



SCHOOL of
GRADUATE STUDIES
EAST TENNESSEE STATE UNIVERSITY

East Tennessee State University
**Digital Commons @ East
Tennessee State University**

Electronic Theses and Dissertations

Student Works


5-2017

Denoising Tandem Mass Spectrometry Data

Felix Offei

East Tennessee State University

Follow this and additional works at: <https://dc.etsu.edu/etd>

 Part of the [Applied Statistics Commons](#), [Clinical Trials Commons](#), [Genomics Commons](#), [Laboratory and Basic Science Research Commons](#), and the [Statistical Methodology Commons](#)

Recommended Citation

Offei, Felix, "Denoising Tandem Mass Spectrometry Data" (2017). *Electronic Theses and Dissertations*. Paper 3218.
<https://dc.etsu.edu/etd/3218>

This Thesis - Open Access is brought to you for free and open access by the Student Works at Digital Commons @ East Tennessee State University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ East Tennessee State University. For more information, please contact digilib@etsu.edu.

Denoising Tandem Mass Spectrometry Data

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

by

Felix Offei

May 2017

Nicole Lewis, Ph.D.

Robert Price, Ph.D.,

JeanMarie Hendrickson , Ph.D.

Keywords: Protein Identification, Pre-processing, Binning.

ABSTRACT

Denoising Tandem Mass Spectrometry Data

by

Felix Offei

Protein identification using tandem mass spectrometry (MS/MS) has proven to be an effective way to identify proteins in a biological sample. An observed spectrum is constructed from the data produced by the tandem mass spectrometer. A protein can be identified if the observed spectrum aligns with the theoretical spectrum. However, data generated by the tandem mass spectrometer are affected by errors thus making protein identification challenging in the field of proteomics. Some of these errors include wrong calibration of the instrument, instrument distortion and noise. In this thesis, we present a pre-processing method, which focuses on the removal of noisy data with the hope of aiding in better identification of proteins. We employ the method of binning to reduce the number of noise peaks in the data without sacrificing the alignment of the observed spectrum with the theoretical spectrum. In some cases, the alignment of the two spectra improved.

Copyright by Felix Offei 2017

All rights reserved.

ACKNOWLEDGMENTS

First and Foremost, I would like to thank the Almighty God for how far He has brought me and seeing me through this thesis. I would like to acknowledge Dr. Nicole Lewis for her guidance and support throughout this project. At some point it got tough, it was her encouragement and productive feedback that kept me going. I am indebted to her for accepting to be my adviser.

I would also like to thank Dr. Robert Price and Dr. JeanMarie Hendrickson for agreeing to be of my committee.

Finally, I would like to acknowledge my family, friends and loved ones for their immense support, love and prayers.

TABLE OF CONTENTS

ABSTRACT	2
ACKNOWLEDGMENTS	4
LIST OF TABLES	7
LIST OF FIGURES	9
1 INTRODUCTION	11
1.1 Background	11
1.2 Proposed Work	13
1.3 Overview of Thesis	14
2 MASS SPECTROMETRY AND PROTEIN IDENTIFICATION	
METHODS	15
2.1 Tandem Mass Spectrometry	19
2.2 Noise Detection	19
2.3 Protein Identification	21
3 FRAGMENTATION PROCESS	22
3.1 Computing expected fragmented b and y ions	24
4 PRE-PROCESSING OF MASS SPECTROMETRY DATA	28
4.1 Binning	28
4.2 Denoising Data by Binning	29
4.2.1 Percentiles	30
4.2.2 Window Width	31
4.3 Pre-processing Steps	31
5 EXPERIMENTAL RESULTS	36

5.1	Short Peptides	37
5.1.1	Example 1	37
5.1.2	Example 2	40
5.1.3	Example 3	43
5.2	Long Peptides	46
5.2.1	Example 1	46
5.2.2	Example 2	49
5.3	Percentile Evaluation	52
5.3.1	Percentile Evaluation for Peptides with Short Sequences Approach	52
5.3.2	Percentile Evaluation for Peptides with Long Sequences Approach	56
5.4	More Examples	59
6	DISCUSSION	60
	BIBLIOGRAPHY	62
	VITA	67

LIST OF TABLES

1	A table listing all 20 amino acids along with their abbreviations. . . .	22
2	A list of all 20 amino acids with their masses, measured in daltons. . .	24
3	A list of different ion types and their corresponding offset values. . . .	25
4	A table comparing the distances for the peptide <i>LSDYGVQLR</i> before and after binning. The values in bold indicate the distances that were reduced.	39
5	A table comparing the distances for the peptide <i>FGSELLAK</i> before and after binning. The values in bold indicate the distances that were reduced.	42
6	A table comparing the distances for the peptide <i>VINELTEK</i> before and after binning.	45
7	A table comparing the distances for the peptide <i>ENLMQVYQQAR</i> before and after binning. The values in bold indicate the distances that were reduced.	48
8	A table comparing the distances for the peptide <i>DLVHAIPLYAIK</i> before and after binning. The values in bold indicate the distances that were reduced.	51
9	A table comparing the distances for the peptide <i>VINELTEK</i> before and after binning. The bolded values indicate the distances that increased after using a threshold of 70%.	53

10	A table comparing the distances for the peptide <i>VINELTEK</i> before and after binning. The bolded values indicate the distances that increased after using a threshold of 80%.	54
11	A table comparing the distances for the peptide <i>LSDYGVQLR</i> before and after binning. The bolded values indicate the distances that increased after using a threshold of 90%.	55
12	A table comparing the distances for the peptide <i>ENLMQVYQQAR</i> before and after binning. The bolded values indicate the distances that increased after using a threshold of 20%.	57
13	A table comparing the distances for the peptide <i>DLVHAIPLYAIK</i> before and after binning. The bolded values indicate the distances that increased after using a threshold of 50%.	58
14	A table showing results of some peptides using our method.	59

LIST OF FIGURES

1	This figure shows the basic components of a mass spectrometer. . . .	15
2	This figure, taken from <i>Frontiers in Microbiology</i> [22], shows the work flow in a MALDI-TOF mass spectrometry.	18
3	Line plot of pairs of intensities and m/z values for a given peptide. .	20
4	Theoretical spectrum for the peptide <i>VINELTEK</i> using only b and y ions. 1 represents the presence of an ion and 0 represents the absence of an ion. The solid lines are b ions and the dashed are y ions. . . .	23
5	A simplified diagram of the fragmentation of b ions. The first b ion indicated by b_1 is V and the last b ion is <i>VINELTE</i> indicated by b_7 .	26
6	A simplified diagram of the fragmentation of y ions. The first y ion indicated by y_1 is K and the last y ion is <i>INELTEK</i> indicated by y_7 .	27
7	An example of binning method applied to reduce a data set.	29
8	Figure (a) shows the observed spectrum before applying the threshold. Figure (b) shows the observed spectrum after applying the threshold.	33
9	Figure (a) shows the observed spectrum before applying the threshold. Figure (b) shows the observed spectrum after applying the threshold.	35
10	Figure (a) shows the observed and theoretical spectrum before binning. Figure (b) shows the observed and theoretical spectrum after binning with k th percentile chosen to be 60%.	38
11	Figure (a) shows the observed and theoretical spectrum before binning. Figure (b) shows the observed and theoretical spectrum after binning with k th percentile chosen to be 60%.	41

12	Figure (a) shows the observed and theoretical spectrum before binning. Figure (b) shows the observed spectrum after binning with k th percentile chosen to be 60%.	44
13	Figure (a) shows the observed and theoretical spectrum before binning. Figure (b) shows the observed and theoretical spectrum for the peptide <i>ENLMQVYQQAR</i> after binning.	47
14	Figure (a) shows the observed and theoretical spectrum before binning. Figure (b) shows the observed and theoretical spectrum for the peptide <i>DLVHAIPLYAIK</i> after binning.	50

1 INTRODUCTION

Proteins are complex compounds that carry out the daily functions of life. One important component of proteins is to defend our bodies against infection and keeping our bodies in good condition. According to the Protein Data Bank (PDB), there are records of about 80,000 entries that is made up of proteins with their biological macromolecular structures identified [1]. However, there remain some proportion of proteins with unknown functions that have yet to be identified. This is so, because researchers have not yet been able to link their sequence and structure level to known functions [2]. One factor which has contributed to unidentified proteins is noisy data. Being able to identify proteins will help researchers immensely to find out if there is a genetic disease (such as diabetes) in an organism or the existence of a bacteria infection (like Rocky Mountain spotted fever caused by the bacteria *Rickettsia rickettsii*). Researchers have been able to identify 1% - 10% of microbes in the ecosystem. While clinical proteomics is important, the need for protein identification in environmental proteomics is dire. There is 90%-99% of microbes in the entire universe that has not been identified or cultured. Being able to identify these microbes by means of protein identification will aid in environmental proteomics and possibly clinical proteomics if certain bacteria can be identified [7].

1.1 Background

We define proteome as the protein content of a cell, a tissue or an entire organism in a defined state. Proteomics describes the overall study of protein expression and function. It is known that, the human genome roughly contains 30,000 genes [3]. The

body of humans is made up of millions of cells. Each of these cells comes with a set of instructions and these instructions define us. These instructions can be likened to a recipe book for the body and they are known as the genome, which is also made up of DNA [4]. Moreover, the proteomes of mammalian cells, tissues, and body fluids are complex and display a wide dynamic range of proteins concentration, one cell can contain between one and more than 100,000 copies of a single protein [12].

Proteomics has been identified as one of the growing areas of research of the genomics. Genomics have to do with the overall analysis of gene expression using different well known techniques to identify, determine and distinguish proteins, as well as to store, communicate and interlink protein and DNA sequence and mapping information from genome projects [12]. There are numerous articles on the application of proteomics in different fields like biochemical and clinical to study and treat various kinds of diseases [25]. Genomics provided the blueprint of successful study of the genes, which is now the main focus of the study of proteomics. There have been some successful ways of applying the study of proteomics and they include mass spectrometry-based proteomics, array based proteomics, structural proteomics, proteome informatics and clinical proteomics [6].

In clinical proteomics, which is a sub-discipline of proteomics, researchers apply technologies on specimens such as proteins or group of proteins to aid diagnosing types of diseases with the sole aim of early diagnosis [16]. These groups of proteins can be significant biomarkers. Biomarkers in general are molecules that show signs of normal or abnormalities processes like diseases or unrecognisable conditions found in the body. Some of the types of these molecules that can act as biomarkers include

DNA (genes), hormones and proteins. Notably, these biomarkers are present in blood, urine or other bodily fluids [13].

There have been several methods which have been developed to aid in the identification of proteins. Current methods of protein identification comes with limitations such as noisy data, limited number of known genome sequences and incomplete ion sequences, restricting the accuracy of protein identification. In this paper, we focus on one of the key challenges faced in protein identification, noisy data, in hopes to enhance the identification of proteins. To resolve this, we employ a preprocessing procedure to reduce noise in the data.

