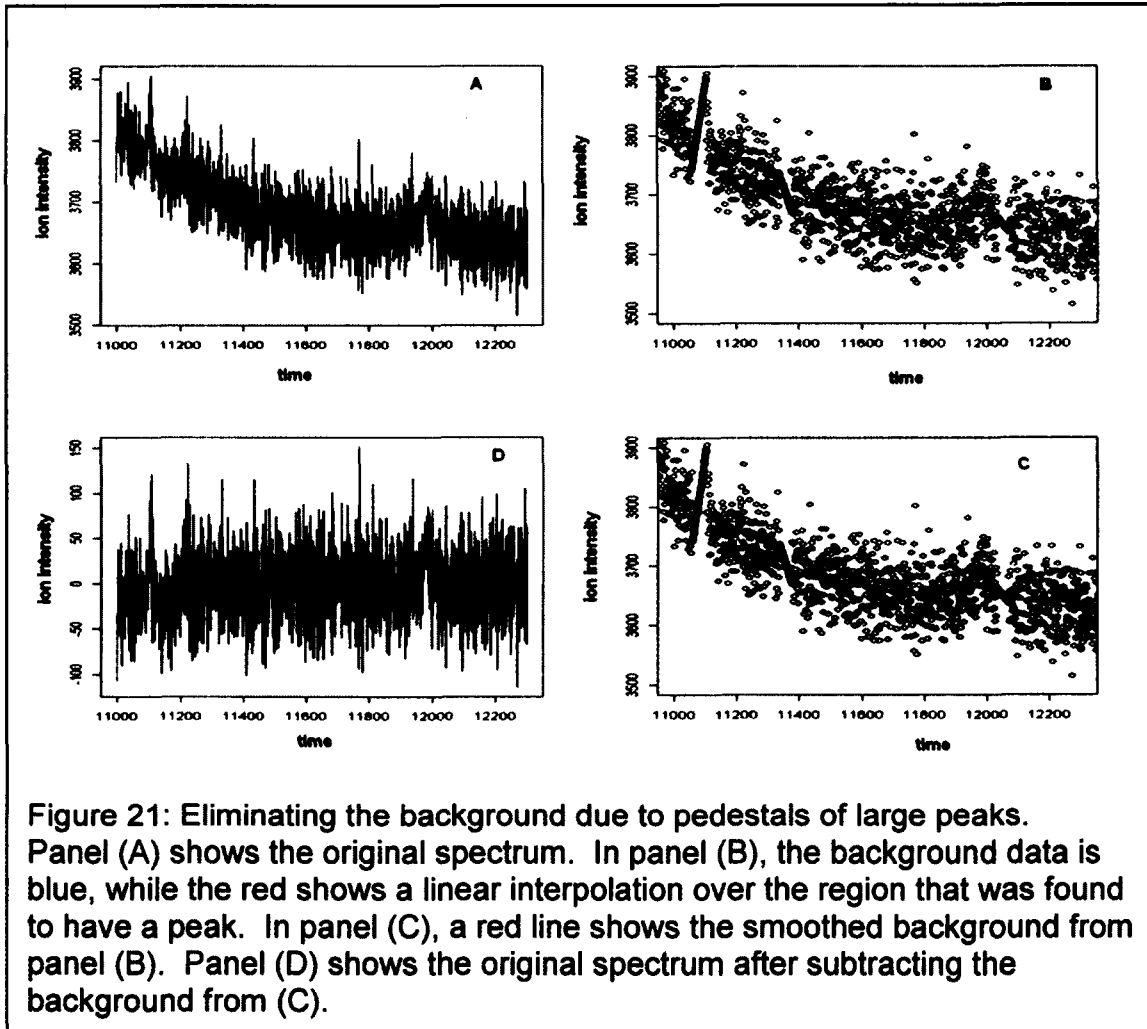


Figure 21: Eliminating the background due to pedestals of large peaks. Panel (A) shows the original spectrum. In panel (B), the background data is blue, while the red shows a linear interpolation over the region that was found to have a peak. In panel (C), a red line shows the smoothed background from panel (B). Panel (D) shows the original spectrum after subtracting the background from (C).

which regions might have peaks, and should therefore be excluded. The red dots are the interpolated points covering the excluded regions. The moving average of the blue and red points gives the estimated background, which is shown in panel C as a red line. Panel D shows the spectrum after the estimated baseline has been subtracted. The slowly varying background has been removed while all peaks remain. This baseline subtraction technique does not subtract signal attributable to slowly varying structure. The baseline correction flattens regions with no peaks while retaining the peak shape where there are peaks. Hence, very broad peaks will not be mistakenly eliminated as background. Other model-based approaches

are only effective with limited data sets and they require a complete understanding of physical aspect of the instruments and the chemical nature of the samples. This parameter-free baseline elimination method requires much less prior information about the data and could be widely used in place of the cumbersome model-based approaches.

### **3.3. Local Baseline Correction**

The previously discussed background subtraction methods work well to remove slowly varying background, either due to the long tail of the matrix signal, or to the pedestals that persist near large peaks or dense collections of peaks. However, they do not work when a small peak is close to a large peak, overlapping the rapidly decreasing edge of the large peak. A rapidly changing background can distort the true shape of the peak, giving rise to errors in both the amplitude and the peak position.

In this section, we present a local background correction that can eliminate the background due to the problematic large peak, even though that peak's line shape is not well described beyond the  $\pm\frac{1}{2}$ FWHM. After our initial peak picking and alignment (to be described in chapter 4) of all spectra, we obtain a master peak list which tells where each peak should be in each spectrum and whether or not that peak was detected. Much like the earlier approach of using the regions where no peaks were found to model the background, here we will use the spectra where specific peaks were *not* found to model the background that they

sit upon. We will start with a special case involving only two peaks, one sitting on the side of a large peak. From there, we will extend the approach to cases where peaks occur in clusters, and apply a similar approach to modeling the effect on each peak of the baseline due to the others.

### ***A Small Peak Adjacent to A Large Peak***

When a small peak sits near a large peak, the large peak may introduce a significant pedestal that is not corrected by our previous methods. Typically, this happens when both peaks fit within the exclusion window that was used to determine the baseline. Then, the pedestal under the small peak is also contained inside the exclusion window, so it is not removed. To correct this, we have corrected the baseline near the small peak by using an average over all the spectra in which the small peak was not discovered. Although this average may include some spectra in which the small peak is present, but not detected, this average will generally match the shape of the baseline due to the large peak pedestal, so that it can be subtracted.

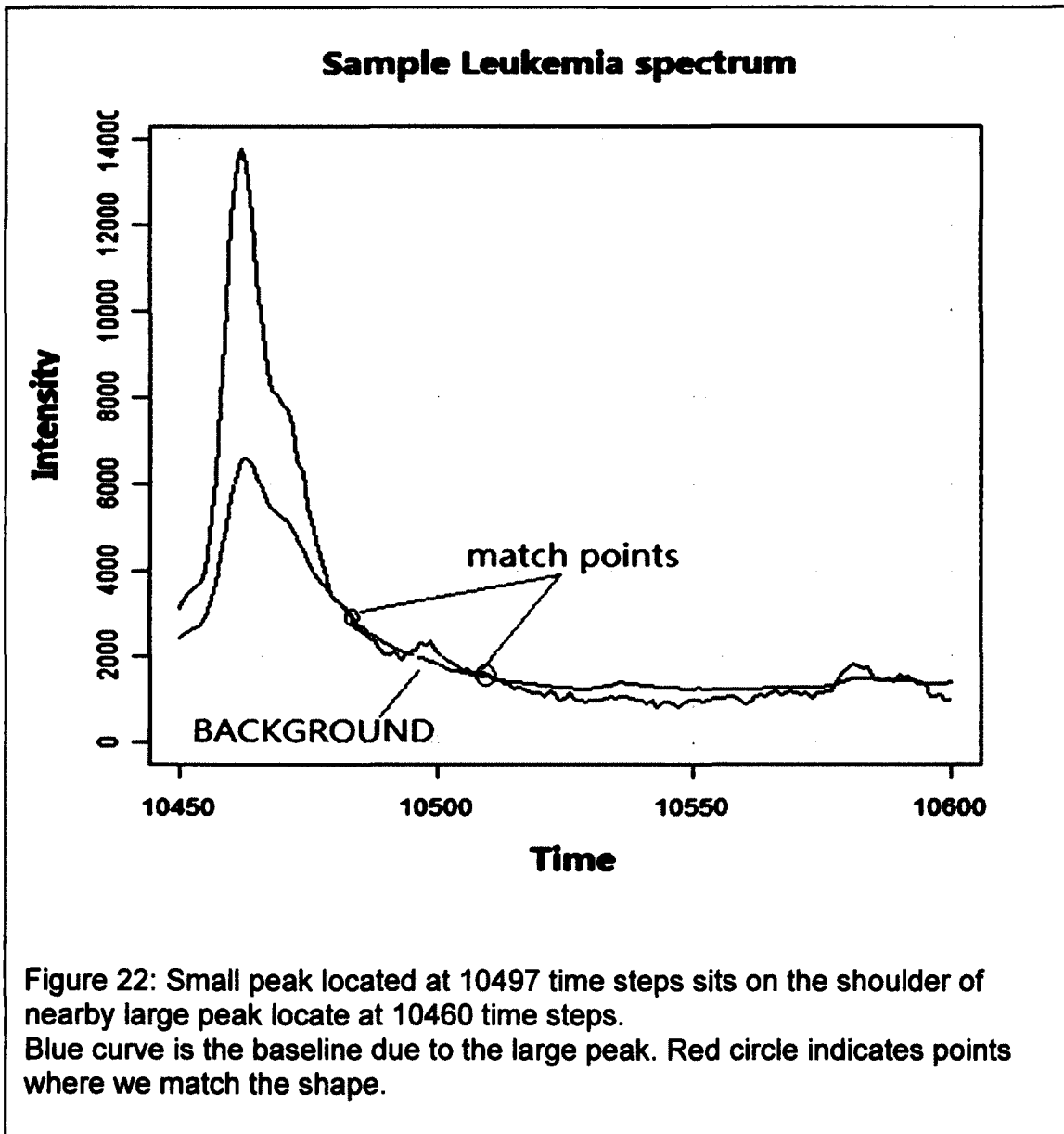
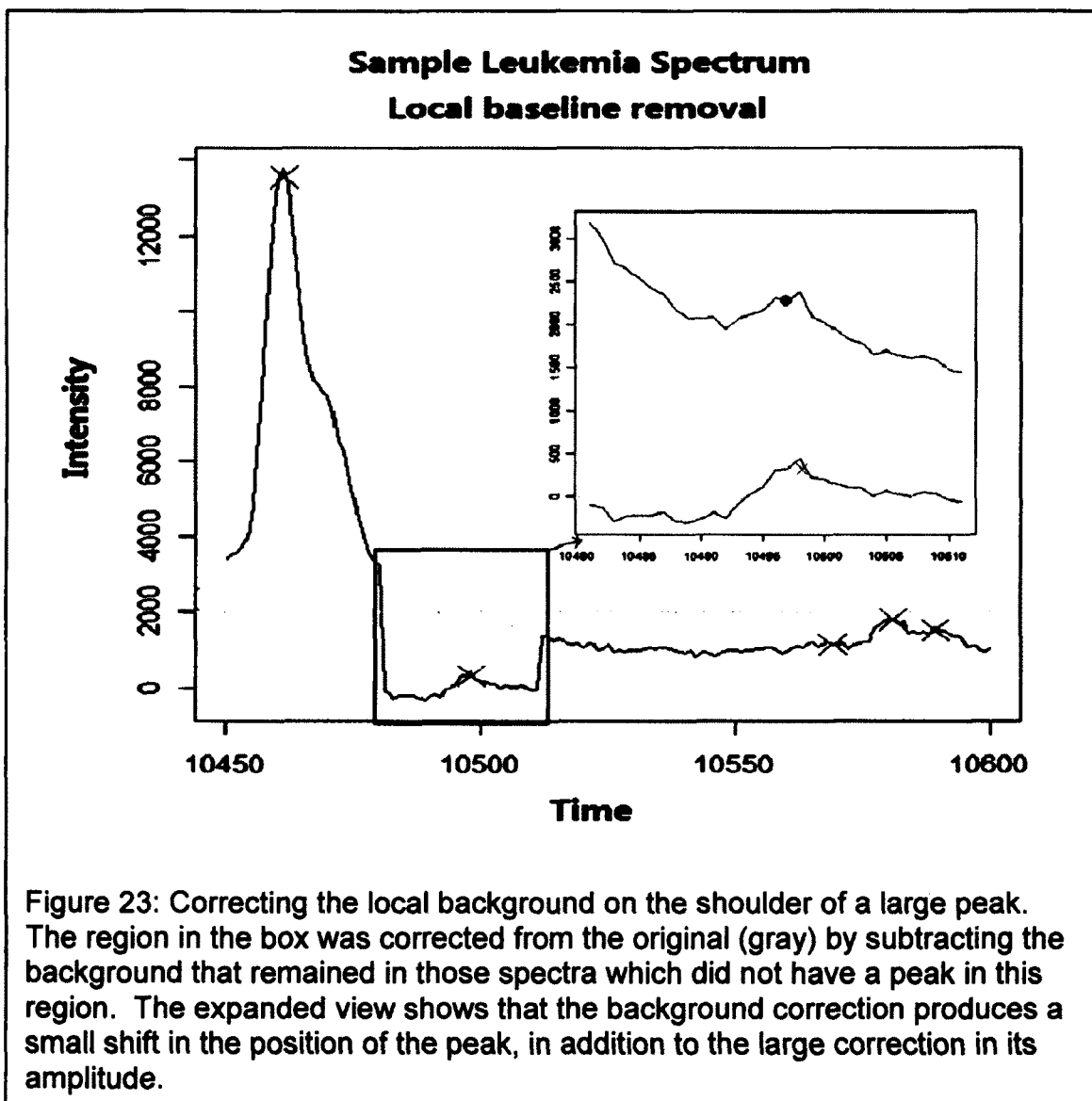


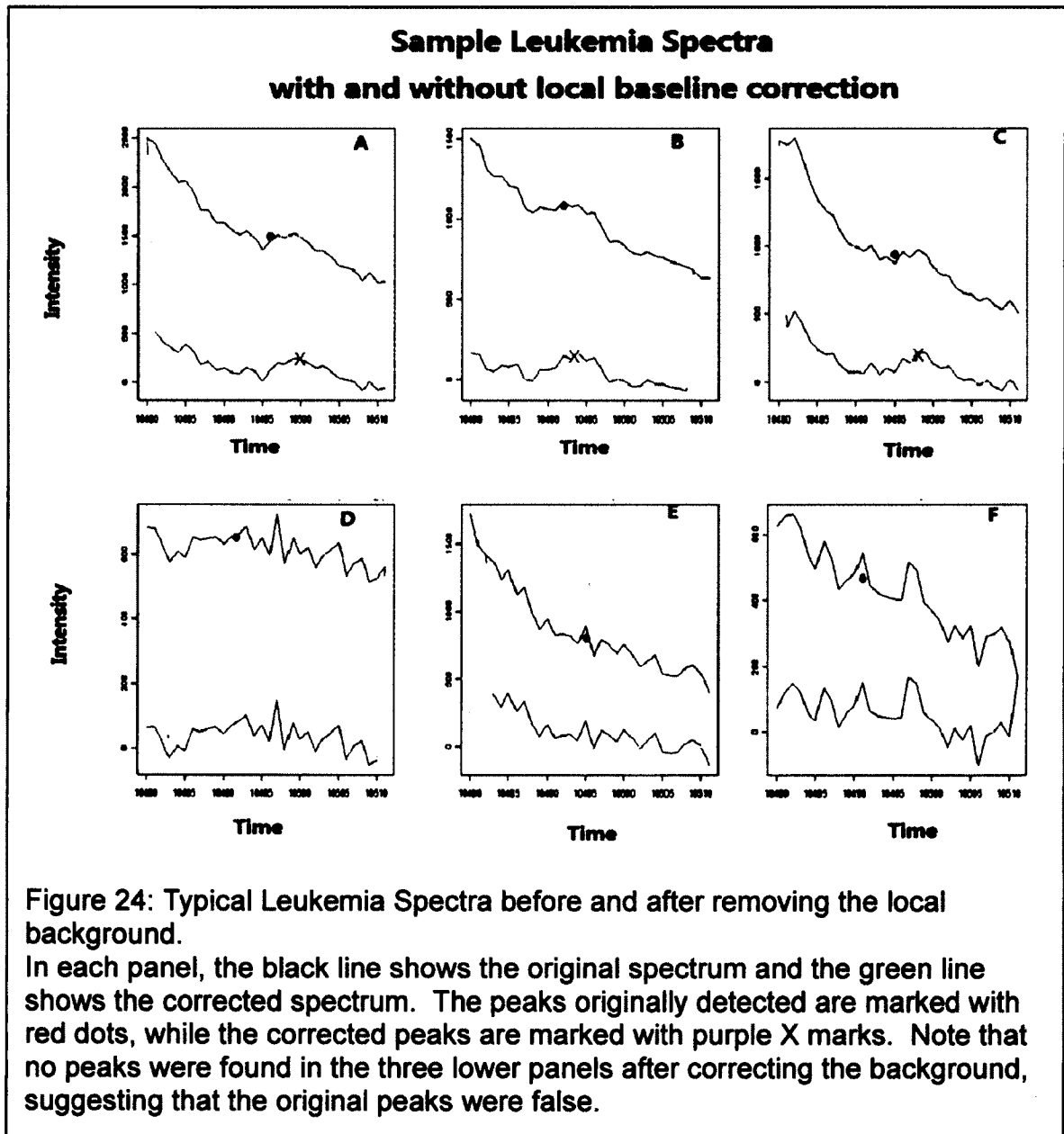
Figure 22 shows such an example where a small peak, near 10,497 time steps, is sitting on the shoulder of a large peak at 10,460 time steps. As we can see from Figure 22, the wing of the large peak still contributes a slowly decreasing baseline of nearly 2000 at the location of the small peak. The blue curve is our constructed local background due to the tail of the large peak around 10,460 time step. First we calculate the average of spectra without report of the small peak



near 10,497 time steps as the background shape, and then adjust the background to each spectrum by matching the background shape over the averages of five end points.

Figure 23 shows the local background removal result. The expanded view shows the original spectrum as a black line, with a red dot at the location of the detected peak. The lower, green line shows that same region after subtracting

the baseline created from the average of all spectra where no peak was detected near 10,497. The new peak position and amplitude is indicated by the purple X. There were two major changes, the new peak amplitude no longer contains the ~2500 contribution from the large peak, and the position has moved slightly to



the right, reflecting the removal of a *sloping* baseline due to the large peak.

Figure 24 shows six other cases of applying this local background correction in the region adjacent to a large peak. In all six, the black line shows regions of the original spectra, with red dots at the location of detected peaks, while green line shows the spectra after applying the local baseline correction, with purple X marks indicating the new peaks. In the top three spectra, the new peak positions and amplitudes better represent the true peak, while in the lower three spectra, no peaks were found after the baseline correction, indicating that those three spectra had false peaks detected because the pedestal level was high, and the baseline fluctuations mimicked a peak.

#### ***Local Baseline Correction near Clusters of Peaks***

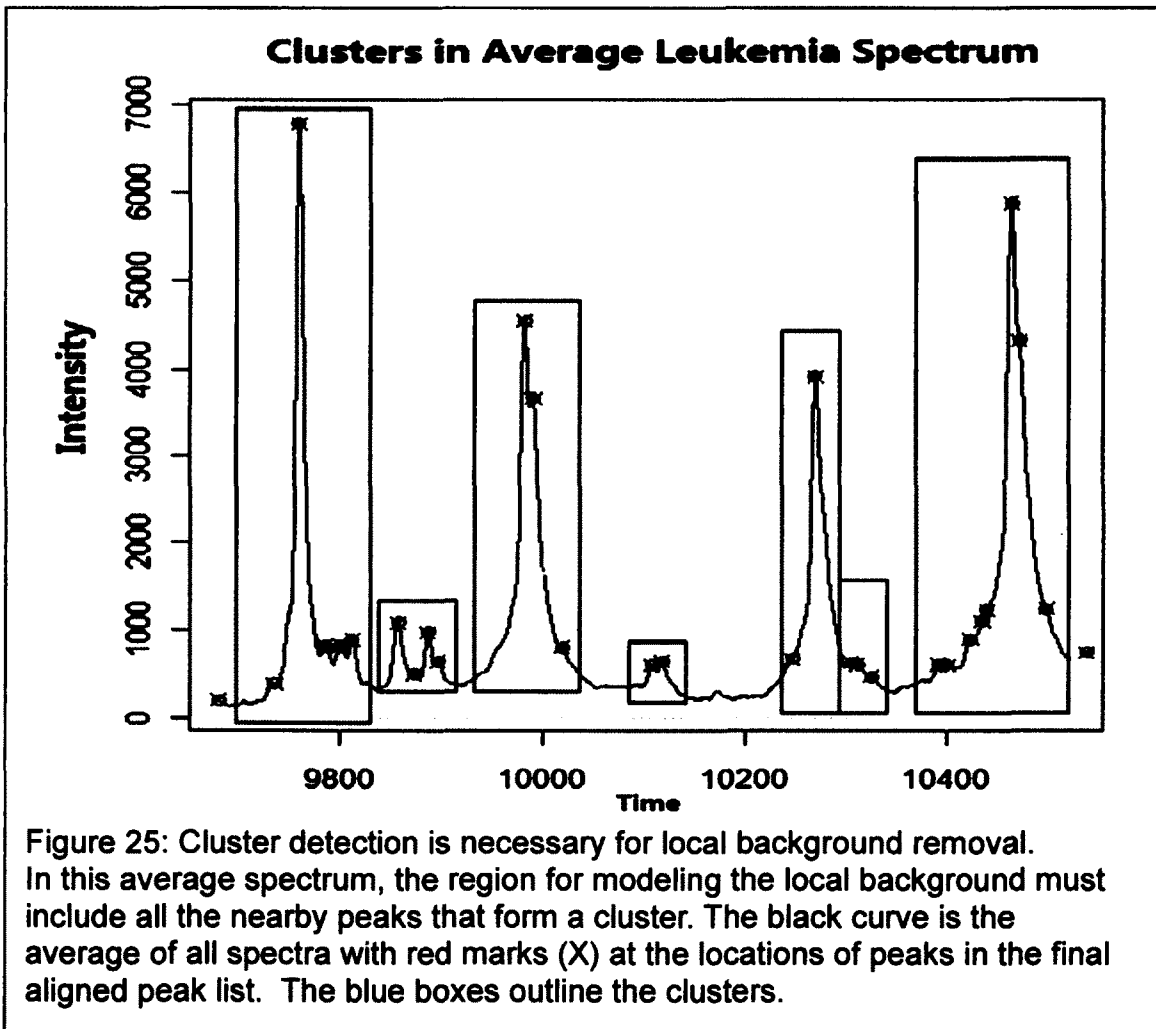
This improved baseline modeling method is particularly important in regions with a high peak density. However, in many of those cases, the baseline will have contributions from more than one nearby large peak. In such cases, we have found that we must treat clusters of peaks, rather than the pairs of peaks considered in the previous section. For example, when one matches the model baseline to the average of the spectra without a given peak, it is crucial that the end points of that modeled baseline not be located in the vicinity of a peak, or else the amplitude of that peak will dominate the process of matching that baseline to the spectrum. However, once one uses clusters of peaks, then there will generally be peaks within the modeled baseline that one subtracts. This means that in addition to matching the end points of the modeled baseline, one



must also match the amplitude of the variation within this region, so we also set an additional parameter to insure that within the cluster, we match the end points and wherever the signal is largest.

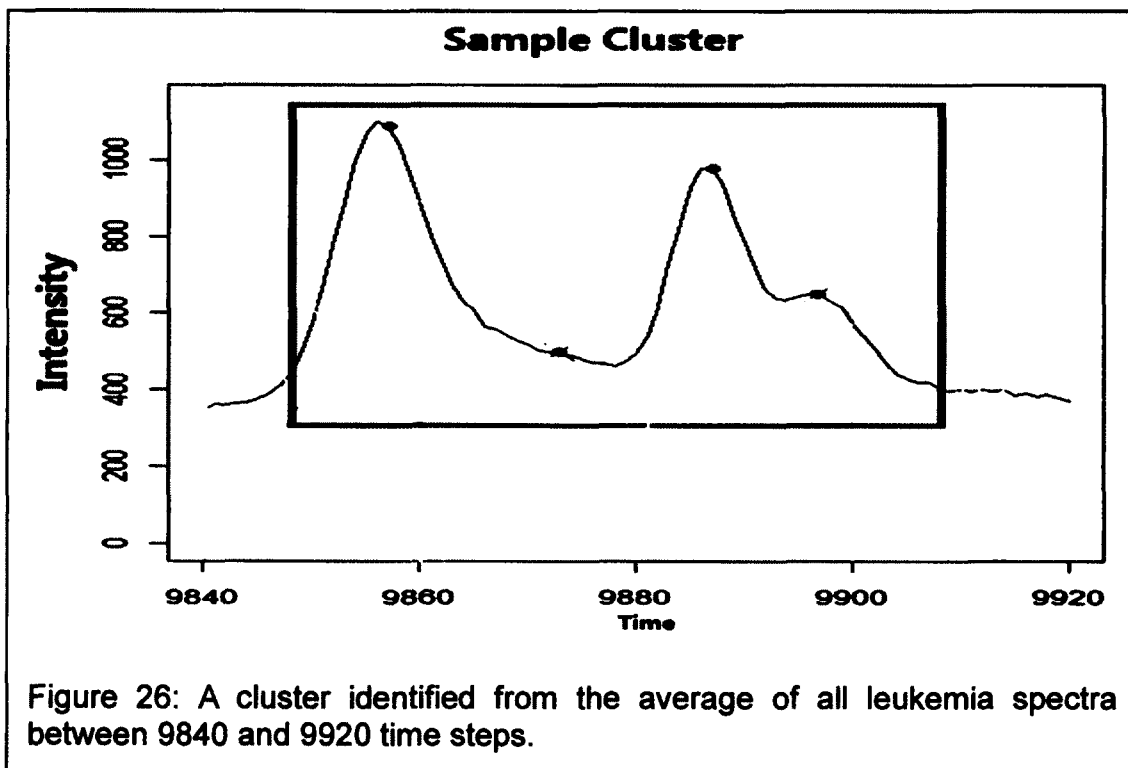
### ***Determine Local Baseline Region***

Before applying the automatic local background correction method, we first



identify the locations where local background corrections are needed. Peaks within a half FWHM distance are considered to be too close to having a locally raised background. Thus, a cluster, where local background corrections are

needed, is determined wherever peaks are within half FWHM away to each other. The cluster boundaries are determined as FWHM away from the first and last peaks belonging to the cluster. Figure 25 is a graph showing part of the result of the final determined clusters on average leukemia spectrum at region 9700-10600. Red Xs shows positions of the aligned peaks, which are peaks supposed to be found at all the spectra. Blue rectangles are the determined clusters. Figure 26 is a closer look at the cluster of region 9840-9920. There exists a local baseline with an intensity of around 300, which likely result from large peaks some distance nearby.



Once the clusters are identified, we go through each cluster to do the local background correction for each aligned peak and for each spectrum by adjusting every spectrum locally to the average spectra without local large peaks. The averages of five end points and maximum of the region are used to match the background shape. After correcting local background, the peak picker algorithm is applied again to get more accurate report of peak information just around the region where original aligned peak were reported. The same process is applied to each aligned peak within each cluster.

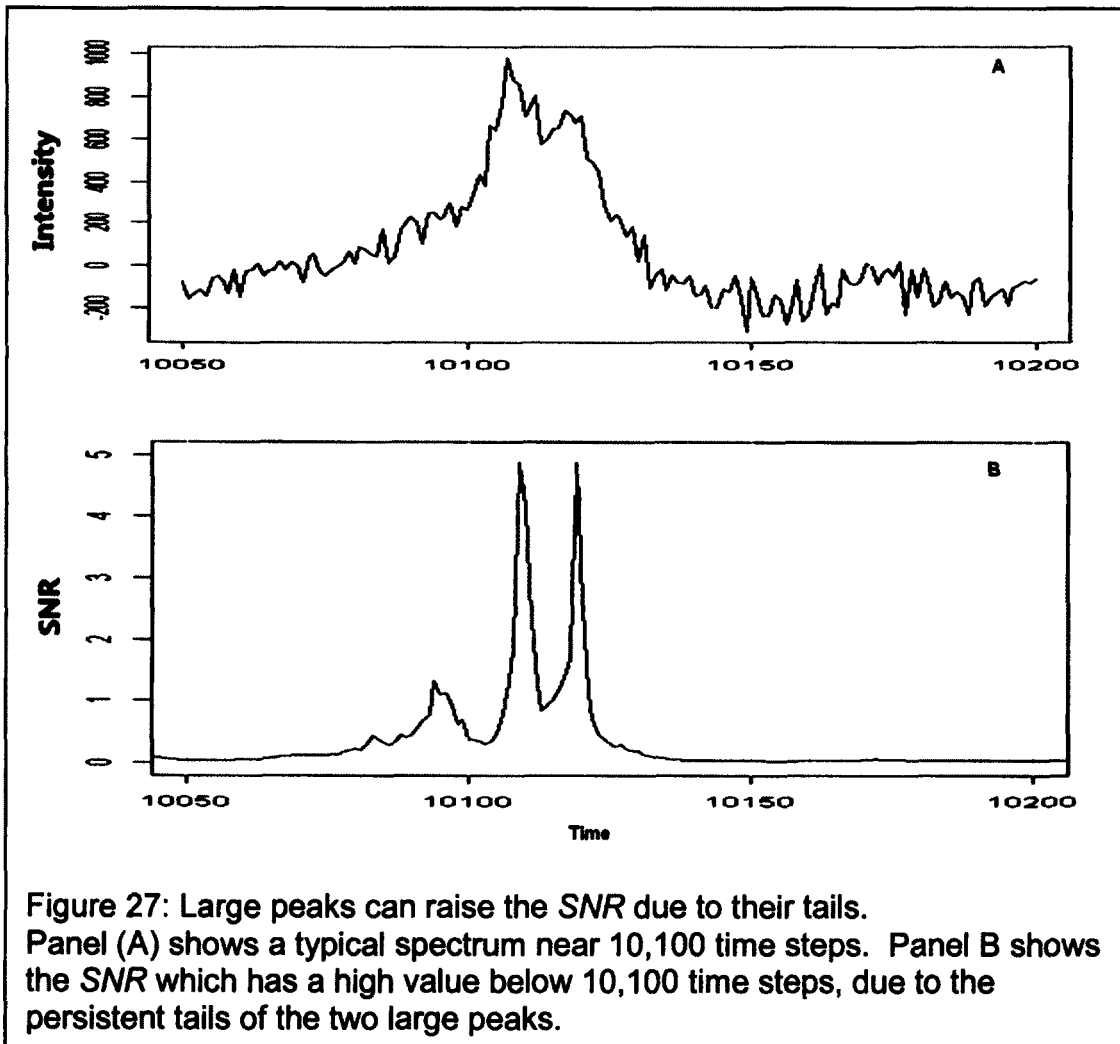
In this section, we have introduced a non-supervised local background correction method; this method is based on the idea that the spectra data belonging to the same group should be similar. By applying this method, we can get more accurate peak information.