1.2 Proposed Work

We are going to apply a preprocessing method called binning to denoise a spectrum. Data in its raw form is soiled, especially data from the mass spectrometer. It can be incomplete, thus missing some key values like m/z values and their intensities (more about m/z values and intensities in Chapter 3). Data can also be noisy (containing some random errors) or showing some irregularities in its attributes. Noisy data can therefore hinder accurate decisions which might result in misleading interpretation. The main objective of the binning method is to smooth a sorted noisy data, fix inconsistencies and reduce data but deliver same detailed results.

The data used in this study was produced by the Pacific Northwest National Laboratory (PNNL). Due to the complex structure of proteins (proteins consist of chains of amino acids, each with different chemical properties), we shall be working with peptides. Protein and peptides both consist of chains of amino acids but peptides

are known to consist about 2 to 50 amino acids where as a protein can contain about 50 or more amino acids [18]. We can easily therefore perform analysis on proteins using peptides.

1.3 Overview of Thesis

The thesis is organized as follows. Chapter 2 discusses how to obtain a protein sample and identifies some current methods used in protein identification. Chapter 3 provides a basic introduction to the idea of protein fragmentation and spectrum types. Chapter 4 describes the pre-processing method used to reduce noisy data. In Chapter 5, we provide results using our method on real data. We conclude the thesis in Chapter 6.

2 MASS SPECTROMETRY AND PROTEIN IDENTIFICATION

METHODS

The use of mass spectrometry (MS) in recent years, has become a significant approach in biological research for peptide identification [29]. There are various kinds of mass spectrometers, but one important component employed by them all, they possess magnetic and electric fields that apply a certain force on the charged particles originating from the samples to be examined. Although there are different kinds of mass spectrometers, the process involved are the same. A basic mass spectrometer has an ion source where the ions are created from the sample, a mass analyzer in which the ions produced by the ion source are isolated based on their masses, and an ion detector which sends a signal from the isolated ions [30]. Figure 1 illustrates the basic components of a mass spectrometer.



Figure 1: This figure shows the basic components of a mass spectrometer.

To elaborate on the basic process of the mass spectrometer, a sample in the form of a gas, dried form or liquid is placed into a vacuum chamber. The sample is then hit with electrons to create ions. As a result of this impact, the electrons collide in the chamber producing enough kinetic energy to displace one or more of the electrons

to produce positive ions. This process is called ionization. The mass spectrometer is noted for working with positive ions. The ions produced need to flow freely hence why they travel through a vacuum chamber. The positive ions are accelerated through the mass spectrometer where they encounter an ion repeller, carrying a slight positive charge. Like poles repels and thus this ion repeller repels away the positive ions. Most of these positive ions produced carry a charge of +1. A magnetic field is formed by an electromagnet deflects the ions by different amounts. The rate at which these ions deflect is determined by the mass of the ion and its charge. Lighter ions travel faster and are thus deflected more than heavy ions [31].

The lighter ions successfully make it through the mass spectrometer and are detected by an ion detector. Some ions end up hitting the walls of the mass spectrometer, receiving electrons in the process and are neutralized. In the course of time, they are eliminated by the vacuum pump. The output from the mass spectrometer is a spectrum representing the ions that hit the detector with their mass-to-charge ratio. In the spectrum, peaks can be recognized by plotting intensities versus the mass-to-charge ratio (m/z). The spectrum can contain thousands of peaks that have to under go processing to be able to identify the peptide.

There are two methods commonly used by the mass spectrometer to identify proteins. One of the well known methods is MALDI-TOF, matrix-assisted laser desorption and ionization-time of flight mass spectrometry. In MALDI-TOF, the sample is first ionized by a laser. Once they are ionized, they receive electric charges. These charged ions are accelerated by an electric field in the mass spectrometer. The time of flight of these ions are monitored and reported. The output from the detector is

a spectrum of mass of the ions and their respective intensities. Figure 2 shows a graphical representation of the work flow in a MALDI-TOF mass spectrometry. One of the key features of the MALDI-TOF procedure is its ability to work perfectly with smaller quantities of samples and also its capacity to process samples in the shortest possible time.

Surface-enhanced laser desorption/ionization (SELDI-TOF) mass spectrometry is well known for quantifying proteins of low-molecular weights. With this method, there is a sample matrix, known as protein chip that plays an effective role in the purification and ionization of the sample [17]. There are 3 parts that make up the SELDI-TOF mass spectrometer; protein chip, a mass analyser and software to analyze the data. The sample is first transferred on to a chromatographic surface where the protein chips are incubated. The chromatographic surface absorbs the proteins of interest based on their individual properties. This absorption can either be done by an electrostatic interaction or by partition. The particles are then analyzed by the TOF (time-of-flight) mass spectrometer and the output will be a spectrum comprised of mass-to-charge (m/z) ratio values and their corresponding intensities [19].

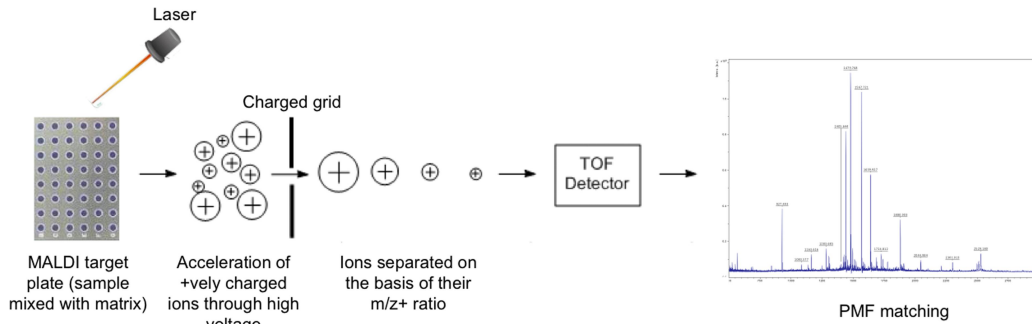


Figure 2: This figure, taken from *Frontiers in Microbiology* [22], shows the work flow in a MALDI-TOF mass spectrometry.

In mass spectrometry, the ions with distinct masses are separated in a flight tube. The mass spectrometer employs a quadratic transformation method to calculate the m/z from the flight time. The coefficients of the quadratic transformation can be determined with a small number of particles from the sample (normally between 3 and 7) with known masses are used to generate a spectrum. The peaks matching the known masses in the spectrum are then determined. The coefficients of the quadratic transformation are then determined by the method of least squares, given a set of (*time*, *mass*) [14]. The process by which the observed flight time is mapped to the m/z values is called calibration. The pairs of intensities and m/z values is then plotted in a spectrum. Data produced from the mass spectrometer can be used to plot a spectrum of peaks, where one peak represents the pair of intensity (vertical axis) and m/z values (horizontal axis) of the peptide in the current sample. Figure 3 is a graphical display of a spectrum for a given peptide. The peaks we see in this plot shows signal peaks and noise peaks. Signal peaks are defined as peaks whose observed m/z value is closest to the theoretical m/z value within a 0.5 Da tolerance.

Usually these signal peaks are the ones with large intensity values.

2.1 Tandem Mass Spectrometry

Tandem mass spectrometry, commonly known as tandem MS/MS is a mass spectrometry comprised of two stages of the mass spectrometer. There are two mass analyzers in tandem MS/MS. The first mass analyzer is responsible for isolating ions of specific m/z values representing a particular peptide coming from the ion source. Tandem MS/MS breaks down precursor ions into product ions also known as fragment ions. The product ions are the ions that show the chemical composition of the precursor ions. Those ions successfully isolated by the first mass analyzer are accelerated into a collision cell chamber holding inert gas. This is where the fragmentation of the ions take place. The process is termed collision induced dissociation (CID) or collision activated dissociation (CAD). The second mass analyzer measures the m/z values of the fragmented ions to obtain the sequence of the peptide.

Some of the commonly used types of tandem MS/MS include triple stage quadrupoles (TSQ), quadrupole time-of-flight (QTOF), quadrupole-linear ion trap (QTRAP), 3D and linear ion traps [15].

2.2 Noise Detection

In mass spectrometry, only a small portion of peaks in a spectrum are essential for peptide identification. Most of the peaks in the spectrum are noise and can hinder the process of peptide identification. Removing these noise peaks will aid in better identification of peptides.

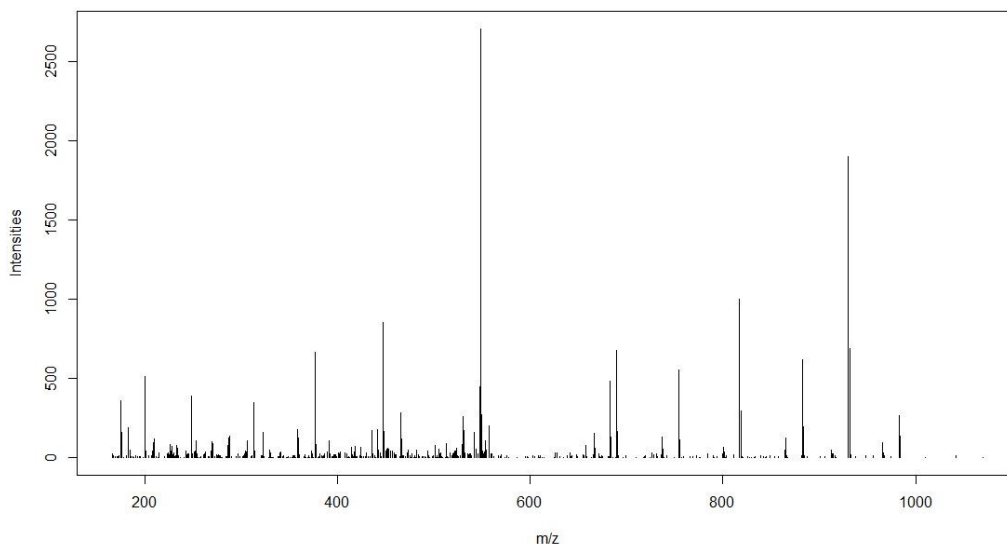


Figure 3: Line plot of pairs of intensities and m/z values for a given peptide.

There are two main sources of noise in mass spectrometry data: chemical noise, contaminants and random noise. Random noise is depicted by small peaks which are uniformly scattered in the masses of the ions. Random noise is known to be caused by electrical distortion in the mass spectrometer. In this thesis, we focus on random noise. It can easily be removed by preprocessing methods such as smoothing and binning. Chemical noise has a pattern which is very identical to signal peaks. The process of removing chemical noise can be very cumbersome. Chemical noise has the effect of shifting peaks, thus hindering peptide identification. In a previous work, they proposed wavelet decomposition can be used for removing chemical noise [32].

2.3 Protein Identification

With the improvements in using mass spectrometry technology, peptide identification has become less challenging. Most methods that are available make comparison of a theoretical spectrum and observed spectrum which is forecasted based on a sequence from a protein database [27]. Two such methods are peptide mass fingerprint (PMF) analysis and de novo sequencing in protein identification [33].

In PMF, an unknown protein is dissolved into short peptides with a proteolytic enzyme [8]. The mass of these short peptides can easily be measured using the MALDI-TOF mass spectrometer. Once the mass of the peptides are determined, they are compared to a database, which contain theoretical spectra of peptides with known sequences [9]. The match which produces the highest score is therefore considered to be the identity of the unknown peptide [9].