### **3.4. Determination of The Optimum Signal To Noise Threshold**

We have reversed our normal procedure to use our standard peak picking algorithm to identify the regions where there are no peaks so that we can then model the baseline and eliminate it. However, even with this improved baseline correction, a small component of peak pedestals remain, and this degrades the signal to noise calculation because the higher baseline level appears to increase the signal, making this estimate of the *SNR* no longer appropriate to identify peak regions. Thus, assuming a constant *SNR* limit may introduce spurious peaks.

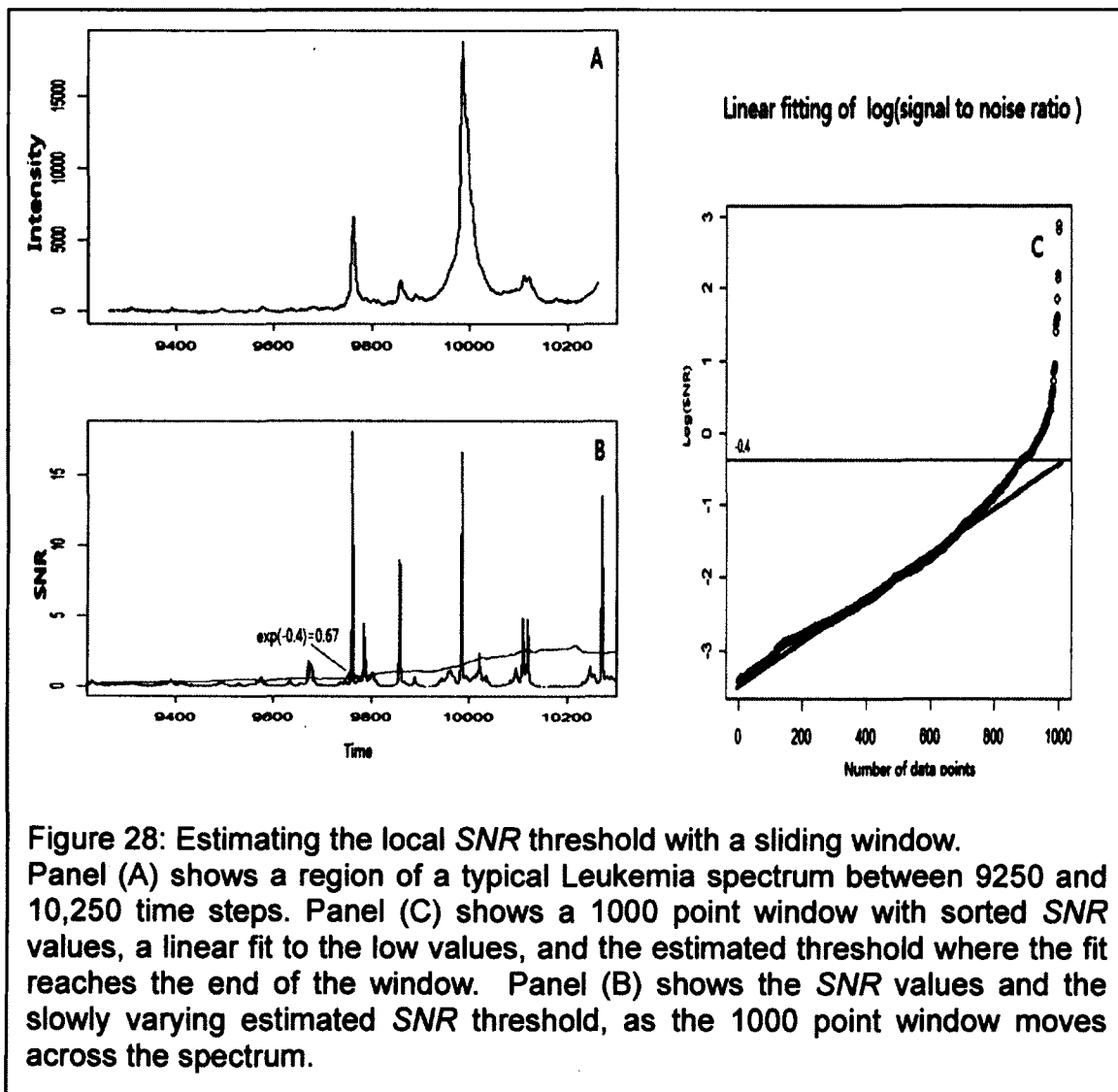
We have therefore introduced a variable (but slowly varying) *SNR* limit to determine the peak region. We describe this procedure in this section.

The peak identification method discussed before requires an appropriate pre-determined peak shape. Here we use a simple normal distribution to model the peak shape, which captures most of the characteristics of the upper half of a typical peak. For the tail of a peak, especially at high intensity, we do not have a good model of the shape. Moreover, clusters of closely spaced peaks would add



even more complications to the modeling of peak tail shapes. But the tails of the high intensity peaks effectively increase the baseline, and this makes the *SNR* appear artificially high. Figure 27 shows an example peak region where we can see how sensitive *SNR* is to the wide pedestals.

This situation depends largely on the local signals. Clearly, setting a constant threshold value to determine peak regions will pick spurious peaks at high signal



regions while missing some true peaks at low signal regions. To correct this, we have introduced a slowly varying *SNR* limit suitable for different local signal types. The method is similar to the use of a moving average: we calculate *SNR* limit value based on the data within a 1000 points window which moves through the entire spectrum. For each window position, a linear fit to the central part of the rank-ordered log *SNR* estimates the *SNR* limit value. Figure 28 shows the calculation of *SNR* within one window. Panel (A) is a spectrum from 9250 to 10250, the blue line in panel (B) is the corresponding *SNR*; the red line is the calculated locally varying *SNR* limit.

Each *SNR* value has a corresponding *SNR* limit value; the *SNR* limit value is based on the trend of the *SNR* for nearby 1000 data points. In Figure 28, to get the *SNR* limit value at time 9750, we first sort the *SNR* of  $\pm 500$  points, and fit the log of the low *SNR* values as shown in panel (C), this produce a  $\ln(\text{SNR})$  limit value of -0.4, so that the *SNR* limit value is 0.67 for data point at 9750.

## **Results**

Even though the signal to noise ratio is artificially high in some tail regions, the changing *SNR* limit also becomes a little higher to avoid the picking of spurious peak. When we go back to the region with high *SNR* at 10050 to 10200 at time as shown in Figure 29, our *SNR* limit value (red line in panel B) is high enough to pick the “right” peaks.

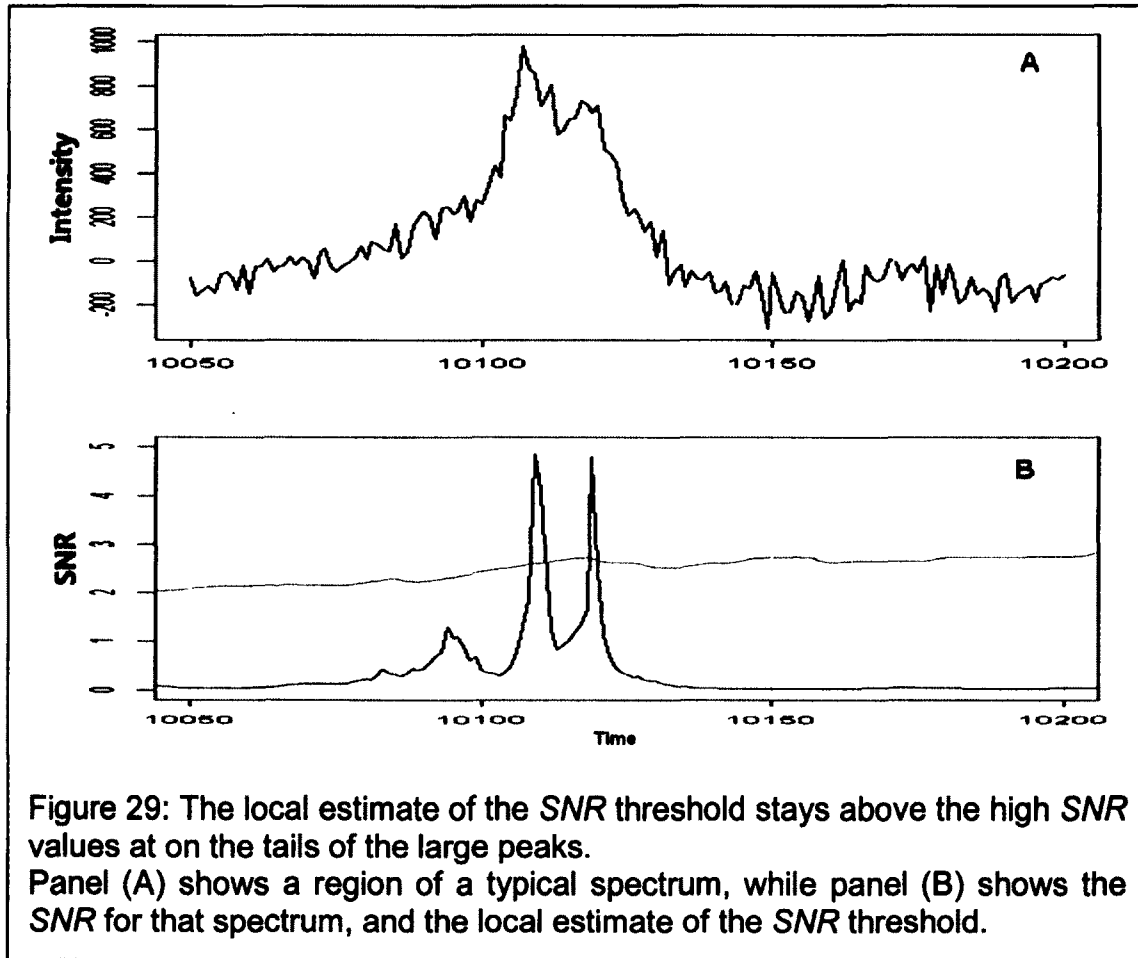
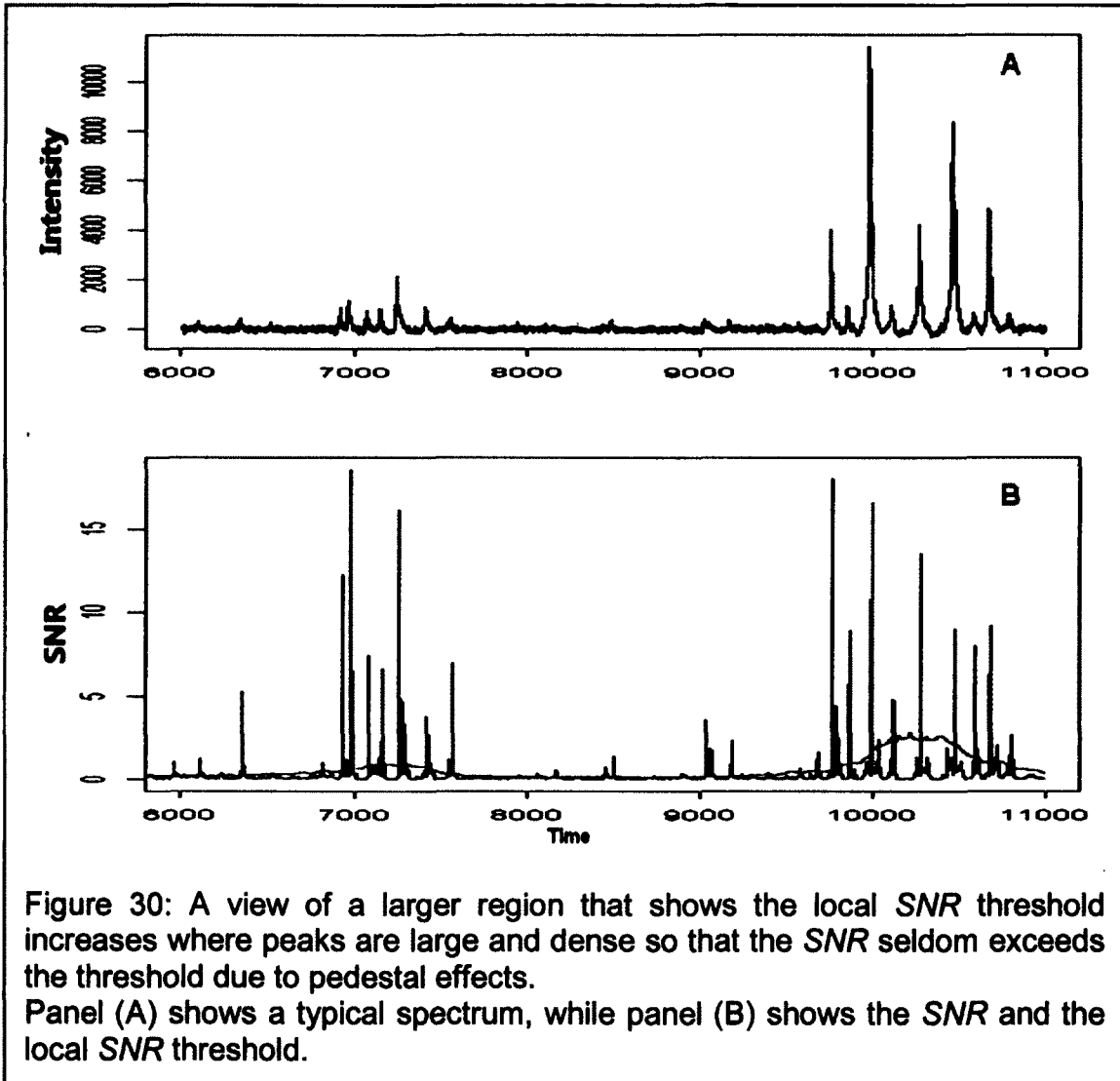


Figure 29: The local estimate of the *SNR* threshold stays above the high *SNR* values at on the tails of the large peaks. Panel (A) shows a region of a typical spectrum, while panel (B) shows the *SNR* for that spectrum, and the local estimate of the *SNR* threshold.

What is more, we could also see how well our *SNR* represents the signal fluctuations. Only looking at the original spectrum, even we may see that there are signs of two peaks. Due to the overlapping of peak tails, the structure of two peaks is no clear. However, when looking from the corresponding signal to noise ratio, the double peak structure becomes much more distinctive. In a word, our signal to noise ratio not only reflects the peak information in our spectrum, but also clarifies the overlapping peak structures.

Looking at the whole region of the spectrum as shown in Figure 30, it is clearly demonstrated how the *SNR* limit automatically changes with *SNR*. To the region



with large signal, or large SNR, e.g. time region from 9500 to 11000, we also have a large SNR limit, which avoided picking up too many spurious peaks. While at the same time, our automatically changing SNR limit value becomes small for a region with small SNR value in order to pick even very small peaks in the spectrum.



With the improved algorithm, the changing *SNR* limit based on floating-baseline peak picker, we can confidently identify very small signals in the spectra, even those which are almost hidden in the noisy background.

### 3.5. Maximum Likelihood Method

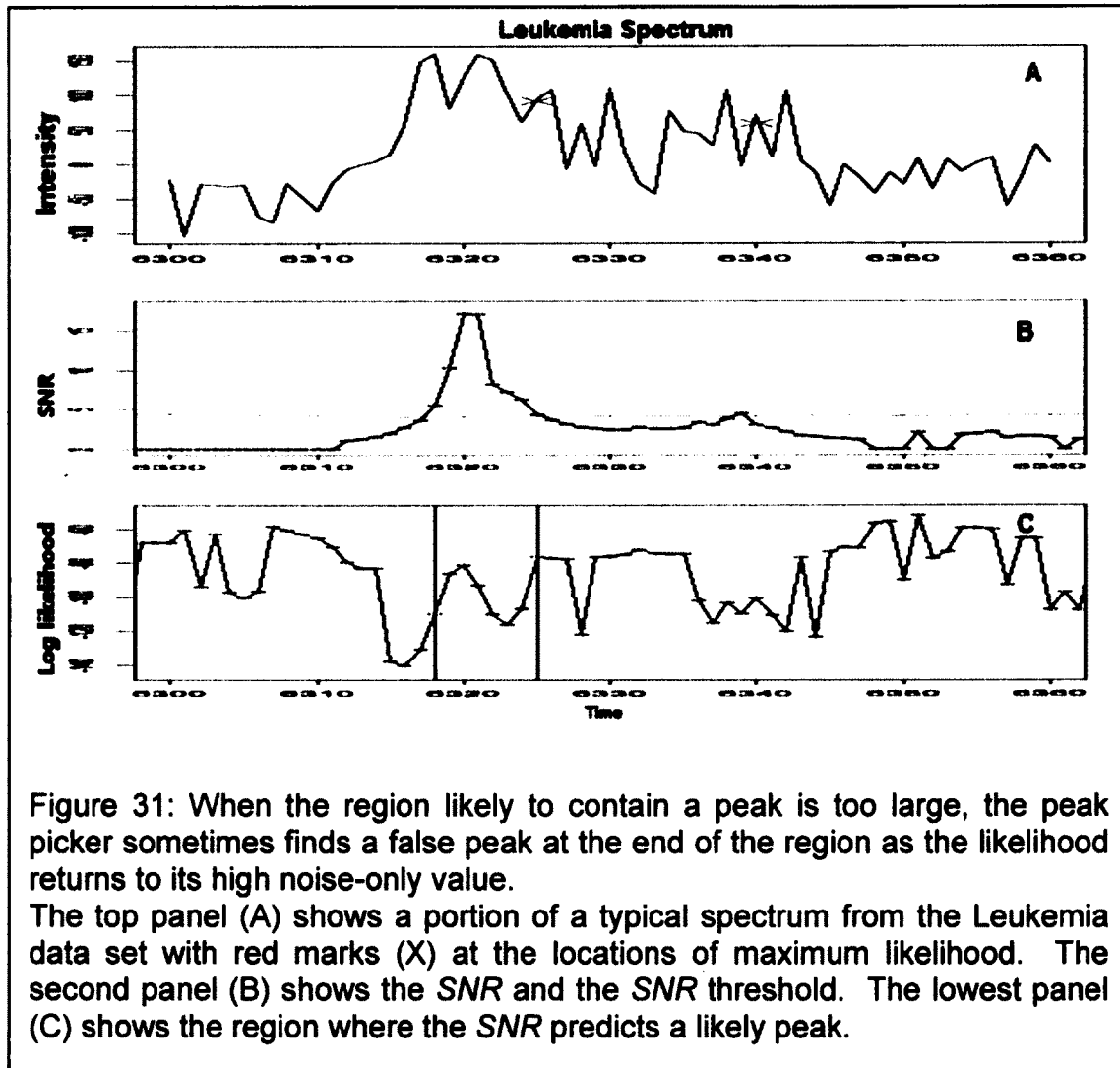
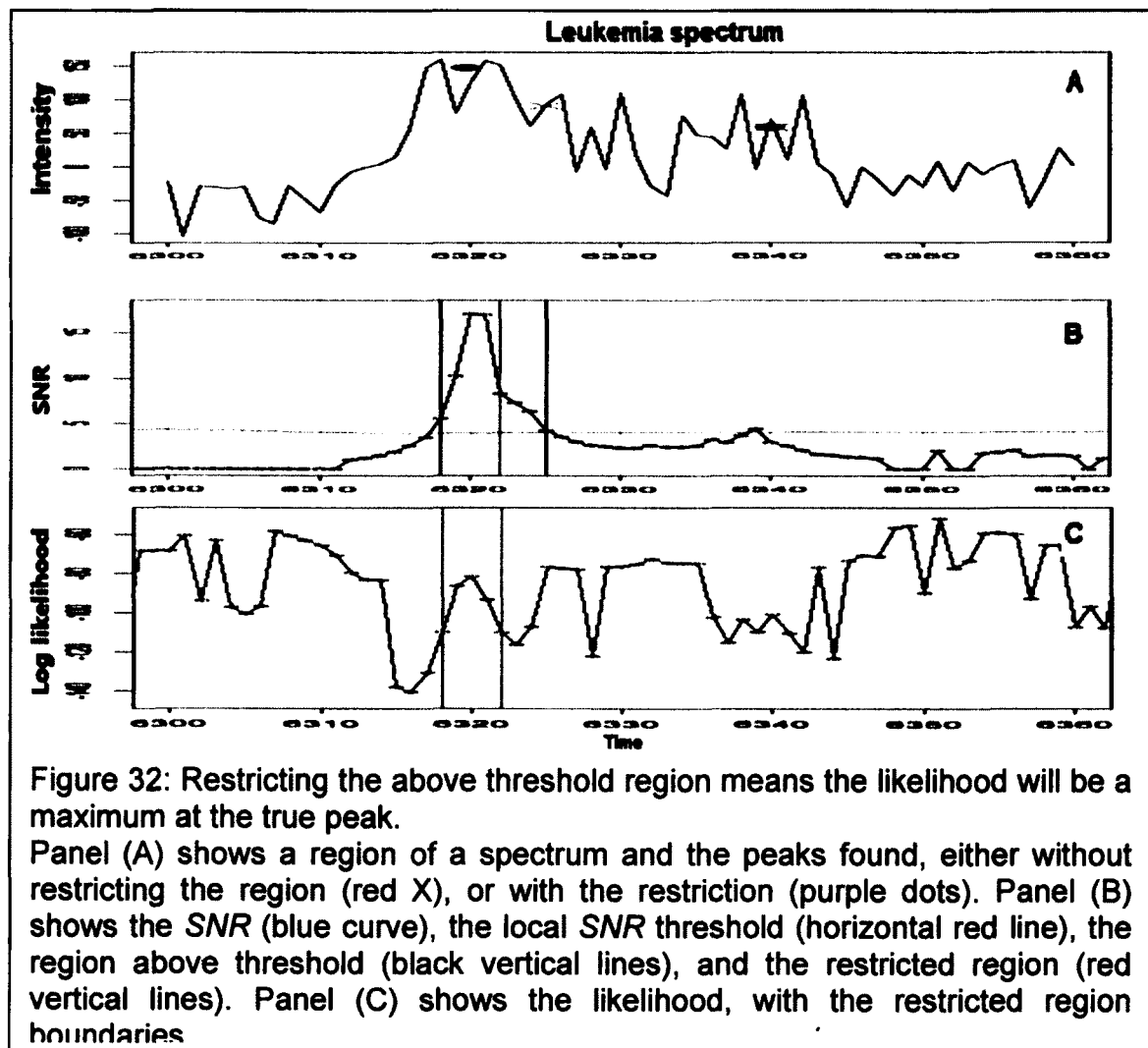


Figure 31: When the region likely to contain a peak is too large, the peak picker sometimes finds a false peak at the end of the region as the likelihood returns to its high noise-only value.

The top panel (A) shows a portion of a typical spectrum from the Leukemia data set with red marks (X) at the locations of maximum likelihood. The second panel (B) shows the *SNR* and the *SNR* threshold. The lowest panel (C) shows the region where the *SNR* predicts a likely peak.

Even with the varying local *SNR* threshold, the likely peak region for some small peaks can be too large. For example, Figure 31 shows a peak picking example where the very small peak near 6320 time steps (in panel A) has noise that

produces an apparent shoulder on the amplitude estimation (in panel B). Because this shoulder then puts the *SNR* threshold above the limit (red line in panel B), the likely peak region extends over into the region best fit by noise alone. Furthermore, the likelihood is maximum at the edge of the region, where the signal is mostly noise. The net result is that the peak picker identifies the edge for the region as the peak location (identified by a red X in panel A). To correct this, we have reduced the likely peak region to a window of a single



FWHM centered on the maximum *SNR* in the region.

Figure 32 shows the result of limiting the likely peak region. This time, by limiting the possible peak region to half FWHM around the strongest *SNR* value, we dramatically reduced the uncertainty in the peak picking. Panel A again shows a portion of the spectrum, with markers showing the original peak position (red X) and the improved position (purple dot) resulting from the reduced likely peak region where the high likelihood of the pure-noise region has been excluded.

## Summary

Our peak picking algorithm reduces a pure spectrum to a list of the important peak positions and amplitudes in an accurate and automatic way. The work flow for finding peaks is as follows:

1. Read in raw spectra data, and set peak picking information: FWHM of peaks = 10, peak picking range from 5800 to 11000 in time.
2. Prepare for baseline subtraction: put a sliding window isolating  $N$  ( $N=FWHM$ ) data points throughout the whole spectrum. Calculate *SNR* and likelihood values based on a “floating baseline” signal model within each window. Get reported peak information including peak positions and peak intensities.
3. Construct the baseline based on smoothed background without reported peak points from step 2. Subtract baseline.

4. Put a sliding window isolating  $N$  ( $N=FWHM$ ) data points at each point throughout the spectrum without baseline from step 3. Calculate  $SNR$  using “non-floating” signal model. Estimate  $SNR$  limit based on local  $SNR$ .

5. Identify possible peak regions by comparing  $SNR$  to local  $SNR$  limit value from step 4.

6. Find peaks in region identified from step 5 and extract peak information by maximum likelihood method.

After all of these steps, peak positions and intensities together with the uncertainties are reported. A detailed peak picking result from a leukemia spectrum is shown in Figure 33. Red Xs on plot represent calculated peak positions and intensities.

7. Get a master peak list of all reported peaks using alignment method (details of the alignment method is in chapter 4 ).

8. Based on the master peak list from step 7, correct the local baseline when necessary, for example when peaks from the master peak list are too close to each other. Adjust reported peak intensities and positions according to the corrected local baseline.

After all of the steps, a peak list containing peak position, peak intensities and uncertainties is ready to use.

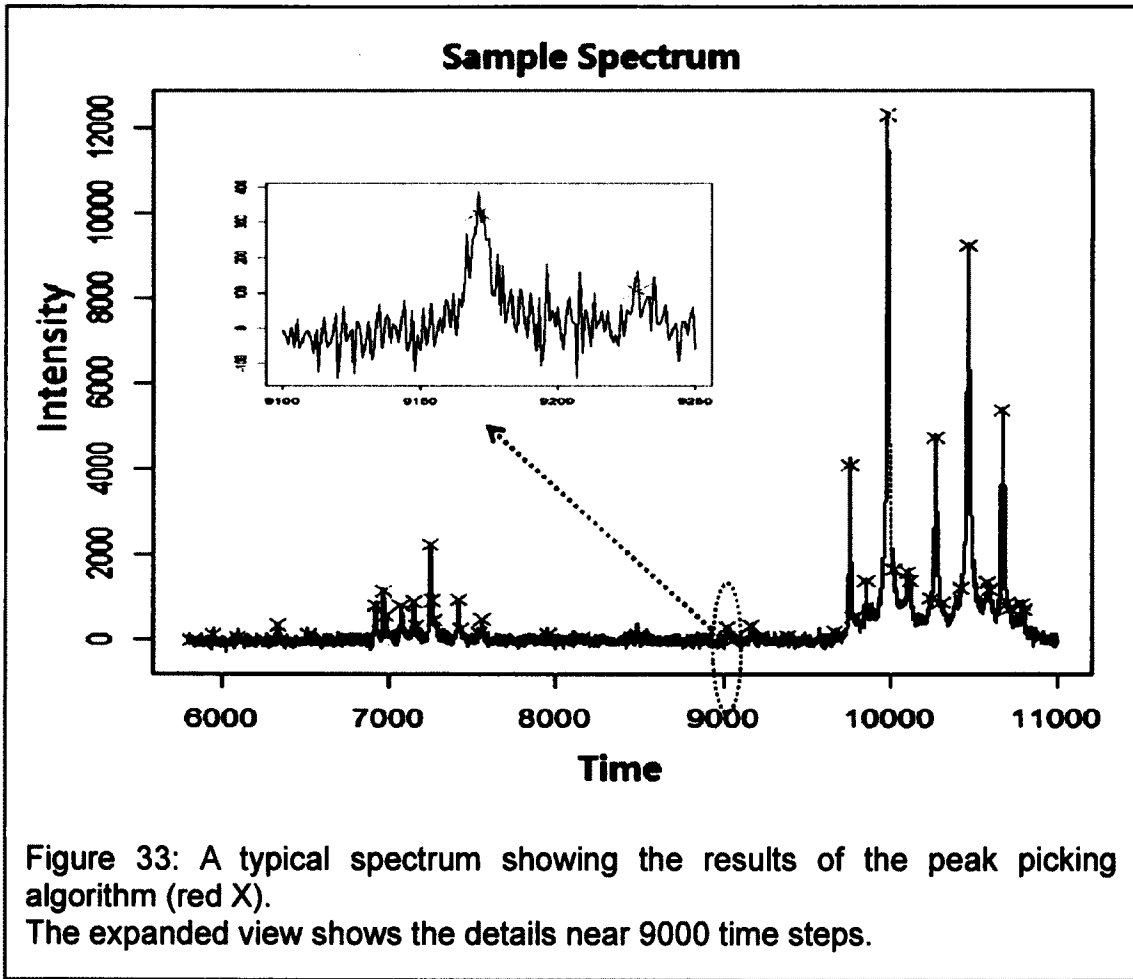


Figure 33: A typical spectrum showing the results of the peak picking algorithm (red X).  
The expanded view shows the details near 9000 time steps.