De novo sequencing make use of the b and y ions produced by collision induced dissociation (CID) [10]. This process use the mass difference between two fragment ions to compute the mass of an amino acid present in a candidate peptide. The computed mass between the amino acids becomes the residue. The process can be used to compute all residues between the ions. If the b and y ion series in the spectrum can be determined, then the sequence of the peptide can be identified [10]. The peptide sequences are used to map a predicted data from sequence databases [11]. In this approach, a scoring function is used to scan the observed peptide and the candidate peptide to check for a match.

3 FRAGMENTATION PROCESS

The most important part of peptide identification is to align an observed spectrum to a theoretical spectrum of a candidate peptide. In this chapter, we present how a peptide is fragmented using the mass spectrometer. A peptide is a combination of two or more amino acids connected in a chain. Twenty distinct types of amino acids make up these peptides. A protein's structure and function is determined by the sequence of the amino acids. Table 1 shows an alphabetical order of the amino acids with their matching 3 letter and 1 letter code. To ease notation, we will be using the 1 letter code for the remaining of the thesis. The theoretical spectrum of a peptide is a collection of peaks with each peak located at the m/z value of each ion type.

In order to find the theoretical spectrum, the candidate peptide is first split into all possible ion combinations. Collision-induced dissociation (CID) is the uttermost way for parent ion fragmentation.

Table 1: A table listing all 20 amino acids along with their abbreviations.

Amino Acid	3 Letter Code	1 Letter Code	Amino Acid	3 Letter Code	1 Letter Code
Alanine	ALa	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic Acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamine	Gln	Q	Serine	Ser	S
Glutamic Acid	Glu	E	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Try	Y
Isoleucine	Ile	I	Valine	Val	V

The most common fragment ions formed by collision induced dissociation (CID) are the b -ions and y -ions [20]. Figure 4 shows a theoretical spectrum using only the b and y ions. Each of the ions is located at its corresponding m/z value on the theoretical spectrum. It must be noted that, the distances between the peaks can also be used to make conclusions about sequences of the peptide. In a spectrum, the peaks observed shows a reflection of the fragment ions produced from the mass spectrometer.

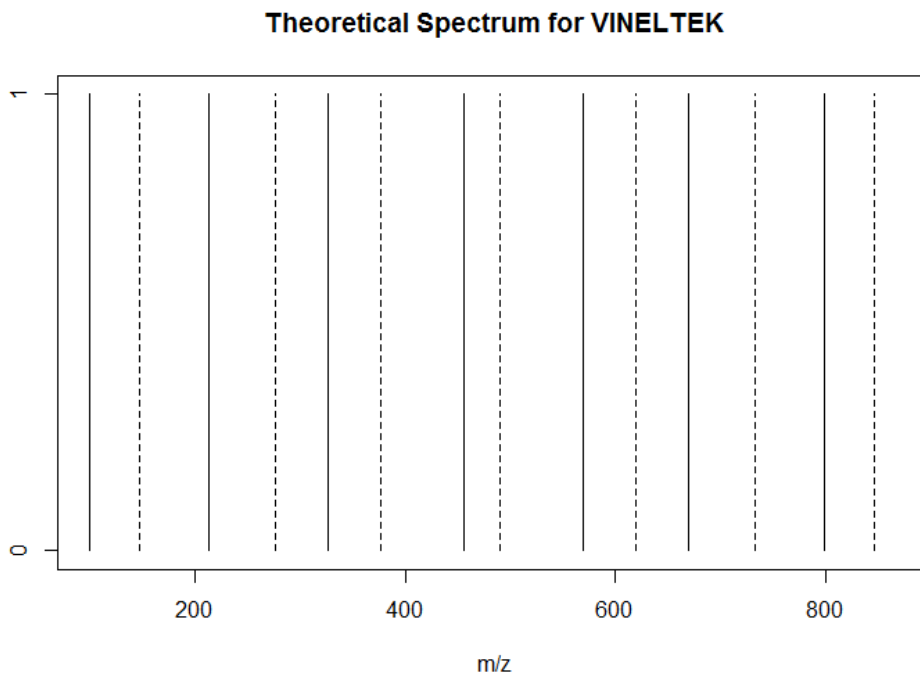


Figure 4: Theoretical spectrum for the peptide *VINELTEK* using only b and y ions. 1 represents the presence of an ion and 0 represents the absence of an ion. The solid lines are b ions and the dashed are y ions.

The b ions extend from the N-terminus. The N-terminus or the amino terminus

is the beginning of a peptide and it is terminated by an amino acid with a free amine group ($-NH_2$) [23]. The y ions serve as a complement to the b ions and they extend from the carboxyl terminus, also called the C-terminus. The C-terminus is the end of the peptide terminated by a free carboxyl group ($-COOH$). The charge of y ion is maintained on the carboxyl terminus.

Table 2: A list of all 20 amino acids with their masses, measured in daltons.

Amino Acid	Mass	Amino Acid	Mass
A	71.0371	M	131.04
C	103.009	N	114.043
D	115.027	P	97.0528
E	129.043	Q	128.059
F	147.068	R	156.101
G	57.0215	S	87.032
H	137.059	T	101.048
I	113.084	V	99.0684
K	128.095	W	186.079
L	113.084	Y	163.063

3.1 Computing expected fragmented b and y ions

In a typical spectrum, the peaks can be differentiated by the mass of the amino acid. Table 2 shows a list of the twenty amino acids and their mass in daltons (Da). These peaks are a representation of the fragment ions formed through the collision process of a mass spectrometer. The mass of a particular ion is determined by $\sum_{j=1}^n m(p_j) + \delta_l$ where $m(p_j)$ represents the mass of the amino acid located in the j th position, n is the number of amino acids in the ion sequence, $p(j)$ is the amino

acid located in the j th position, l represents the type of ion present δ_l is the offset value for the type of ion. The offsets correspond to peaks showing ion types produced by the tandem mass spectrometer [24]. Table 3 shows a list of the various ions types along with their terminus, their corresponding offset values and how to compute the mass of the ion.

Table 3: A list of different ion types and their corresponding offset values.

Ion	Terminus	Offset Value	Position
b	N	0.85	(M + 0.85)
$b - H_2O$	N	-17.05	(M - 17.05)
a	N	-27.15	(M - 27.15)
$b - NH_3$	N	-16.15	(M - 16.15)
$b - H_2O - H_2O$	N	-35.20	(M - 35.20)
$b - H_2O - NH_3$	N	-34.20	(M - 34.20)
$a - NH_3$	N	-44.25	(M - 44.25)
$a - H_2O$	N	-45.15	(M - 45.15)
y	C	18.85	(M + 18.85)
$y - H_2O$	C	0.90	(M + 0.90)
y^2	C	20.05	(M + 20.05)/2
$y - NH_3$	C	1.90	(M + 1.90)
$y^2 - H_2O$	C	2.30	(M + 2.30)/2
$y - H_2O - NH_3$	C	-16.10	(M - 16.10)
$y - H_2O - H_2O$	C	-17.15	(M - 17.15)

Consider the peptide *VINELTEK* whose sequence consists of 8 amino acids. This peptide has $n - 1$ b ions and $n - 1$ y ions. Recall that, the beginning of any peptide is on the N-terminus, and thus V is the first b ion, denoted b_1 , with a mass of $99.0684 + 0.85 = 99.9184$ Da. The second b ion is VI with a mass of $99.0684 + 113.084 + 0.85 = 213.0024$ Da. In the same manner, we compute the subsequent masses of the

other b ions in the order VIN , $VINE$, $VINEL$, $VINELT$, and $VINELTE$ with their respective masses 327.0454, 456.0884, 569.1724, 670.2204, and 799.2634 daltons. Figure 5 shows a simplified diagram of the fragmentation of the b ions found in the peptide $VINELTEK$.

The fragmentation of the y ions starts at the C-terminus and the first y ion, denoted y_1 , is K . The mass of y_1 is computed to be $128.095 + 18.85 = 146.945$ Da. The second y ion denoted y_2 is EK with a mass of $129.043 + 128.095 + 18.85 = 275.988$ Da.

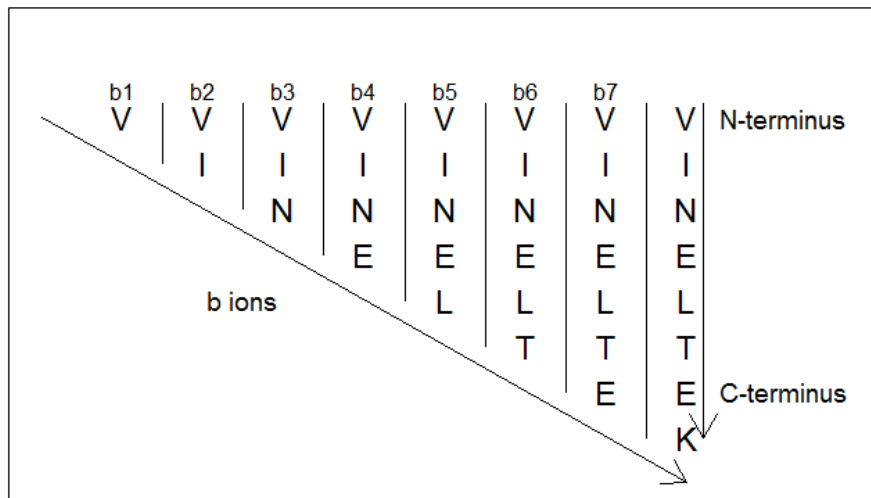


Figure 5: A simplified diagram of the fragmentation of b ions. The first b ion indicated by b_1 is V and the last b ion is $VINELTE$ indicated by b_7 .

The subsequent y ions, in order, are TEK , $LTEK$, $ELTEK$, $NELTEK$, and

INELTEK with their respective masses to be 377.036, 490.12, 619.163, 733.206, 846.29 Daltons. Figure 6 shows a simplified diagram of the fragmentation of the y ions found in the peptide *VINELTEK*.

To find the total weight of a candidate peptide, we add the mass of all the amino acids present in the peptide plus the mass of one hydrogen and water molecule. Consider the peptide *VINELTEK*. To compute the total weight, we add the following masses representing the amino acids in order: $99.0684 + 113.084 + 114.043 + 129.043 + 113.084 + 101.048 + 129.043 + 128.095 + 18.010565$ (water) $+ 1.00794$ (hydrogen) $= 945.5269$ da.

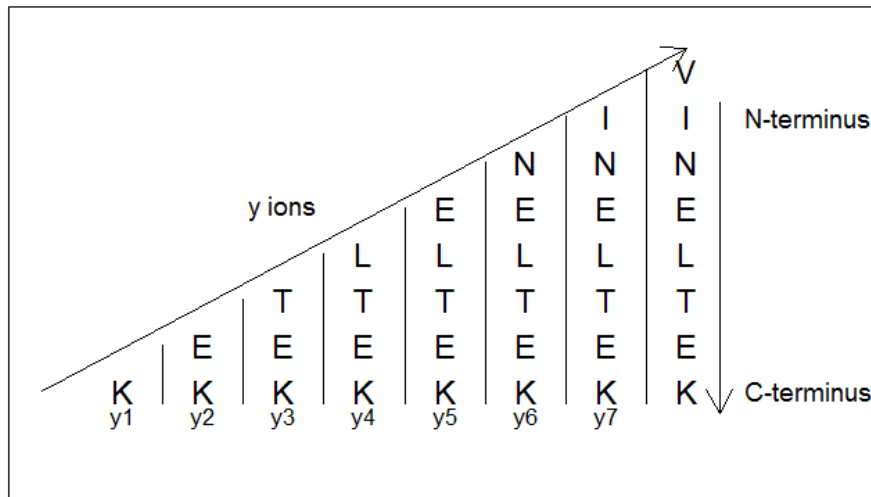


Figure 6: A simplified diagram of the fragmentation of y ions. The first y ion indicated by y_1 is K and the last y ion is *INELTEK* indicated by y_7 .

4 PRE-PROCESSING OF MASS SPECTROMETRY DATA

Data produced by mass spectrometer are influenced by errors due to sample preparation, sample insertion into the experiment and the instrument itself. It therefore becomes a challenge to use the raw data in peptide identification because of noise, instrument distortion and saturation, wrong calibration and m/z measurement errors. We propose a binning method to reduce the number of noise peaks in the data.

4.1 Binning

One common method of preprocessing mass spectrometry data is binning. Binning is a method used to group large quantitative numbers into small bins. This method seeks to reduce the amount of data by grouping adjacent m/z values of the data into small bins. This causes a reduction in the data thus reducing the number of noise peaks present in the spectrum. It becomes problematic when you have to examine the spectra by setting a sliding window. This is because, if the window width is too large, we may accidentally remove some signal peaks from the data set. This will cause the observed spectrum and theoretical spectrum not to be aligned. In the case of a small window with, the outcome of data reduction is not given; thus, the number of peaks intended to be reduced will not happen.

A bin will consists of N m/z (mass-to-charge) values and their associated intensities. The structure of the bin is in the form $[(I_j, m/z_j), \dots, (I_N, m/z_N)]$, where I_j is the j th intensity value and m/z_j is the j th m/z value for $j = 1, \dots, N$. They are then joined to form $(I, m/z)$ vector pairs. The m/z value of the bin is calculated by taking the mean of all the N original m/z data values and the intensity value of I is

calculated by using an aggregate function (such as the maximum intensity or sum) of all the N original intensity values. Figure 7 shows a small data set where binning was applied (adapted from [26]). The total number of observations in the original data was 9 and after binning, it reduced to 4. We use this binning method in conjunction with thresholding to denoise large data sets.

Original Data		After Binning (Window Width = 1)	
m/z	Intensity	m/z	Intensity
174.3213	2.51136	174.7544	4.46839
175.1874	4.46839		
176.7977	2.71992	177.1994	3.59747
177.0584	3.59747		
177.7420	2.8503		
178.1917	6.03303	178.5766	14.22540
178.9615	14.22540		
180.0874	2.08024	180.27315	19.33010
180.4589	19.33010		

Figure 7: An example of binning method applied to reduce a data set.

4.2 Denoising Data by Binning

In order to use binning, the data must be sorted. We know that peaks with large intensities are usually signal peaks. Therefore, we want to ensure these peaks remain at their current m/z location. We first compute a percentile of the observed intensity, denoted p . The percentile chosen allows us to keep peaks with large intensities. That is, any peaks above the threshold value will be kept. Binning will be applied to the remaining peaks. After binning, the binned data set will be combined with the pairs of m/z values and intensities whose intensity values are larger than the threshold

value to form the final denoised data set.

4.2.1 Percentiles

After our data is sorted, we must threshold out the peaks with large intensities. It was found that the same percentile p to determine the threshold value does not work for all peptides. Upon investigation, it was found that peptides with short sequences behaved differently than peptides with longer sequences. Although the mass spectrometer cannot determine the length of the peptide sequence, it can determine the total weight of the peptide. It must be noted that, the length of a peptide or the number of amino acids does not determine the size of the peptide but rather the total weight. We consider a peptide to be short if the total weight is found to be less than 1100 Da and a peptide is considered long if the total weight is 1100 Da or more.

Consider a peptide, *VINELTEK* consisting of 8 amino acids. The total weight of this peptide is 945.5269 Da, thus this peptide would be classified as a short peptide. The threshold value for the 60th percentile for the peptide *VINELTEK* is 21.07. And so all intensity values that fall below this threshold value with their corresponding m/z values will be binned.

Consider another peptide *AFNEMQPIVDR* with a total mass weight of 1319.642 Da and 654 pairs of m/z values and corresponding intensities. Since the total weight is 1100 Da or more, this peptide would be classified as a peptide with a long sequence. The threshold value for the 10th percentile was computed to be 8.257594 Da.

4.2.2 Window Width

Since observed spectra are unique, the same window width cannot be applied to all sequences. We need to calculate the average distance for all m/z values. A general rule of thumb is to round the average value to the nearest 0.5 Da to get the window width. For instance, if the average distance between m/z values of the ions is 0.2, then a window width of 0.5 Da would be appropriate for binning. If the average distance for all m/z values in the spectrum is than 1.75, then a window width of 2 Da should be employed.

Consider the peptide *VINELTEK* with 347 pairs of m/z values and intensities. Computing for the average distance between m/z values, we get an average of 2.07, thus a window width of 2.5 Da should be applied. Now consider the peptide *AAAAPVTGPLADFPIQETITFDDFAK* with 573 pairs of m/z values and intensities. We found the average distance between the m/z values to be 2.77 and so a window width of 3.0 should be used for binning.

4.3 Pre-processing Steps

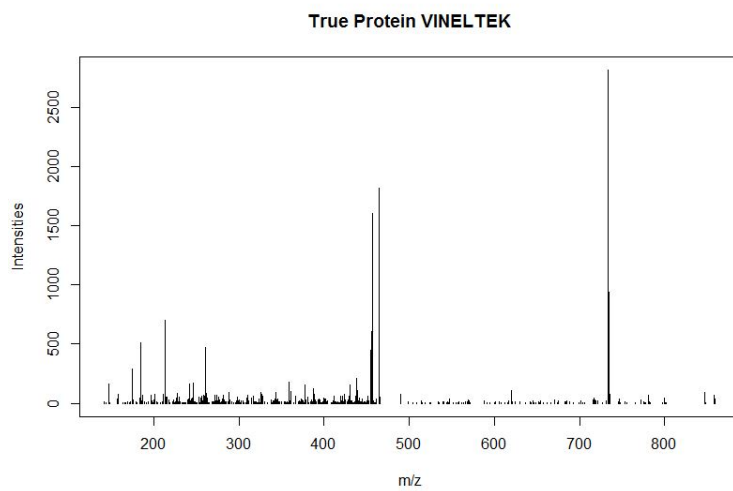
To denoise the observed spectrum for peptides, consider the following steps:

1. Find the total weight of a given peptide and compare to 1100 Da. If the total weight is less than 1100 Da, the peptide is classified as a short peptide otherwise it is classified as long peptide.
2. Determine the threshold value. Use $p = 0.6$ for short peptides and $p = 0.1$ for long peptides.

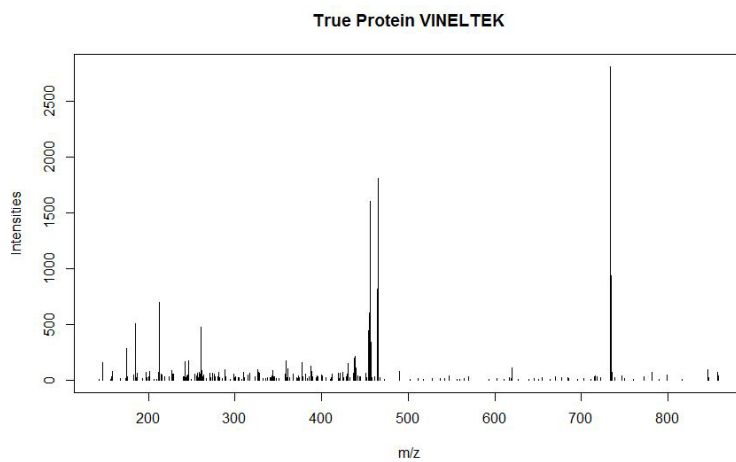
3. Compare all intensities to the threshold value. For intensity values greater than the threshold value save until the final step.
4. Compute the average distance between m/z values and determine the appropriate window width by rounding to the nearest 0.5.
5. Apply binning to the m/z values whose corresponding intensity values are below the threshold values.
6. Combine the binned data to the pairs of m/z values and intensities whose intensities are larger than the threshold value.

Consider the peptide *VINELTEK* with 348 pairs of m/z values and corresponding intensity values. The threshold value was determined to be 21.07. Thus, any pair of m/z values whose corresponding intensity value is larger than 21.07 was saved until the final step and all other pairs were binned. The average distance between m/z values was found to be 2.07 and thus a window with of 2.5 was used. After applying the method, the spectrum was reduced from 348 to 226 pairs of m/z values and intensities. Figure 8 (a) shows the observed spectrum before applying our method. Figure 8 (b) shows the observed spectrum after applying our method. You can see quite a bit of the noise has been reduced from the spectrum.

Consider the peptide *AFNEMQPIVDR* with 654 pairs of m/z values and corresponding intensities. The threshold value for the 10th percentile was computed to be 8.257594. Any pair of m/z values whose corresponding intensity value is larger than 8.257594 was saved until the final step and all other pairs were binned. The average distance between m/z values was found to be 1.70 and thus a window with of



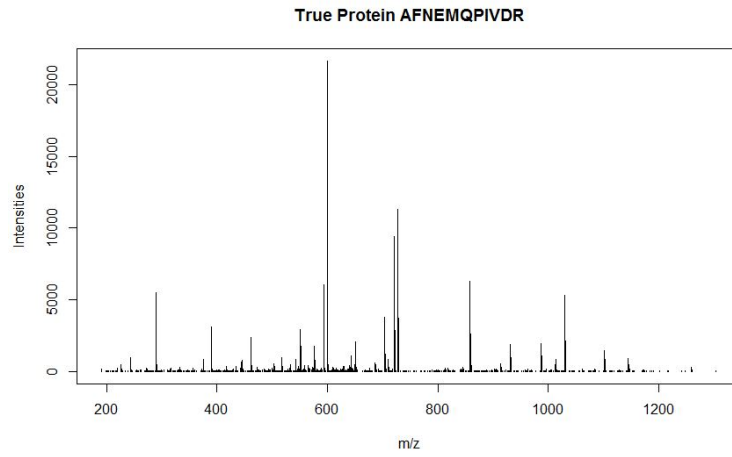
(a)



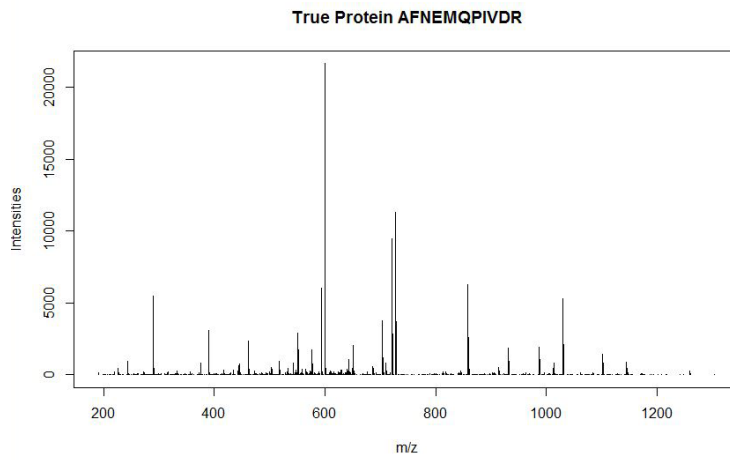
(b)

Figure 8: (a) shows the observed spectrum before applying the threshold. (b) shows the observed spectrum after applying the threshold

2.0 was used. After applying the method, the spectrum was reduced from 654 to 619 pairs of m/z values and intensities. Figure 9 (a) shows the observed spectrum before applying our method. Figure 9 (b) shows the observed spectrum after applying our method. One can see some of the noise has been reduced in the spectrum.