# Chapter 4 Peak Alignment

## 4.1. Introduction

Peak detection reduces a single spectrum to its essential information about the important proteins and protein fragments within that sample. However, a biomarker discovery experiment will typically include hundreds of samples; each will then be summarized by its own list of peak positions and amplitudes. Peak alignment is the process of correlating the peaks found in any one spectrum to those found in the other spectra. If the spectral data were noise-free and entirely reproducible, this process would not be difficult; however, three things complicate this process: (1) spectra taken at different times may not be precisely calibrated; (2) experimental error will introduce some slight shift in each peak positions; and (3) some peaks will be false peaks, *i.e.* they will have been identified by the peak picker, but they will not represent true peaks associated with proteins or protein fragments. Our alignment procedure is designed to minimize each of these issues.

The calibration problem can give rise to two separate effects. If the electronic start signal is varies from the true laser firing time (usually due to electronic noise generated by the laser firing), then each spectrum might have a time shift compared to the others. Our alignment procedure begins by adjusting this so

that the most common peaks line up with each other. We typically find time shifts of about two time steps (and as large as 5 time steps) to be necessary. The other likely calibration problems are associated with small changes in the acceleration voltage. This would show up in the spectra as a scaling necessary to match the largest peaks at long times (or high  $m/z$ ) when the short time peaks have already been aligned by a start time adjustment. We have found no evidence of this in the data set discussed here, so we have not allowed any scaling correction in our current alignment procedure.

To deal with the other two issues, experimental peak position error and false peaks, we use a procedure that chooses which locations are most likely to represent clusters of peaks that are common in many spectra. This represents the bulk of the effort in alignment, because it requires an automated method of deciding when any given peak is sufficiently close to be said to belong to one cluster rather than another cluster. This is especially difficult when the other cluster has not yet been identified.

## **4.2. Method**

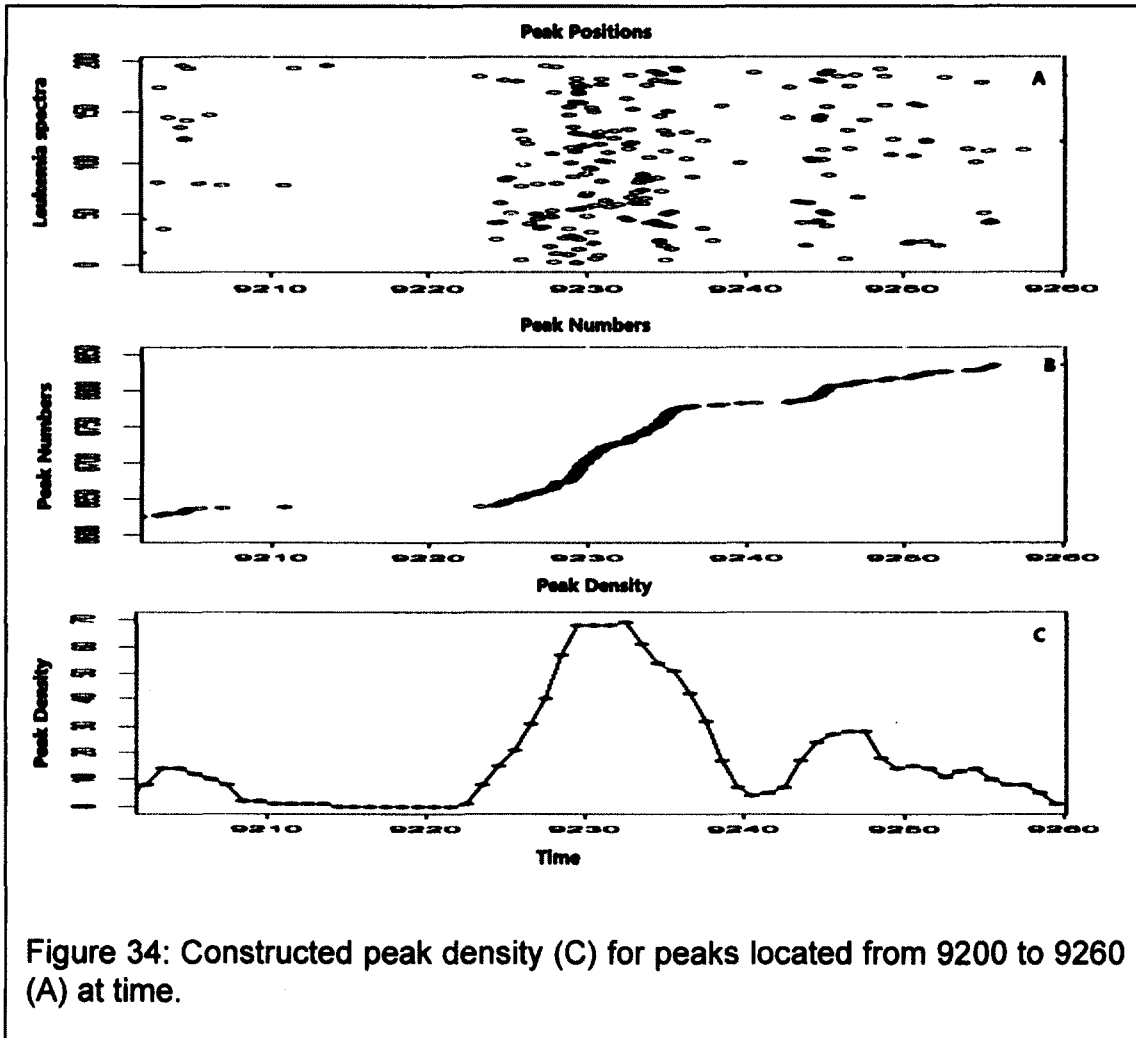
We use an alignment routine to match peaks representing the same  $m/z$  from different samples based solely on their positions. We do not use peak amplitudes because they are expected to have a wide variation among different patients. Not all peaks will eventually be associated with the final list of peaks. If a peak only appears in a small number of samples, we will eliminate it as either a

false peak (representing a coincidental noise fluctuation that looked like a peak) or a peak that is specific to only a very small number of patients. In either case, it is not a good candidate for a biomarker, and so we will eliminate it from our final master list.

There are two major sub-steps in this alignment procedure. The first step is to adjust the starting time value for each spectrum. The second step is to then group the high peak density regions to determine the final master peak positions based on the shifted peaks. Both steps use window bins to identify likely peak locations from the maxima in the density of peak positions. We set window bins to define each individual feature, i.e., common peaks which should be found at the same group of spectra. We use peak density values as reference to mark the place where windows bins should be set. Our general procedure is to set window bins at high density regions, and then subdivide those regions until they only contain peaks belonging to the same group.



The top panel of Figure 34 shows an example of listing the detected peak



positions for the region between 9200 and 9260 time steps for the leukemia samples in the data set. We sort all the peak positions to get the total peak numbers at times as shown in panel B, then do an interpolation to get regular values for every time click, smooth the result by integrating over half of a FWHM. The derivative is the constructed peak density as shown in panel C. As we can see from Figure 34, high peak density value, e.g., 70 at 9230, indicates large

number of peaks clustered at these time points, thus could be used as the reference to set window bins.

Window bins are first determined by the peak density. Likely peak cluster regions are placed where the peak density is above a manually introduced threshold. If the window bin is larger than a FWHM, then more than one group of peaks are likely in the same window bin, so the bin should be subdivided. We subdivide the bin by putting a smaller bin of  $\pm\text{FWHM}/2$  around the maximum density value. We repeat this process to further subdivide the remaining part of the high density region if necessary.

After setting window bins to the desired size, the next step is to check peaks within each window; and choose only one peak if two peaks from the same spectrum are assigned to the same group. Finally, we eliminate any bins where too few spectra contribute peaks. The master peak list is then the average position of all measured values in each bin.

The master peak list gives us a summary of which peaks should be present in the group of spectra. Some spectra may not have detected a peak at each of the master list positions. For those spectra, we record the average signal intensity at the expected peak position to generate a complete data set.

### **Starting Time Shift**

To correct the start time shift, we generate a peak list based on the most commonly occurring peaks, and get the starting time shift values by a least square fitting to minimize the differences among these peak positions.

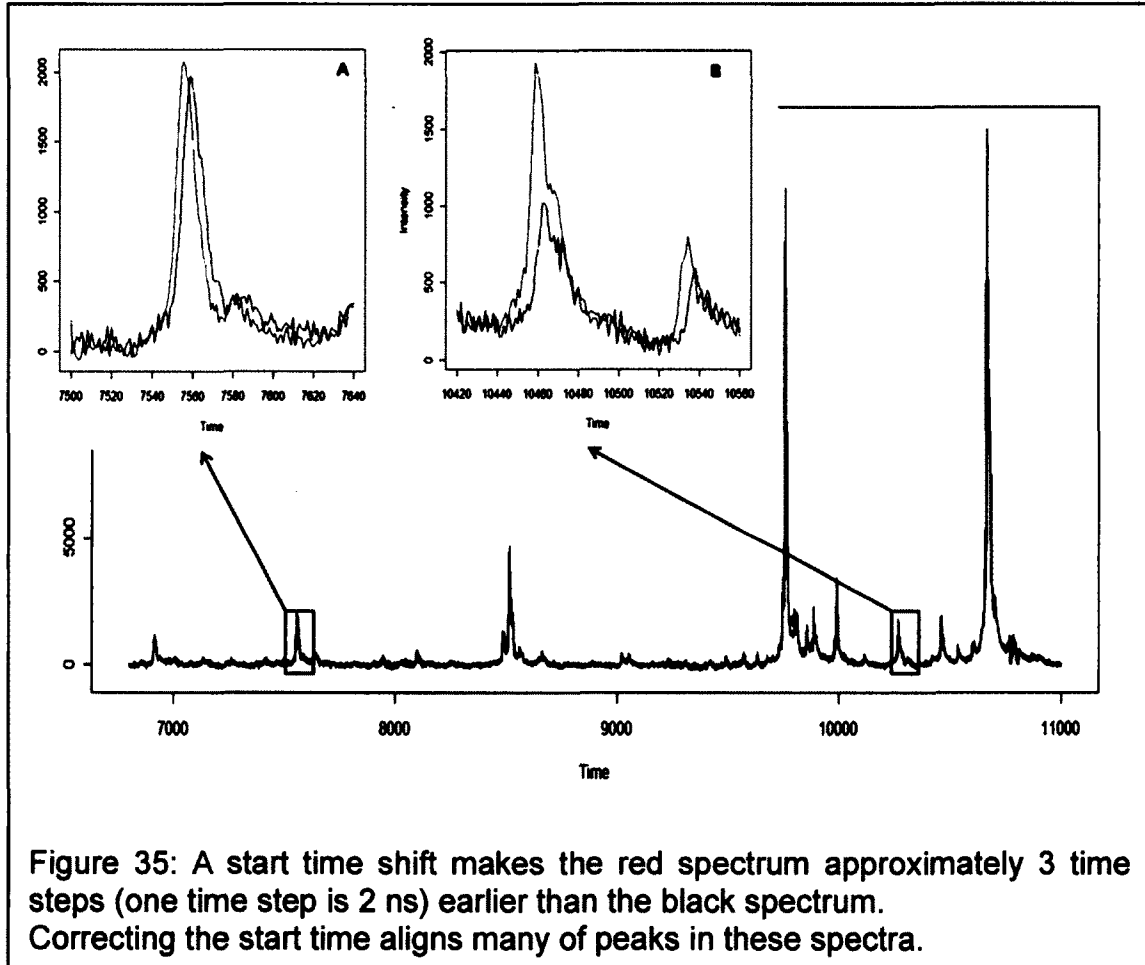


Figure 35 shows an example of a starting time error between two spectra. The two spectra are plotted in different colors (red and black) in the region between 6800 to 1100 time steps. The insets show magnified views of the spectrum region from 7500 to 7640 (panel A) and from 10420 to 10500 (panel B). In both panels, the two spectra look similar, but the red spectrum is shifted approximately

3 time steps early. Shifting the start time of one of these spectra aligns many of the features of both spectra.

The starting time shift values for the 196 leukemia spectra are shown in Figure 36. For these spectra, the starting time shift ranges from -2.5 to 4.5. Note that although the spectra are measured at integer time steps, the net starting time shift is not an integer. This effectively removes the  $\pm\frac{1}{2}$  step measurement error introduced by digitizing the time.