(a)



(b)

Figure 9: (a) shows the observed spectrum before applying the threshold. (b) shows the observed spectrum after applying the threshold

5 EXPERIMENTAL RESULTS

The data used for the analysis was produced by the Pacific Northwest National Laboratory (PNNL) and is available to the general public. Each of these peptides consist of pairs of m/z values and their respective intensities. We present examples in which our binning method is applied in order to denoise the spectrum. The data set consists of 1,206 peptides with amino acids sequences ranging in length from 7 to 20 amino acids. Tandem mass spectrometry was used to determine the proteomic profiles of the samples. The m/z values range from 100 Da to 2000 Da with corresponding intensities from 1 to 3000.

To ensure our method does not alter the alignment between the observed spectrum and the theoretical spectrum, we compare the distances between the theoretical m/z value based on the b and y ions and the closest observed m/z value to the distances between the theoretical m/z value based on the b and y ions and the closest observed m/z value after applying our method. If these distances remain the same or reduce, this implies our method successfully does not change the alignment or improves the alignment, respectively. If these distances increase, the alignment has shifted and will hinder the identification of that peptide. In that case, our method would be unreliable.

5.1 Short Peptides

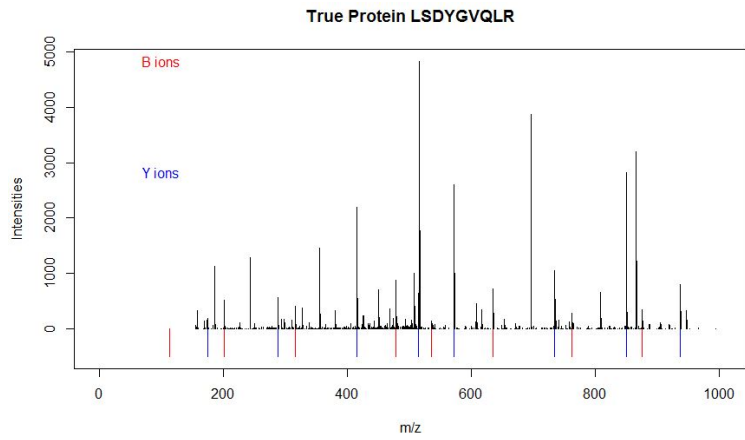
5.1.1 Example 1

To ensure the reliability of our method, we applied our method to several peptides with short amino acid sequences. Consider the short peptide *LSDYGVQLR* whose total weight is 1050.558 with 470 pairs of m/z values ranging from 155 to 995 Da and corresponding intensities ranging from 1 to 4900.

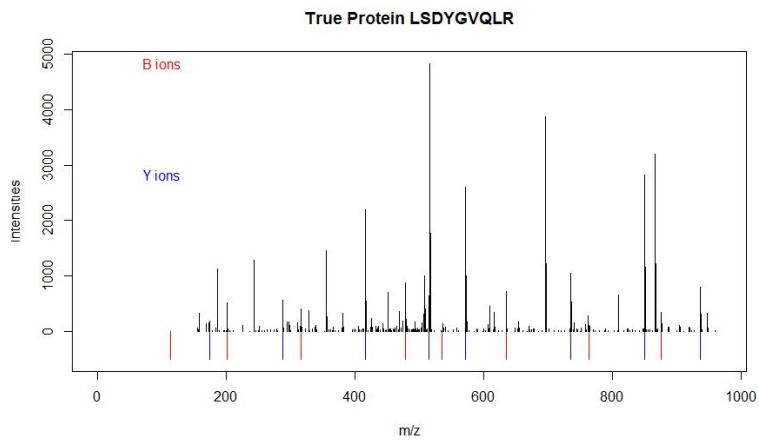
Using a 60th percentile, the threshold value was found to be 28. The average distance between m/z values is 1.8 Da and thus we set the window width to be 2.0 Da. After our applying our method, the observed spectrum was reduced from 470 pairs to 294 pairs of m/z values and their corresponding intensities.

Figure (a) in Figure 10 shows the observed and theoretical spectrum before our method is applied and Figure (b) shows the observed and theoretical spectrum after our method was applied. One can see some of the noise peaks are removed successfully without affecting the alignment of the observed and theoretical spectrum.

Table 4 lists the closest observed m/z values for both before and after our method is applied, the theoretical m/z value, the distances between the observed m/z value before denoising and the theoretical m/z value, and the distances between the observed m/z value after denoising and the theoretical m/z value. The values in bold in Table 4 indicate the distances that were reduced. We see all the distances either remained the same or reduced indicating our method is doing well.



(a)



(b)

Figure 10: (a) shows the observed and theoretical spectrum before binning. (b) shows the observed and theoretical spectrum after binning with kth percentile chosen to be 60%.

Table 4: A table comparing the distances for the peptide *LSDYGVQLR* before and after binning. The values in bold indicate the distances that were reduced.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
155.2550	113.9340	155.2550	41.3209	41.3209
174.3954	174.9510	175.0938	0.5556	0.1427
200.1734	200.9660	201.0907	0.7926	0.1247
288.1805	288.0350	288.1805	0.1455	0.1455
315.6753	315.9930	316.2448	0.3177	0.2517
416.2673	416.0940	416.2673	0.1733	0.1733
479.0506	479.0560	479.0506	0.0054	0.0054
515.3530	515.1624	515.3530	0.1906	0.1906
536.1034	536.0775	536.1034	0.0259	0.0259
572.2805	572.1839	572.2805	0.0966	0.0965
635.0522	635.1459	635.0522	0.0937	0.0937
735.3122	735.2469	735.3122	0.0653	0.0653
763.1658	763.2049	763.1658	0.0391	0.0391
850.3229	850.2739	850.3229	0.0490	0.0490
876.1458	876.2889	876.1458	0.1431	0.1431
937.3237	937.3059	937.3237	0.0178	0.0178

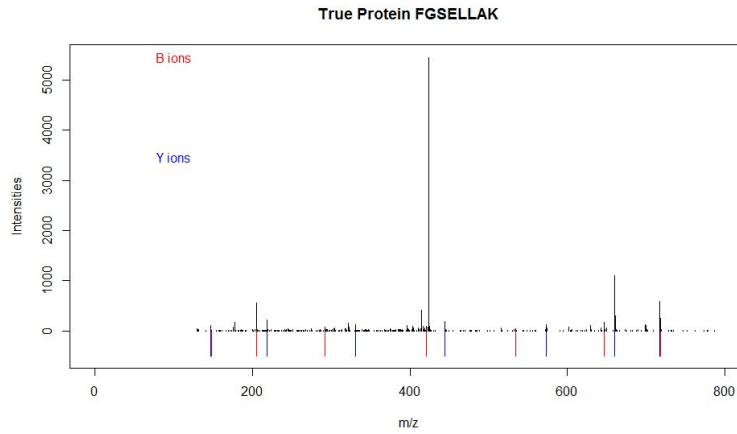
5.1.2 Example 2

Consider the short peptide *FGSELLAK* whose total weight is 864.4831 with 269 pairs of m/z values ranging from 129 to 787 Da and corresponding intensities ranging from 1 to 5455.

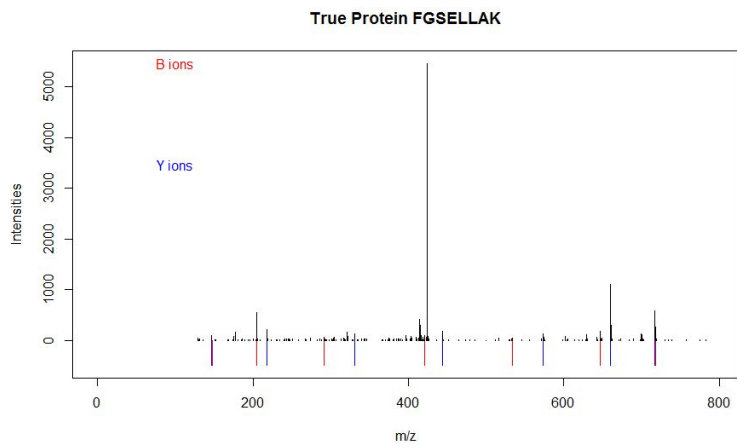
Using a 60th percentile, the threshold value was found to be 14. The average distance between m/z values is 2.4 Da and thus we set the window width to be 2.5 Da. After applying our method, the observed spectrum was reduced from 269 pairs to 173 pairs of m/z values and their corresponding intensities

Figure 11 (a) shows the observed and theoretical spectrum before our method is applied and Figure 11 (b) shows the observed and theoretical spectrum after our method is applied. One can clearly see some of the noise peaks are removed without disturbing the alignment of the two spectra.

Table 5 lists the closest observed m/z values for both before and after our method is applied, the theoretical m/z value, the distances between the observed m/z value before denoising and the theoretical m/z value, and the distances between the observed m/z value after denoising and the theoretical m/z value. We see all the distances either remained the same or reduced indicating our method is doing well.



(a)



(b)

Figure 11: (a) shows the observed and theoretical spectrum before binning. (b) shows the observed and theoretical spectrum after binning with k th percentile chosen to be 60%.

Table 5: A table comparing the distances for the peptide *FGSELLAK* before and after binning. The values in bold indicate the distances that were reduced.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
147.0104	146.9450	147.0101	0.0654	0.0654
147.0104	147.9180	147.0101	0.90759	0.90759
204.9052	204.9395	204.9052	0.0343	0.0343
218.1224	217.9821	218.1224	0.1403	0.1403
292.9860	291.9715	292.0566	1.0145	0.0851
331.2053	331.0661	331.2053	0.1392	0.1392
421.2489	421.0145	421.2489	0.2344	0.2344
444.2705	444.1501	444.2705	0.1204	0.1204
534.1363	534.0985	534.1363	0.0378	0.0378
573.2606	573.1931	573.2606	0.0675	0.0675
647.0905	647.1825	647.0905	0.0920	0.0920
660.2869	660.2251	660.2869	0.0618	0.0618
717.3460	717.2466	717.3460	0.0994	0.0994
718.3046	718.2196	718.3046	0.0850	0.0850

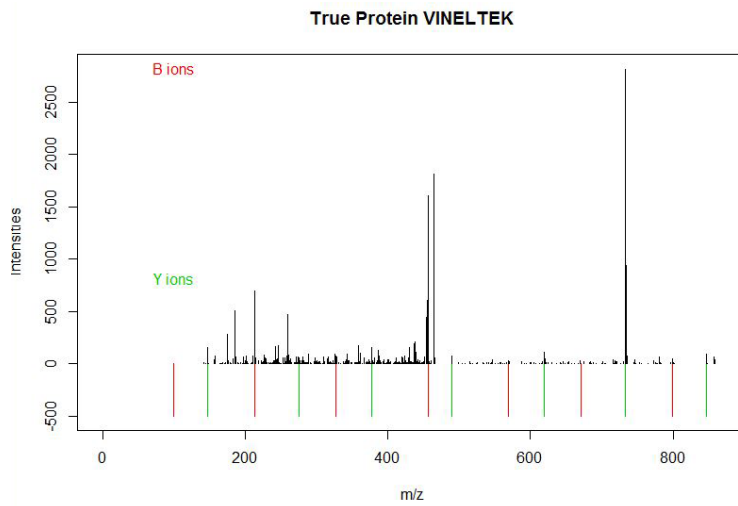
5.1.3 Example 3

Now consider the peptide *VINELTEK* whose total weight is 945.5269 with 348 pairs of m/z values ranging from 140 to 860 Da and corresponding intensities ranging from 1 to 2820.

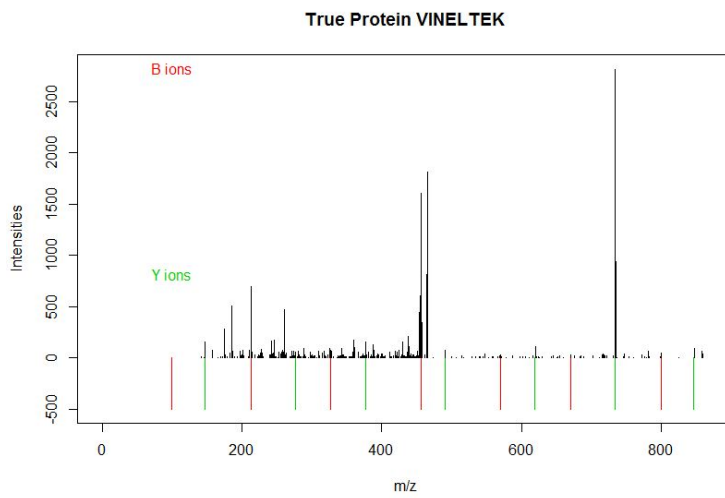
Using a 60th percentile, the threshold value was found to be 21. The average distance between m/z values is 2.07 Da and thus we set the window width to be 2.5 Da. After our applying our method, the observed spectrum was reduced from 347 pairs to 213 pairs of m/z values and their corresponding intensities.

Figure 12 (a) shows the observed and theoretical spectrum before our method is applied and Figure 12 (b) shows the observed and theoretical spectrum after our method is applied. Once again we see the noise peaks have been successfully removed without ruining the alignment of the observed and theoretical spectrum.

Table 6 lists the closest observed m/z values for both before and after our method is applied, the theoretical m/z value, the distances between the observed m/z value before denoising and the theoretical m/z value, and the distances between the observed m/z value after denoising and the theoretical m/z value. In all cases but the first, the distances remained the same. Although the distance increased for the first observed m/z value closest to the theoretical m/z value, this is not of concern because the first b and y ions are rarely captured in the mass spectrometer. Therefore the alignment will not be disturbed.



(a)



(b)

Figure 12: (a) shows the observed and theoretical spectrum before binning. (b) shows the observed spectrum after binning with kth percentile chosen to be 60%.

Table 6: A table comparing the distances for the peptide *VINELTEK* before and after binning.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
141.9028	99.9184	144.7715	41.9844	44.8531
147.0816	146.9450	147.0816	0.1366	0.1366
213.0144	213.0024	213.0144	0.0120	0.0120
276.1071	275.9880	276.1065	0.1191	0.1185
327.0514	327.0454	327.0514	0.0060	0.0060
377.2013	377.0360	377.2013	0.1653	0.1653
456.2416	456.0884	456.2416	0.1532	0.1532
490.2245	490.1200	490.2245	0.1045	0.1045
569.1671	569.1724	569.1671	0.0053	0.0053
619.3815	619.1630	619.3815	0.2185	0.2185
670.1361	670.2204	670.1361	0.0843	0.0843
733.1962	733.2060	733.1962	0.0098	0.0098
799.2451	799.2634	799.2451	0.0183	0.0183
846.3795	846.2900	846.3795	0.0895	0.0895

5.2 Long Peptides

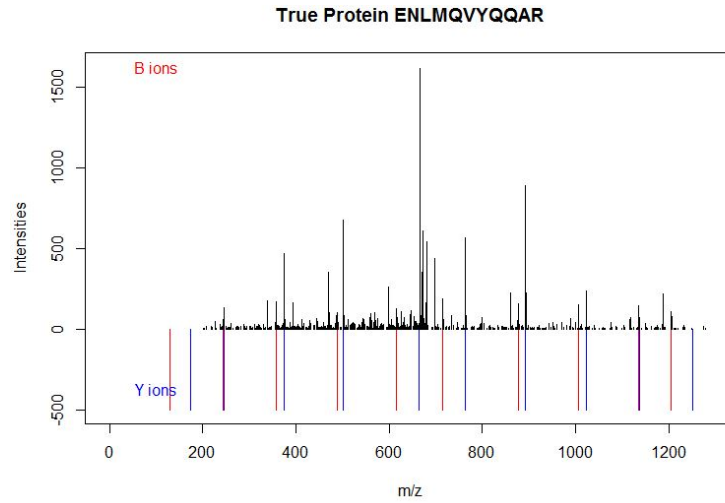
5.2.1 Example 1

Now we will consider peptides in which we classify as being long. First, consider the peptide *ENLMQVYQQAR* whose total weight is 1379.675 with 590 pairs of m/z values ranging from 200 to 1280 Da and corresponding intensities ranging from 1 to 1700.

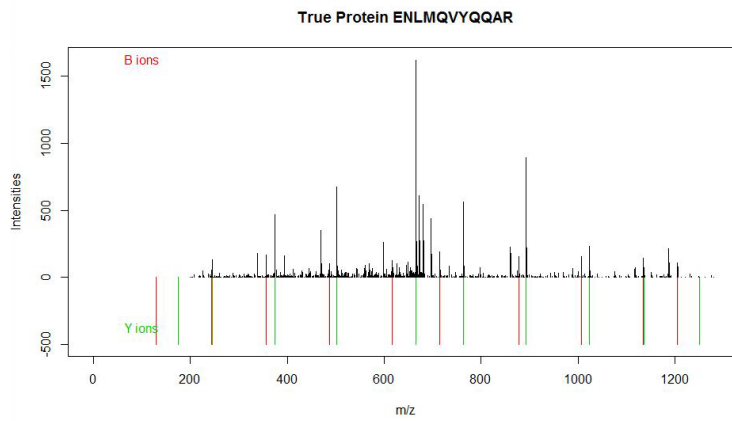
Using a 10th percentile, the threshold value was found to be 4. The average distance between m/z values is 1.8 Da and so we set the window width to be 2.0 Da. After our applying our method, the observed spectrum was reduced from 590 pairs to 548 pairs of m/z values and their corresponding intensities.

Figure 13 (a) shows the observed spectrum before our method is applied and Figure 13 (b) shows the observed spectrum after our method is applied. We see that a few noise peaks have been removed and the alignment remains the same.

Table 7 lists the closest observed m/z values for both before and after our method is applied, the theoretical m/z value, the distances between the observed m/z value before denoising and the theoretical m/z value, and the distances between the observed m/z value after denoising and the theoretical m/z value. We see in most cases the distances either remained the same or was reduced. The first two increase but as we mention in Section 5.1.3 this not a concern and the alignment is still intact.



(a)



(b)

Figure 13: (a) shows the observed and theoretical spectrum before binning. (b) shows the observed and theoretical spectrum for the peptide *ENLMQVYQQAR* after binning.

Table 7: A table comparing the distances for the peptide *ENLMQVYQQAR* before and after binning. The values in bold indicate the distances that were reduced.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
200.0935	129.8930	201.5275	70.2005	71.6345
200.0935	174.9510	201.5275	25.1425	26.5765
243.8989	243.9360	243.8989	0.0371	0.0371
246.0997	245.9881	246.0997	0.1116	0.1116
357.1619	357.0200	357.1619	0.1419	0.1419
374.2530	374.0471	374.2530	0.2059	0.2059
488.0701	488.0600	488.0701	0.0101	0.0101
502.2154	502.1061	502.2154	0.1093	0.1093
616.1608	616.1190	616.1608	0.0418	0.0418
665.2491	665.1691	665.2491	0.0800	0.07998
715.1245	715.1874	715.1245	0.0629	0.0629
764.2294	764.2375	764.2294	0.0081	0.0081
878.1607	878.2504	878.1607	0.0897	0.08969
892.3641	892.2965	92.3641	0.0676	0.06757
1006.1984	1006.3094	1006.1984	0.1110	0.1110
1023.3416	1023.3365	1023.3416	0.0051	0.0051
1134.2614	1134.3684	1134.2614	0.1070	0.1070
1136.3928	1136.4205	1136.3928	0.0277	0.0277
1205.1498	1205.4055	1205.1498	0.2557	0.2557
1248.4108	1250.4635	1248.4108	2.0527	2.0527

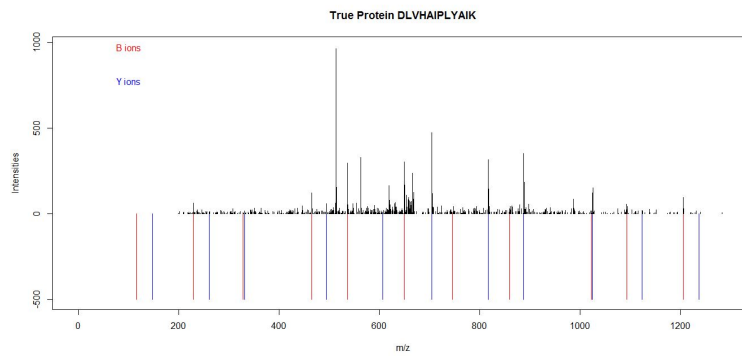
5.2.2 Example 2

Let us consider the peptide *DLVHAIPLYAIK* whose total weight is 1352.794 with 520 pairs of m/z values ranging from 199 to 1285 Da and corresponding intensities ranging from 1 to 965.