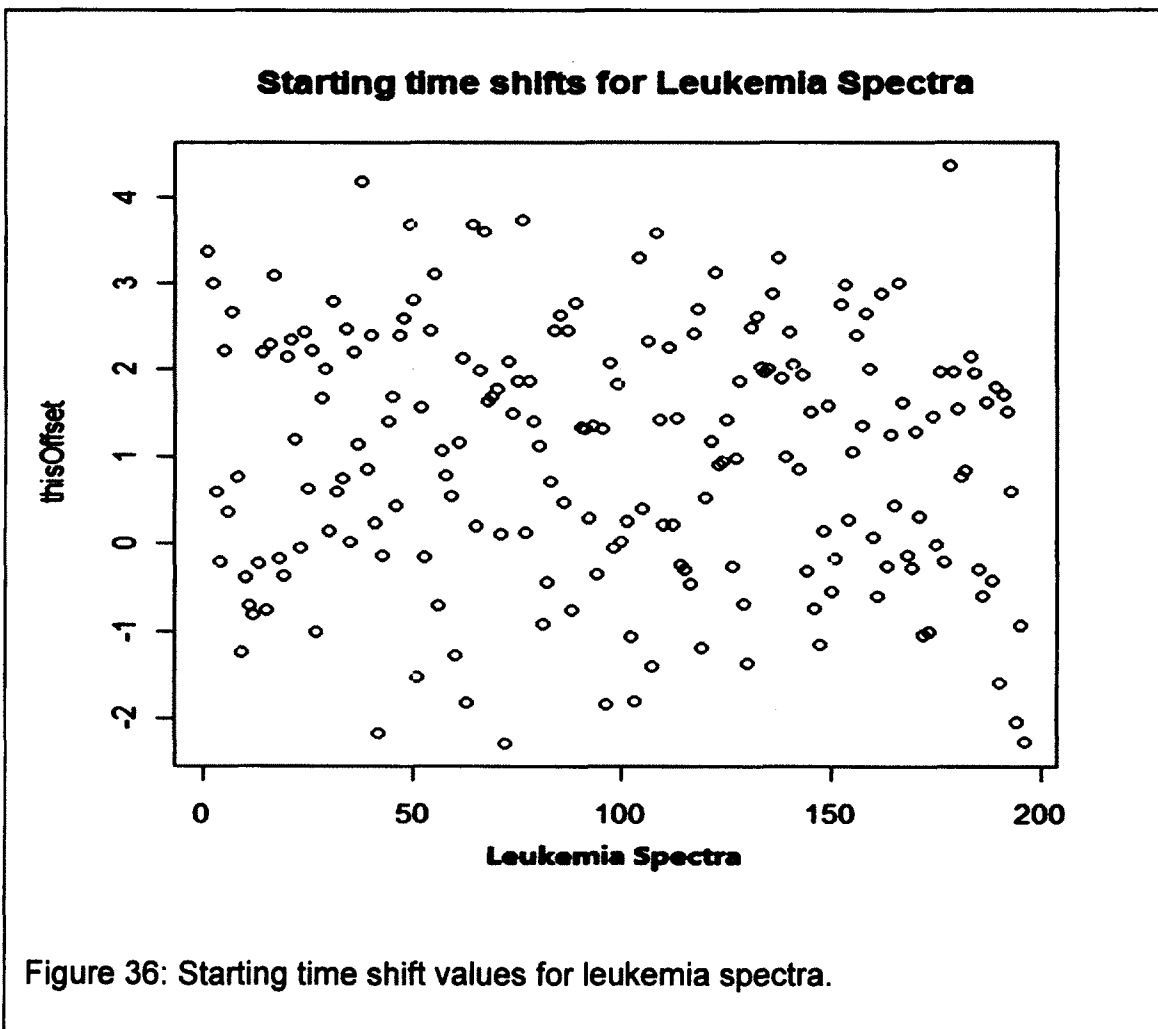
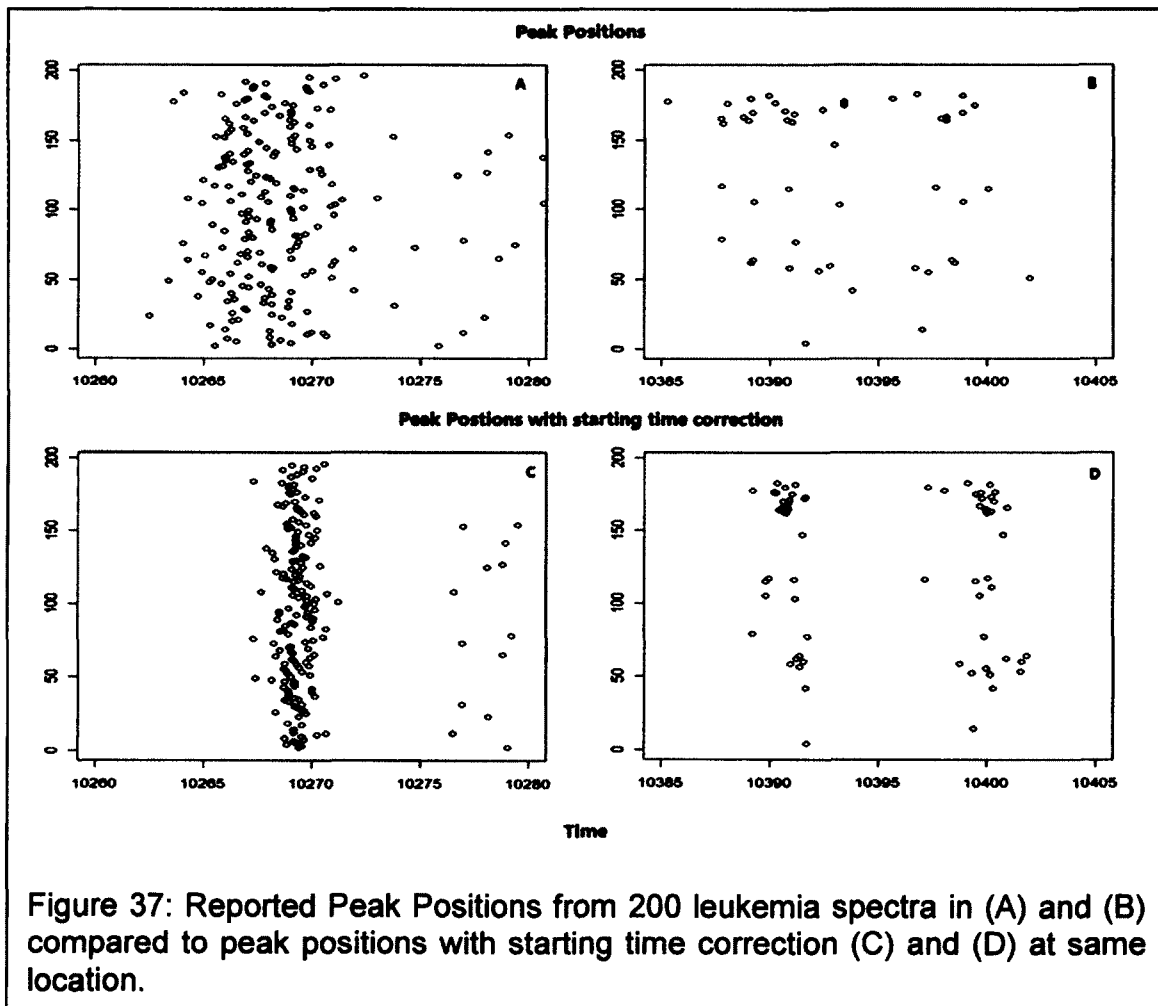


Figure 36: Starting time shift values for leukemia spectra.

This algorithm only uses the peak position data developed by the peak picking algorithm described in earlier chapters and requires no other information (such as an expected target spectrum, for example). However, it could also be manually supervised by selecting a specific set of reference peaks for alignment, if a standard set does exist throughout the full spectra range. We have found this unnecessary, although the peaks used for calibrating the machine could have served that purpose. Using our start-time adjustment method, peak misalignment

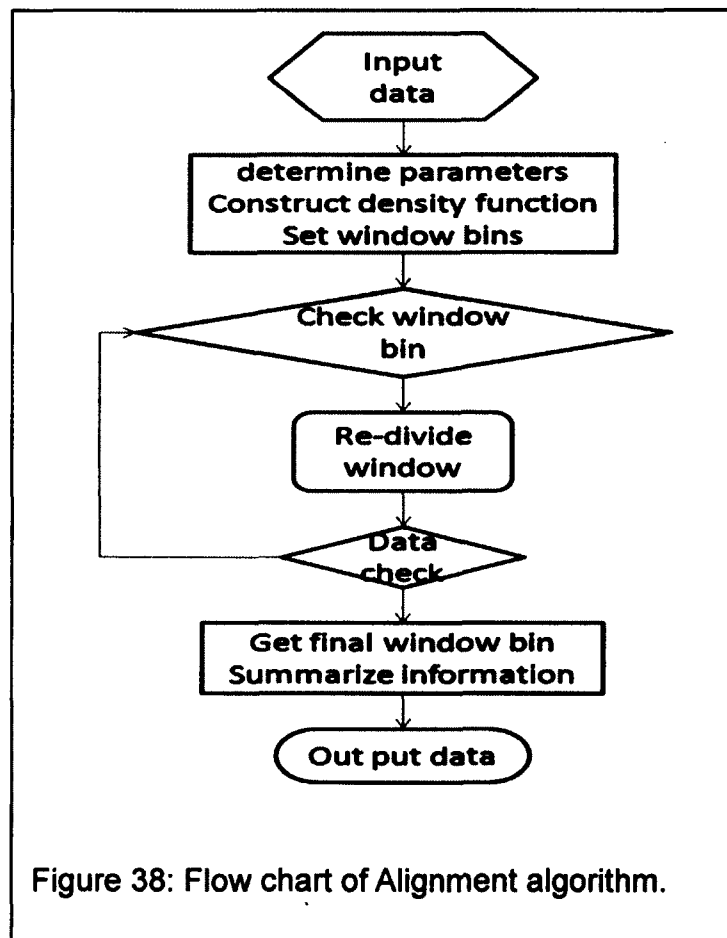


was reduced from about 1-2 time steps to about half a time step, globally.

Figure 37 shows the peak positions for the leukemia spectra in two regions near 10,300-10,400 time steps before (top panels, A and B) and after (lower panels, C and D) correcting the start times. The peak around 10270 is used as one of the reference peaks for the time shift alignment, but the other peaks were not. The tight bunching the peaks around 10340 and 10400 demonstrate that the starting time adjustment was a valid step.

#### 4.2.1. Peak Binning:

The second step of peak alignment calculates the overall distribution of shifted peaks and dynamically identifies boundaries of regions where many peaks have similar position. Here, we identify and quantify significant features, discarding data that do not seem to be part of a feature. A general overview of the preprocessing and



analysis strategy is shown in Figure 38.

The input data is the shifted peak positions combined from all spectra. A corresponding density function is constructed as the reference to pick the common peaks. Master peaks should be picked from the region with high peak density.

The major difference of this peak alignment step and the starting time correction step is how the threshold is set for picking the high peak density. In the start-time correction step, a set of highly reproducible peak groups should be chosen and set as reference to do the axis shift. Thus, a high threshold should be set to pick those standard peaks which are supposed to be found at most spectra. While in constructing the master peak list, the purpose is to extract as many important features within the data set as possible. Therefore, the peak density threshold value should be set lower than what is used for starting time correction so as to extract the common features among the same group of spectra. For a large data set, we decide on conversion of a peak found in 5% of all the spectra should be considered as a feature to be paid attention to. In contrast, we required peaks used for the time-shifting part of alignment to appear in at least 20% of the spectra.

Window bins are used to separate the high peak density region from low peak density region. However, some peaks are difficult to resolve from one another when they are very close to each other and included in the same window. We

may see groups of peaks which are supposed to belong to different master peaks are in the same large window bin. Those large window bins need to be divided, but then two problems arise: first, how to determine a window size is too large; second, what is the condition to halt further subdivision? Any window larger than the full peak width is likely to contain more than one peak, so we check those windows. Then each window bin size is determined based on the peak density within that window. Starting from the highest peak density value in the distribution within the first pass of window, we descend down either side of the meta-peak until the distribution increases again. The place where the trend of distribution just changes is the sub-bin setting point. This process defines a sub-window bin in which different groups of peaks are set apart; it is flexible enough to adjust for different lengths of sub window bins.

One time subdividing sometimes is not enough for very large windows; examining the remaining part within window is necessary. If the remaining part is still considered to be too large, second subdividing will take place. We perform the process in a loop until no further subdividing is necessary.



Once all the beans have set, we first check that each bin includes at most one peak from each spectrum, and then eliminate any bins that represent peaks that occur in 5% or fewer from the total samples, according to the previous mentioned criteria.

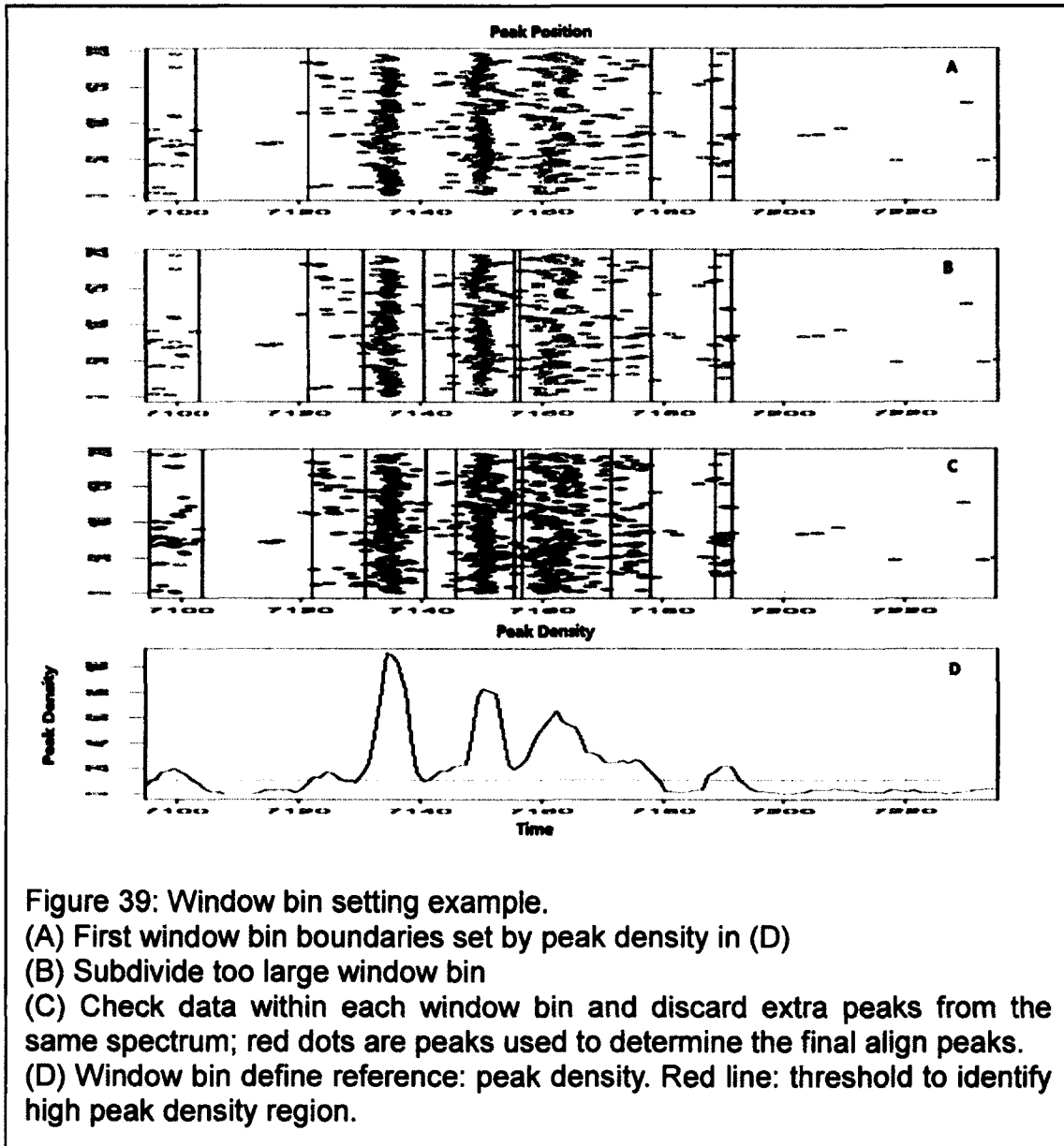
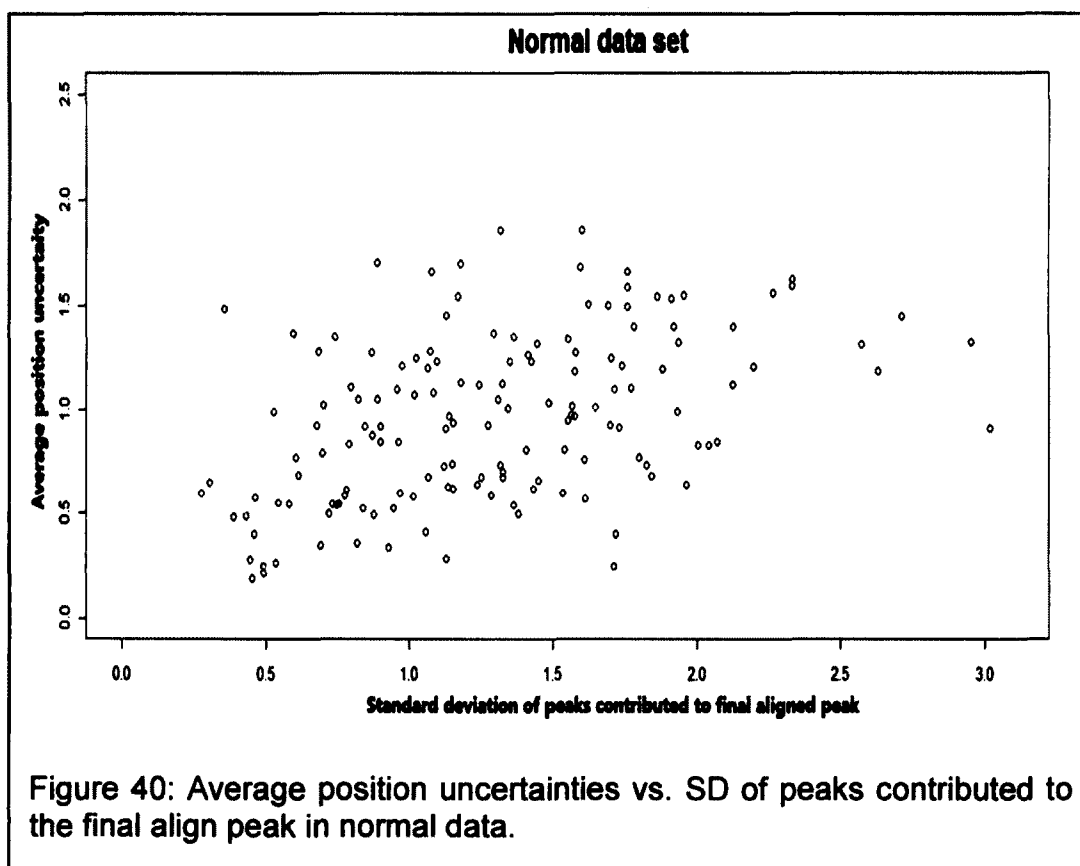


Figure 39 is a detailed example about the window bin setting process for a high peak density region from 7120 to 7180, panel D is the corresponding peak density for this region. Black dots in panel A, B and C represent peak positions from all the spectra, vertical blue lines are the left bounds for window bins while green lines are right bounds. A blue line and a green line compose a window bins. Panel A shows the result of setting the window bins after choosing high peak density region, the first window bin is 60 time step wide, this is too large compared to the desired window bin size 10. Subdividing this window bin is necessary. We subdivide this window bin by putting a smaller window  $\pm 5$  time steps around maxima density value within the large window. Re-subdivision procedure is repeated within this large window bin until each window bin contains only one master peak, the final window bins are set as shown in panel B. Next is to identify peaks that were present across most spectra within the same group and make sure that only one peak from same spectrum is included in the same peak group. The red dots in panel C are those peaks. Discarding peaks that were selected from few spectra, a final master peak list is generated based on the peaks (red dot in panel C) within each window.

Finally, peaks must then be matched across spectra to allow calculation of time deviations and recording of intensity information, we assume the closest peak is the correct match. Additionally, the algorithm determines which spectra are missing from each peak group, and then raw data are integrated to fill in intensity values for each of the missing peaks.

Figure 40 shows the relationship of average reported position uncertainties and standard deviation of peaks contributed to final master peaks for each align peak. Small peak uncertainty is related with small standard deviation as a high confidence in the peak finding.



### 4.3. Summary

We have developed a peak alignment procedure that aligns peaks in different SELDI TOF-MS spectra. Several steps are taken to construct common peaks from peak list; a detailed flow chart is shown in Figure 41:

**Major steps are:**

- 1. Prepare data:** combine all the picked information in an ordered way, and then construct peak density function as a measure for peak reproduction.
- 2. Correct starting time shift:** select the most well behaved common peaks as standards to shift each spectrum to correct errors (usually of only a time step or two) in the start time signal which can result from the electronic noise created when the laser fires.
- 3. Get master peak list:** determine the final master peak list based on the "shifted" spectra. Set window bins to group near-by peaks. Window bins are divided several times until the window size is sufficiently small and each window contains only a single feature. Finally, we have identified peaks that were present across most spectra within the same group while also avoiding peaks that were selected in few spectra.
- 4. Complete peak information:** determine master peak positions according to peaks within each window bin, add peak intensity value for picked master peaks or fill in peak intensity value for missing picked master peak in each spectrum. Get metrics, such as X and Y, of a peak group that will serve as features.

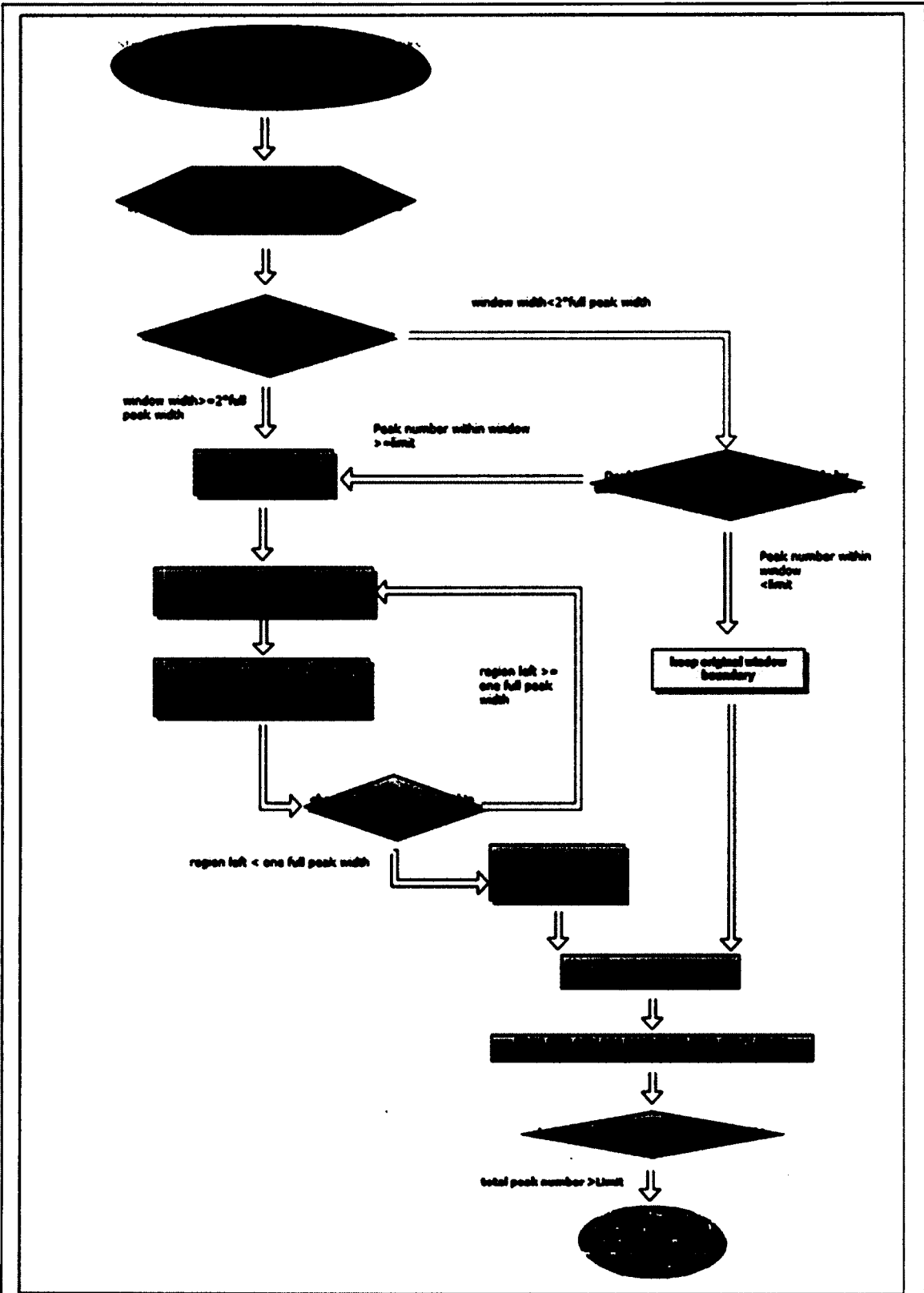
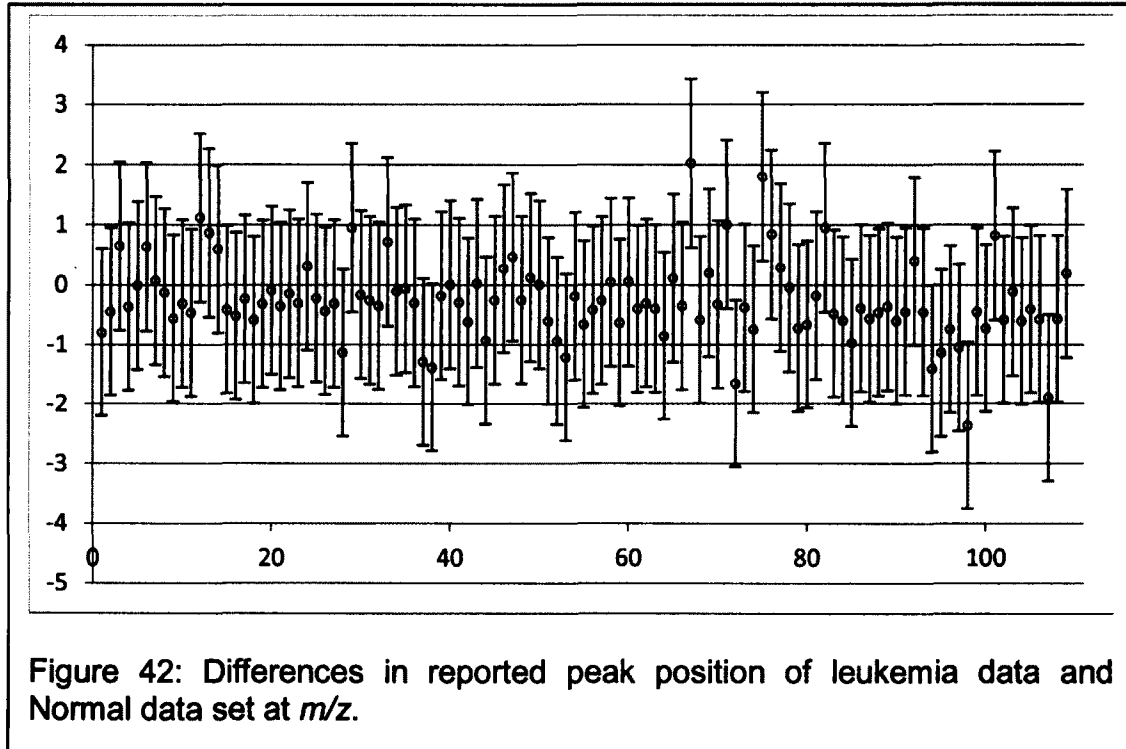


Figure 41: Overview of alignment

After the alignment process, we will have a master peak list representing all the spectra in a uniform manner. After processing a set of  $n$  spectra and identifying and quantifying  $p$  peaks per spectrum, we are left with an  $n \times p$  matrix of peak expression levels.

There are 147 final master aligned peaks in the EVMS leukemia spectra data while there are 157 final master aligned peaks in normal spectra data. 110 master aligned peaks are considered to be the same peaks as the difference in position is within 3 time lag. The difference of the Intensity of these similar peaks is range from 0.34 to 2.08. Detailed data could be found at Table 1.

Figure 42 shows the difference of position shifts in  $m/z$  between the same peaks from leukemia data and normal data. The difference in positions ranges from -2



to 2 at  $m/z$ , very small and thus these peaks could be considered as the same ones. The two sets of data are generated independently based on the method we have introduced; it shows the power of our method to capture the important information from raw spectra.