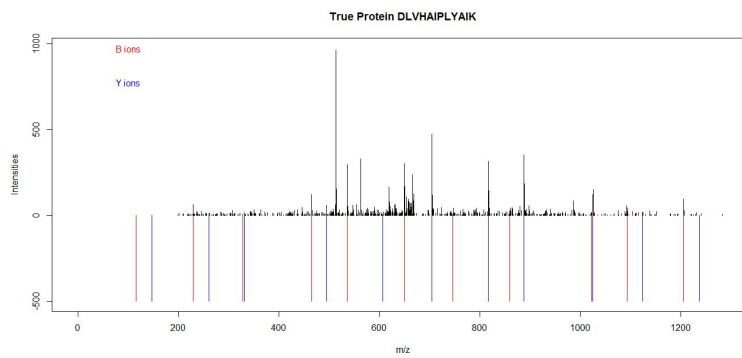
Using a 10th percentile, the threshold value was found to be 3. The average distance between m/z values is 2.0 Da and thus we set the window width to be 2.5 Da. After our applying our method, the observed spectrum was reduced from 520 pairs to 498 pairs of m/z values and their corresponding intensities.

Figure 14 (a) shows the observed spectrum before our method is applied and Figure 14 (b) shows the observed spectrum after our method is applied. Once again, we see a few of the noise peaks have been successfully removed.

Table 8 lists the closest observed m/z values for both before and after our method is applied, the theoretical m/z value, the distances between the observed m/z value before denoising and the theoretical m/z value, and the distances between the observed m/z value after denoising and the theoretical m/z value. In all cases the distances either remained the same or reduced indicating our method is doing well.



(a)



(b)

Figure 14: (a) shows the observed and theoretical spectrum before binning. (b) shows the observed and theoretical spectrum for the peptide *DLVHAIPLYAIK* after binning.

Table 8: A table comparing the distances for the peptide *DLVHAIPLYAIK* before and after binning. The values in bold indicate the distances that were reduced.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
199.2601	115.8770	199.2601	83.3831	83.3831
199.2601	146.9450	199.2601	52.3151	52.3151
228.9623	228.9610	228.9623	0.0013	0.0013
260.0568	260.0290	260.0568	0.0278	0.0278
328.8438	328.0294	327.8862	0.8144	0.1432
331.3174	331.0661	331.3174	0.2513	0.2513
465.1366	465.0884	465.1366	0.0482	0.0482
494.3502	494.1291	494.3502	0.2211	0.2211
536.3278	536.1255	536.3278	0.2023	0.2023
607.4730	607.2131	607.4730	0.2599	0.2599
649.1726	649.2095	649.1726	0.0369	0.0369
704.3751	704.2659	704.3751	0.1092	0.1092
746.8922	746.2623	746.8922	0.6299	0.6299
817.3657	817.3499	817.3657	0.0158	0.0158
859.1126	859.3463	859.1126	0.2337	0.2337
888.2925	888.3870	888.2925	0.0945	0.0945
1022.2666	1022.4093	1022.2666	0.1427	0.1427
1025.3823	1025.4460	1025.3823	0.0637	0.0637
1093.0952	1093.4464	1093.0952	0.3512	0.3512
1124.4200	1124.5144	1124.4200	0.0944	0.0944
1206.2695	1206.5304	1206.2695	0.2609	0.2609
1240.4637	1237.5984	1240.4637	2.8653	2.8653

5.3 Percentile Evaluation

Extensive experimentation showed the 60th percentile and the 10th percentile worked well for short and long peptide sequences, respectively. Other values were explored to find the optimized value. For short peptide sequences, we looked at the 70th, 80th, and 90th percentiles. For long peptide sequences, we looked at the 20th and 50th percentiles.

5.3.1 Percentile Evaluation for Peptides with Short Sequences Approach

Consider the short peptide *VINELTEK* with 348 pairs of m/z values and intensities. In our initial example, using a 60th percentile, all distances remained the same or were reduced and spectrum was reduced to 226 pairs of m/z values and intensities. When applying a 70th percentile, the data was reduced further from 226 pairs to 191 pairs of m/z values and intensities but three of the distances were increased as one can see in Table 9. This will cause the observed spectrum not to be aligned with the theoretical spectrum, which could greatly hinder the identification of this peptide. Any shift more than 0.5 Da (the standard tolerance level in most peptide identification methods) is considered severe. From Table 9, we see that distances that were increased are greater than 0.5. Similar results were seen for other peptides with short sequences when using a 70th percentile.

Table 9: A table comparing the distances for the peptide *VINELTEK* before and after binning. The bolded values indicate the distances that increased after using a threshold of 70%

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
141.9028	99.9184	144.7715	41.9844	44.8531
147.0816	146.9450	147.0816	0.1366	0.1366
213.0144	213.0024	213.9703	0.0120	0.9679
276.1071	275.9880	275.9107	0.1191	0.0773
327.0514	327.0454	327.0514	0.0060	0.0060
377.2013	377.0360	377.2013	0.1653	0.1653
456.2416	456.0884	456.2416	0.1532	0.1532
490.2245	490.1200	490.2245	0.1045	0.1045
569.1671	569.1724	569.1671	0.0053	0.0053
619.3815	619.1630	619.3815	0.2185	0.2185
670.1361	670.2204	672.2233	0.0843	2.0029
733.1962	733.2060	733.1962	0.0098	0.0098
799.2451	799.2634	799.2451	0.0183	0.0183
846.3795	846.2900	846.3795	0.0895	0.0895

Consider the peptide *VINELTEK* again with 348 pairs of m/z values and intensities. In our initial example, using a 60th percentile, all distances remained the same or were reduced and spectrum was reduced to 226 pairs of m/z values and intensities. When applying a 80th percentile, the data was reduced further from 226 pairs to 160 pairs of m/z values and intensities but four of the distances were increased as one can see in Table 10. From Table 10, we see that the four distances that increased are greater than 05. Similar results were seen for other peptides with short sequences when using this percentile value.

Table 10: A table comparing the distances for the peptide *VINELTEK* before and after binning. The bolded values indicate the distances that increased after using a threshold of 80%

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
141.9028	99.9184	144.7715	41.9844	44.8531
147.0816	146.9450	147.0816	0.1366	0.8265
213.0144	213.0024	213.0144	0.0120	0.0120
276.1071	275.9880	275.1071	0.1191	0.1191
327.0514	327.0454	327.0514	0.0060	0.0060
377.2013	377.0360	377.2013	0.1653	0.1653
456.2416	456.0884	456.2416	0.1532	0.1532
490.2245	490.1200	490.2245	0.1045	0.1045
569.1671	569.1724	573.7797	0.0053	4.6073
619.3815	619.1630	619.3815	0.2185	0.2185
670.1361	670.2204	672.2233	0.0843	2.0029
733.1962	733.2060	733.1962	0.0098	0.0098
799.2451	799.2634	799.2451	0.0183	0.0183
846.3795	846.2900	846.3795	0.0895	0.0895

Now consider the peptide *LSDYGVQLR* with 470 pairs of m/z values and intensities. In our initial example, using a 60th percentile, all distances remained the same or were reduced and spectrum was reduced to 294 pairs of m/z values and intensities.

Table 11: A table comparing the distances for the peptide *LSDYGVQLR* before and after binning. The bolded values indicate the distances that increased after using a threshold of 90%.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
155.2550	113.9340	156.8959	41.3209	42.9619
174.3954	174.9510	170.0961	0.5556	4.8549
200.1734	200.9660	201.0907	0.7926	0.1247
288.1805	288.0350	288.1805	0.1455	0.1455
315.6753	315.9930	316.2869	0.3177	0.2939
416.2673	416.0940	416.2487	0.1733	0.1547
479.0506	479.0560	479.0506	0.0054	0.0054
515.3530	515.1624	515.3530	0.1906	0.1906
536.1034	536.0775	538.5869	0.0259	2.5094
572.2805	572.1839	572.2805	0.0966	0.0966
635.0522	635.1459	635.0522	0.0937	0.0937
735.3122	735.2469	735.3122	0.0653	0.0653
763.1658	763.2049	763.1658	0.0391	0.0391
850.3229	850.2739	850.1546	0.0490	0.1193
876.1458	876.2889	876.1458	0.1431	0.1431
937.3237	937.3059	937.3237	0.0178	0.0178

When applying a 90th percentile, the data was reduced further from 294 pairs to 183 pairs of m/z values and intensities but four of the distances were increased as one can see in Table 11. From Table 11, we see the four distances that were increased are greater than 05. Similar results were seen for other peptides with short sequences

when using this percentile value.

5.3.2 Percentile Evaluation for Peptides with Long Sequences Approach

Consider the peptide *ENLMQVYQQAR* with 590 pairs of m/z values and intensities. In our initial example, using a 10th percentile, all distances remained the same or were reduced and spectrum was reduced to 548 pairs of m/z values and intensities. When applying a 20th percentile, the data was reduced further from 548 pairs to 529 pairs of m/z values and intensities but three of the distances were increased as one can see in Table 12. From Table 12, we see that the three distances that were increased are greater than 0.5. Similar results were seen for other peptides with long sequences using a threshold of 20%.

We consider another example of the long peptide *DLVHAIPLYAIK* with 520 pairs of m/z values and intensities. In our initial example, using a 10th percentile, all distances remained the same or were reduced and spectrum was reduced to 498 pairs of m/z values and intensities. When applying a 50th percentile, the data was reduced further from 498 pairs to 363 pairs of m/z values and intensities but five of the distances were increased as one can see in Table 13. From Table 13, we see that the five distances that were increased are greater than 0.5. Similar results were seen for other peptides with long sequences with a threshold of 50%.

Table 12: A table comparing the distances for the peptide *ENLMQVYQQAR* before and after binning. The bolded values indicate the distances that increased after using a threshold of 20%.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
200.0935	129.8930	201.5275	70.2005	71.6345
200.0935	174.9510	201.5275	25.1425	26.5765
243.8989	243.9360	243.8989	0.0371	0.0371
246.0997	245.9881	246.0997	0.1116	0.1116
357.1619	357.0200	357.1619	0.1419	0.1419
374.2530	374.0471	374.2530	0.2059	0.2059
488.0701	488.0600	488.0701	0.0101	0.0101
502.2154	502.1061	502.2154	0.1093	0.1093
616.1608	616.1190	616.1608	0.0418	0.0418
665.2491	665.1691	665.2491	0.0800	0.07998
715.1245	715.1874	715.1245	0.0629	0.0629
764.2294	764.2375	764.2294	0.0081	0.0081
878.1607	878.2504	878.1607	0.0897	0.08969
892.3641	892.2965	892.3641	0.0676	0.06757
1006.1984	1006.3094	1006.1984	0.1110	0.1110
1023.3416	1023.3365	1023.3416	0.0051	0.0051
1134.2614	1134.3684	1134.2614	0.1070	0.1070
1136.3928	1136.4205	1136.3928	0.0277	0.0277
1205.1498	1205.4055	1205.1498	0.2557	0.2557
1248.4108	1250.4635	1247.5865	2.0527	2.876967

Table 13: A table comparing the distances for the peptide *DLVHAIPLYAIK* before and after binning. The bolded values indicate the distances that increased after using a threshold of 50%.