Leukemia		Normal		Diff in Position(m/z)	Ratio in Intensity
Position(m/z)	Intensity	Position(m/z)	Intensity		
2758.88	114.6141	2760.47	179.261	-1.59254	1.56
2798.71	434.5866	2799.61	149.0417	-0.89746	0.34
2861.26	155.5834	2859.96	137.2284	1.292487	0.88
2874.7	205.0836	2875.44	202.0462	-0.7364	0.99
2961.72	192.5645	2961.74	162.3469	-0.02271	0.84
2982.16	126.5191	2980.89	211.525	1.265057	1.67
3001.7	114.1613	3001.56	83.61061	0.138164	0.73
3023.7	307.8199	3023.96	206.0026	-0.26139	0.67
3038.7	296.315	3039.82	171.8144	-1.12022	0.58
3044.88	94.02478	3045.51	133.5147	-0.63074	1.42
3137.68	343.5196	3138.62	186.0019	-0.93606	0.54
3155.35	144.7113	3153.11	207.3843	2.235595	1.43
3160.75	106.7598	3159.02	137.5683	1.734803	1.29
3165.98	197.3618	3164.79	228.2176	1.19429	1.16
3249.69	273.7903	3250.53	141.2753	-0.83471	0.52
3268.86	837.6403	3269.9	365.8684	-1.04259	0.44
3283.74	1179.807	3284.22	548.675	-0.47931	0.47
3298.11	609.645	3299.28	253.0078	-1.17543	0.42
3329.07	131.7423	3329.71	102.6225	-0.64384	0.78
3381.75	240.826	3381.94	313.5202	-0.19309	1.3
3442.87	82.76032	3443.59	118.2315	-0.72077	1.43
3452.5	309.3272	3452.81	444.3109	-0.30491	1.44
3495.75	239.7732	3496.36	295.108	-0.61686	1.23
3610.46	218.7629	3609.85	115.2803	0.614855	0.53
3722.81	133.3003	3723.26	156.0588	-0.44837	1.17
3784.01	421.2357	3784.89	308.7442	-0.87398	0.73
3827.7	167.455	3828.33	209.4019	-0.6309	1.25
3897.41	699.3287	3899.69	1058.445	-2.27261	1.51
3920.49	259.7456	3918.59	357.0874	1.900448	1.37
3951.28	1207.315	3951.61	814.3245	-0.33191	0.67
3967.3	1232.728	3967.82	788.4412	-0.52116	0.64
3983.13	1043.489	3983.83	738.0126	-0.70314	0.71
3996.9	357.8637	3995.47	160.694	1.434912	0.45
4076.31	821.7408	4076.53	1138.315	-0.22359	1.39
4145.76	252.0529	4145.89	247.781	-0.13085	0.98
4163.72	427.8508	4164.32	415.0337	-0.60885	0.97
4178.16	306.0153	4180.73	303.3399	-2.57715	0.99



Leukemia		Normal		Diff in Position(m/z)	Ratio in Intensity
Position(m/z)	Intensity	Position(m/z)	Intensity		
4190.81	347.2302	4193.57	374.4131	-2.76553	1.08
4278.61	2002.029	4278.97	1324.663	-0.36675	0.66
4294.57	1564.517	4294.57	1048.162	0.001189	0.67
4311.52	1049.809	4312.1	582.5834	-0.58501	0.55
4480.05	744.3882	4481.28	658.9594	-1.23645	0.89
4627.75	455.3989	4627.7	321.4798	0.046906	0.71
4639.25	216.5044	4641.11	195.9052	-1.85776	0.9
4656.12	841.4731	4656.64	866.3903	-0.52004	1.03
4688.88	203.4551	4688.34	210.3111	0.546688	1.03
4697.82	149.995	4696.88	150.7374	0.936453	1
5033.57	112.7449	5034.08	109.0638	-0.50982	0.97
5092.69	239.6647	5092.45	191.238	0.240855	0.8
5118.25	226.4116	5118.23	156.3084	0.010385	0.69
5145.67	151.258	5146.88	171.8362	-1.20801	1.14
5258.19	202.2392	5260.07	189.4216	-1.87652	0.94
5276.43	162.2263	5278.84	195.9256	-2.41766	1.21
5349.46	774.1606	5349.84	711.7508	-0.38357	0.92
5364.75	364.6307	5366.07	311.505	-1.32651	0.85
5430.95	129.9867	5431.78	109.2275	-0.83716	0.84
5556.54	156.5283	5557.07	163.8407	-0.5316	1.05
5816.15	293.0885	5816.07	282.1563	0.083034	0.96
5878.54	694.3221	5879.82	286.3004	-1.27311	0.41
5887.75	353.9223	5887.66	255.6022	0.094682	0.72
5898.02	379.1293	5898.82	203.0718	-0.80275	0.54
5917.36	2722.283	5917.97	2681.003	-0.617	0.98
5932.48	1404.471	5933.28	1309.8	-0.80089	0.93
5954.54	385.2366	5956.25	291.1736	-1.71126	0.76
6104.36	266.1195	6104.13	261.928	0.22259	0.98
6123.29	426.1981	6124	388.1693	-0.71001	0.91
6139.89	344.0095	6135.85	277.2943	4.047505	0.81
6201.31	287.9949	6202.49	236.7808	-1.17613	0.82
6447.41	142.1659	6447.02	223.4671	0.396324	1.57
6644.98	332.4379	6645.64	280.0892	-0.65533	0.84
6662.05	171.1186	6660.04	181.8805	2.011262	1.06
6677.55	198.8523	6680.85	270.3919	-3.3013	1.36
6691.59	330.9991	6692.36	363.3238	-0.76154	1.1
6865.68	454.2204	6867.18	273.2295	-1.49169	0.6

Leukemia		Normal		Diff in Position(m/z)	Ratio in Intensity
Position(m/z)	Intensity	Position(m/z)	Intensity		
6957.89	253.325	6954.28	263.4481	3.603654	1.04
7200.6	544.8441	7198.92	1053.266	1.679915	1.93
7358.91	264.5894	7358.33	403.7622	0.580647	1.53
7412.56	217.3874	7412.66	173.4448	-0.09999	0.8
7448.1	116.1127	7449.54	165.3601	-1.44405	1.42
7484.2	426.3862	7485.53	789.6282	-1.32665	1.85
7578.64	394.0007	7579.01	819.2108	-0.36494	2.08
7652.03	453.7857	7650.13	456.0407	1.903218	1
7740.07	508.3908	7741.03	514.1341	-0.96115	1.01
7936.86	1151.48	7938.05	1243.361	-1.18833	1.08
7962.11	631.2532	7964.05	436.1237	-1.94003	0.69
7985.22	1033.612	7986	1027.146	-0.77678	0.99
8000.94	701.9094	8002.07	888.5049	-1.13306	1.27
8140.58	8633.804	8141.5	11898.88	-0.92021	1.38
8154.59	2120.05	8155.32	1780.074	-0.72926	0.84
8201.15	1926.563	8202.35	2834.806	-1.19974	1.47
8362.8	1135.008	8363.69	1569.755	-0.89618	1.38
8579.01	596.561	8578.22	992.7151	0.789991	1.66
8616.35	4243.643	8617.26	2915.262	-0.90967	0.69
8677.79	1197.021	8680.6	732.7147	-2.80649	0.61
8687.41	533.1212	8689.69	510.8203	-2.28362	0.96
8877.39	1045.431	8878.88	822.5677	-1.48764	0.79
8897.92	772.6767	8900.03	903.518	-2.10568	1.17
8906	898.5214	8910.71	875.9717	-4.71031	0.97
8947	6261.814	8947.91	5088.069	-0.9098	0.81
8960.1	3392.297	8961.56	2857.034	-1.4602	0.84
9005.7	1480.746	9004.06	845.139	1.635991	0.57
9075.27	736.4276	9076.45	786.3785	-1.17563	1.07
9153.46	1578.662	9153.7	1242.531	-0.24308	0.79
9167.6	1114.109	9168.81	1288.234	-1.21635	1.16
9302.56	6899.431	9303.39	6954.589	-0.82268	1.01
9362.77	1023.65	9363.93	964.1195	-1.15899	0.94
9380.72	712.3056	9384.5	617.7523	-3.77938	0.87
9508.64	992.0342	9509.79	500.5067	-1.15029	0.5
9523.79	828.1956	9523.42	756.5106	0.371913	0.91

Table 1: Comparison of the similar master peaks from normal and leukemia data.

# Chapter 5 Conclusion

Biomarker identification has the potential to assist in early cancer prognosis and diagnosis. MALDI TOF MS provides a precise and rapid way of measuring the relative abundances of the protein present in the complex mixture of biological and chemical samples. We have successfully created an automatic algorithm for optimizing the data processing, transforming the raw mass spectrum data into a high quality representation to maximize the use of samples collected from measurement of protein concentrations in biological samples for biomarker discovery.

## *Peak Picking*

We have been using a peak identification algorithm that maximizes the likelihood of data setting in a filtering mode of operation. We restrict the peak region from noisy background by choosing regions with high signal to noise ratio values. A changing *SNR* limit is used as the *SNR* threshold to determine the possible peak region. The threshold value is determined automatically by a survey of the local signal to noise ratio trend. Studying the *SNR* when there are no peaks, we can predict a relationship between the signal to noise ratio and the probability of there being a peak within the window. The survey approach allows this value to change at different regions, reflecting the variations in the level of background noise. What is more, we did not just pick peak position at the maximum value of

likelihood in a possible peak region; we limit the possible peak region to a reasonable size and reject cases when the maximum likelihood occurs at the edge of that possible region. With this improved algorithm, we dramatically reduce the number of false peaks and get more precise representative information from spectrum.

### ***Background Elimination***

In my work, I have developed algorithms to eliminate background artifacts, this is necessary to get a true “zero baseline” spectra prepared for the peak picking step. Our best representation of the background is determined by smoothing the data in those regions where we determined that peaks were unlikely to be found. The peak picker algorithm, mentioned above, easily finds most peaks, except for very small signals which are obscured by the baseline noise. By using this peak-picker to first determine the regions that should be smoothed to get a baseline, the net effect of any small included peaks is negligible. Once the background has been removed, the peak-picker can use a zero background model, which is more stable in regions where the background level had been elevated.

### ***Alignment Process***

Next is our alignment process which aims at getting summary of sample data belongs to the same group. There are two major sub-steps: first we generate a peak list based on the largest signal peaks, and use this list to adjust the zero time value of each spectrum to correct for start signal variations; second, we

generate a master peak list based on all the peaks of the shifted spectra. Both steps use the same general approach, which identifies likely peak locations from the maxima in the density of peak positions, averaged over all the spectra. This requires setting windows to represent the common peak regions, and then subdividing these windows until they are no larger than a desired size, based on the peak width; finally, we eliminate any bins for which too few spectra contribute peaks. From these windows, we then generate a master peak list as the average position of all measured values in those remaining bins. Thus we have reduced the peak list by restricting it to most common and most distinguished peaks using the alignment routine.

### ***Local Background Removal***

However, information about small peaks sitting on the shoulder of nearby large peaks may be distorted by the influence of that large peak; this may lead to an unnecessary focus on those small peaks causing inaccurate reporting of the positions. To correct that influence, we adjust the local baseline around that small peak according to nearby large peak by removing the shoulder or tail effect of the nearby large peak. This task will easily be achieved by adjusting every spectrum locally to the average spectra contributed by large peaks. We know exact which spectra do not claim to have found that small shoulder peak in that region, averaging these spectra will give a good representative of the local baseline introduced by the nearby large peak. Removing the local baseline,

peak picking and alignment process gives a more precise report about the peak information.

### ***Future Work***

The success of the peak picking method relied largely on the peak shape; we use a constant peak width here, which is acceptable within a narrow mass focusing range, however, a changing peak width maybe needed as we push the peak picking beyond the mass focusing range. What is more, our signal model is a simple model of a peak sitting on noisy background. Overlapped peaks may need to use different signal models. It is then necessary to set up different models containing one or two overlapping peaks, and then choose the most likely model.

## Bibliography

- [1] "Cancer Facts and Figures 2012," American Cancer Society 2012.
- [2] Cho, William CS. "Proteomics technologies and challenges." *Genomics, Proteomics & Bioinformatics* 5.2 (2007): 77-85.
- [3] Cho, William. "OncomiRs: the discovery and progress of microRNAs in cancers." *Molecular cancer* 6.1 (2007): 60.
- [4] Cramer, Daniel W., et al. "Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens." *Cancer prevention research* 4.3 (2011): 365-374.
- [5] Zhu, Claire S., et al. "A framework for evaluating biomarkers for early detection: validation of biomarker panels for ovarian cancer." *Cancer Prevention Research* 4.3 (2011): 375-383.
- [6] Xiao, Zhen, et al. "Proteomic patterns: their potential for disease diagnosis." *Molecular and cellular endocrinology* 230.1 (2005): 95-106.
- [7] Stephens, W. E. "A pulsed mass spectrometer with time dispersion." *Phys. Rev* 69.691 (1946): 46.
- [8] Jensen, Finn V. and Nielson, Thomas D. *Bayesian Networks and Decision Graphs*. New York, NY : Springer, 2007. ISBN 0-387-68281-3.

- [9] Robinson, R. W. Counting unlabeled acyclic digraphs. [ed.] C. H. C. Little. *Combinatorial Mathematics V*. Berlin : Springer, 1977, Vol. 622, pp. 28-43.
- [10] Heckerman, David. *A Tutorial on Learning With Bayesian Networks*. Redmond Washington : Microsoft Research, 1995. MSR-TR-95-06.
- [11] Cover, Thomas M. and Thomas, Joy A. *Elements of Information Theory*. s.l. : John Wiley & Sons, Inc, 1991. 0-471-06259-6.
- [12] Marchetelli, Robert. *Analysis of Quality Control Data*. Department of Physics, College of William and Mary. Williamsburg, VA : s.n., 2005. Internal Report.
- [13] Elisseeff, M. and Pontil, A. *Leave-one-out error and stability of learning algorithms with applications*. [ed.] J. Suykens. s.l. : IOS Press, 2002, NATO-ASI Series on Learning Theory and Practice.
- [14] Kohavi, Ron. *A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection*. Stanford, 1995.
- [15] Karas, Michael, Doris Bachmann, and Franz Hillenkamp. "Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules." *Analytical Chemistry* 57.14 (1985): 2935-2939.
- [16] Westman-Brinkmalm, Ann, and Gunnar Brinkmalm. "A mass spectrometer's building blocks." *Mass Spectrometry: Instrumentation, Interpretation, and Applications* (2008): 15-87.



[23] Li, Jinong, et al. "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer." *Clinical chemistry* 48.8 (2002): 1296-1304.

[24] Liotta, Lance A., and Emanuel F. Petricoin. "Mass Spectrometry-Based Protein Biomarker Discovery: Solving the Remaining Challenges to Reach the Promise of Clinical Benefit." *Clinical chemistry* 56.10 (2010): 1641-1642.

[25] Siuzdak, Gary. *Mass spectrometry for biotechnology*. Academic Press, 1996.

[26] Semmes, O. John, et al. "Discrete serum protein signatures discriminate between human retrovirus-associated hematologic and neurologic disease." *Leukemia* 19.7 (2005): 1229-1238.

[27] Gatlin-Bunai, Christine L., et al. "Optimization of MALDI-TOF MS detection for enhanced sensitivity of affinity-captured proteins spanning a 100 kDa mass range." *Journal of proteome research* 6.11 (2007): 4517-4524.

[28] Baumann, Sven, et al. "Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." *Clinical Chemistry* 51.6 (2005): 973-980.

[29] Grizzle, William E., et al. "Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer." *Disease markers* 19.4-5 (2004): 185-195.

- [30] Roy, Pascal, et al. "Protein mass spectra data analysis for clinical biomarker discovery: a global review." *Briefings in bioinformatics* 12.2 (2011): 176-186.
- [31] Randolph, Timothy W., et al. "Quantifying peptide signal in MALDI-TOF mass spectrometry data." *Molecular & Cellular Proteomics* 4.12 (2005): 1990-1999.
- [32] Antoniadis, Anestis, et al. "Nonparametric Pre-Processing Methods and Inference Tools for Analyzing Time-of-Flight Mass Spectrometry Data." *Current Analytical Chemistry* 3.2 (2007): 127-147.
- [33] Wong, Jason WH, Gerard Cagney, and Hugh M. Cartwright. "SpecAlign—processing and alignment of mass spectra datasets." *Bioinformatics* 21.9 (2005): 2088-2090.
- [34] Kong, Xiaoxiao, and Cavan Reilly. "A Bayesian approach to the alignment of mass spectra." *Bioinformatics* 25.24 (2009): 3213-3220.
- [35] Feng, Yang, et al. "Alignment of protein mass spectrometry data by integrated Markov chain shifting method." *Statistics and Its Interface* 2 (2009): 329-40.

## **Vita**

Qian Si received her Bachelor of Science degree in Physics from Nanjing University in Nanjing, China in 2006. She has been a PhD student in the Department of Physics at the College of William and Mary since 2006. This dissertation was defended on November 26<sup>th</sup> 2013.