Observed m/z before method	Theoretical m/z	Observed m/z after denoising	Difference before denoising	Difference af- ter denoising
199.2601	115.8770	200.9872	83.3831	85.1102
199.2601	146.9450	200.9872	52.3151	54.0422
228.9623	228.9610	228.9623	0.0013	0.00133
260.0568	260.0290	261.2964	0.0278	1.2674
328.8438	328.0294	330.5817	0.8144	2.5523
331.3174	331.0661	331.3174	0.2513	0.2513
465.1366	465.0884	465.1366	0.0482	0.0482
494.3502	494.1291	494.3502	0.2211	0.2211
536.3278	536.1255	536.3278	0.2023	0.2023
607.4730	607.2131	607.4730	0.2599	0.2599
649.1726	649.2095	649.1726	0.0369	0.0369
704.3751	704.2659	704.3751	0.1092	0.1092
746.8922	746.2623	746.8922	0.6299	0.6299
817.3657	817.3499	817.3657	0.0158	0.0158
859.1126	859.3463	860.2671	0.2337	0.9208
888.2925	888.3870	888.2925	0.0945	0.0945
1022.2666	1022.4093	1022.2666	0.1427	0.1427
1025.3823	1025.4460	1025.3823	0.0637	0.0637
1093.0952	1093.4464	1093.0952	0.3512	0.3512
1124.4200	1124.5144	1124.4200	0.0944	0.0944
1206.2695	1206.5304	1206.2695	0.2609	0.2609
1240.4637	1237.5984	1234.8877	2.8653	2.7107

5.4 More Examples

Table 14 lists other peptides in which our method was applied. In each example, the distances of concern either remained the same or reduced. This gave us indication that our method is performing well. In each example, the alignment of the theoretical and observed spectrum either remained the same or was improved by the reduction of some distances.

Table 14: A table showing results of some peptides using our method.

Short sequence peptides	Long sequence peptides
<i>VSGQTVR</i>	<i>ENLMQVYQQAR</i>
<i>TGMSNVSK</i>	<i>DLVHAIPLYAIK</i>
<i>PAVAMLEER</i>	<i>AGSGALTLGQPNSPGVPADFAK</i>
<i>ALNLQDK</i>	<i>AFNEMQPIVDR</i>
<i>FGSELLAK</i>	<i>AAAAPVTGPLADFPIQETITFDDFAK</i>
<i>FNDAVIR</i>	<i>GASQNIIPSSTGAAK</i>
<i>IGENINIR</i>	<i>HSSTISDPDTNVK</i>
<i>ISDIPEFVR</i>	<i>IDVEGSNEMGQLAENLR</i>
<i>VINELTEK</i>	<i>KAVLLPGDLSDESFAR</i>
<i>TLNDAVEVK</i>	<i>TVGKPVETVPQIFVDQK</i>
<i>LSDYGVQLR</i>	

6 DISCUSSION

We presented a method that denoises the spectrum and thus aid in the better identification of peptides. The identification of peptides have proved to be vital in determining diseases in its early stages especially in cancer. Early detection of cancer conclusively leads to a better chance of recovery. With the help of advance technologies being developed, protein identification has become less cumbersome for early discoveries of diseases. Being able to identify peptides that have not been cultured will help researchers immensely. It is the hope that scientists will one day expand their research in developing drugs that pertains to specific diseases made for an individual with minimum unwanted secondary effect.

Our focus in this thesis was on random noise, however it was stated in Chapter 2 that another type of noise is chemical noise. Wavelet thresholding is a type of pre-processing procedure to reduce chemical noise in the data. In this approach, coefficients are computed based on the data and the coefficients are compared with a threshold [35]. It is assumed that that there are n noisy samples of a certain function f ;

$$y_i = f_{t_i} + \sigma \epsilon_i \dots(1)$$

where $i = 1, \dots, n$, ϵ_i are independent and identically distributed (iid) $N(0, 1)$ and σ is the level of the noise. The noise level may be known or sometimes unknown as in figure 29, which shows a signal which is clean and a noisy signal. The wavelet coefficient is calculated from Equation (1) and is given as

$$y_{jk} = w_{jk} + \sigma^2 \epsilon_{jk} \dots(2)$$

where w_{jk} are the wavelet coefficients and ϵ_{jk} are independent and identically dis-

tributed (iid) $N(0, 1)$.

In this approach, coefficients with a minimum magnitude are thought to be pure noise and therefore set to zero. There are two coefficients calculated from Equation 2, detailed coefficients and approximation coefficients. Wavelet coefficient thresholding is most often used on the detail coefficients. The reason being that, the approximation coefficients contain some important signals with less noise [35]. We believe the use of wavelets to remove chemical noise coupled with our method to reduce the random noise, will make peptide identification less cumbersome.

Tandem mass spectrometry produces considerably huge amount of data that is used for analysis of proteins. Data produced by tandem mass spectrometry in its raw form is polluted by noise. The method used in this thesis lessens the effect of the noise inhibiting the process of peptide identification.

BIBLIOGRAPHY

- [1] Berman H.M., Battistuz T., Bhat T.N., Bluhm W.F., Bourne P.E., Burkhardt K., Feng Z., Gilliland G.L., Iype L., Jain S., et al., *Searching for hypothetical proteins: Theory and practice based upon original data and literature*. Acta Crystallogr. D Biol. Crystallogr. 2002;58:899907.
- [2] Lubec G., Afjehi-Sadat L., Yang J.W., John J.P., *The Prospects and Challenges of Proteomics*. Prog. Neurobiol. 2005;77:90127.
- [3] The Catalyst-Chemistry Resource for the Secondary Education Teacher, *The Prospects and Challenges of Proteomics*. Roland Annan
- [4] Welcome Genome Campus, Public Engagement. Your Genome-What is a genome?. Retrieved from: <http://www.yourgenome.org/facts/what-is-a-genome>.
- [5] Celis, J. E.; Ostergaard, M.; Jensen, N. A.; Gromova, I.; Rasmussen, H. H.; Gromov, P., "Human and mouse proteomic databases: novel resources in the protein universe.". FEBS Lett 1998, 430, 64-72.
- [6] Biointelligence: Education, Training and Consultancy Services. Proteomics: Challenges and Approaches. Retrieved from: <https://biointelligence.wordpress.com/tag/scope-of-proteomics/>.
- [7] Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon - The human microbiome project: exploring the

microbial part of ourselves in a changing world. Retrieved from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709439/>.

- [8] IonSource - Mass Spectrometry Educational Resource. Peptide-Mass Fingerprinting. Retrieved from <http://www.ionsource.com/tutorial/protID/fingerprint.htm>.
- [9] Creative Proteomics - Protein Identification. Retrieved from: <http://www.creative-proteomics.com/services/peptide-mass-fingerprinting-pmf.htm>.
- [10] Bioinformatics solutions - De novo sequencing introduction. Retrieved from: <http://www.bioinform.com/denovo-tutorial/>.
- [11] University of Otago - Center for Protein Research - De novo sequencing. Retrieved from: <http://cpr.otago.ac.nz/workflows/de-novo-sequencing/>.
- [12] Celis JE, Gromov P. 2D protein electrophoresis: can it be perfected? *Current Opinion in Biotechnology*: 1999;10(1):1621.
- [13] Seibert V, Ebert MPA, Buschmann T. Advances in clinical cancer proteomics: SELDI-ToF-mass spectrometry and biomarker discovery. *Briefings in Functional Genomics and Proteomics*.: 2005;4(1):1626.
- [14] Coombes, K.R., K.A., and Morris, J.S. (2007). Pre-processing mass spectrometry data. In Dubitzky, M. Granzow, M. and Berrar, D., editors *Fundamentals of Data Mining in Genomics and Proteomics*.: pages 79-99: Kluwer.

- [15] Washington University in St. Louis. Tandem Mass Spectrometry. Retrived from:
<http://msr.dom.wustl.edu/tandem-mass-spectrometry/>
- [16] Domon B, Aebersold R. Challenges and opportunities in proteomics data analysis. *Molecular and Cellular Proteomics.*: 2006;5(10):19211926.
- [17] Papale M, Pedicillo MC, Di Paolo S, et al. *Saliva analysis by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF/MS): from sample collection to data analysis.* Clin Chem Lab Med. 2008;46:8999
- [18] Written by Kara Rogers: What is the difference between a peptide and a protein. Retrieved from: <https://www.britannica.com/demystified/what-is-the-difference-between-a-peptide-and-a-protein>.
- [19] Schipper R, Loof A, de Groot J, Harthoorn L, Dransfield E, van Heerde W. SELDI-TOF-MS of saliva: methodology and pre-treatment effects. J Chromatogr B Anal Technol Biomed Life Sci. 2007;847:4553.
- [20] Christopher A. Crutchfield, Stefani N. Thomas, Lori J. Sokoll and Daniel W. Chan, "*Advances in mass spectrometry-based clinical biomarker discovery*". Clinical Proteomics,2016
- [21] Overview of Mass Spectrometry for Protein Analysis. Retrieved from: <https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-mass-spectrometry.html.html>

- [22] Figure as originally published in Metpally RPR, Nasser S, Malenica I, Courtright A, Carlson E, Ghaffari L, Villa S, Tembe W and Van Keuren-Jensen K (2013). *Front. Genet.* 4:20. doi: 10.3389/fgene.2013.00020. Retrieved from: <http://journal.frontiersin.org/article/10.3389/fmicb.2015.00791/full>
- [23] Reusch, William (5 May 2013). *Peptides and Proteins*. Michigan State University Department of Chemistry
- [24] Dancik, V., Addona, T.A., Clauser, K.R., and Vath, J. E (1999). De novo peptide sequencing via tandem mass spectrometry: A graph-theoretical approach. In *RECOMB '99: Proceedings of the third annual international conference on Computational molecular biology*, number 135-144, New York, NY, USA. ACM Press.
- [25] Hochstrasser DF (1997) Clinical and biomedical applications of proteomics. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser DF, editors. *Proteome research: New frontiers in functional genomics*. Berlin/Heidelberg: Springer-Verlag. pp. 187-220.
- [26] Monroe, M. Peptide sequence fragmentation modeling. Retrieved from: <http://www.alchemistmatt.com/MwtHelp/PeptideFragModelling.htm>
- [27] Jian Liu, Alexander W Bell, John JM Bergeron, Corey M Yanofsky, Brian Carrillo, Christian EH Beaudrie and Robert E Kearney. *Proteome Science Methods for peptide identification by spectral comparison*.

- [28] Nguyen Hung Son. - Data cleaning and Data Preprocessing. Retrieved from:
<https://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf>
- [29] Aebersold, R., and Mann, M. (2003). *Mass spectrometry-based proteomics*. Nature 422: 198-207.
- [30] Mass Spectrometers - A short explanation for the absolute novice. Retrieved from: <http://www.research.uky.edu/core/massspec/jeolnovice.pdf>
- [31] University of California - How the mass spectrometer works. Retrieved from:
http://chem.libretexts.org/Core/Analytical_Chemistry/Instrumental_Analysis/Mass_Spectrometry/How_the_Mass_Spectrometer_Works
- [32] Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Retrieved from:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418498/>
- [33] Protein Chemistry Lab (Texas A and M University).
Protein Identification-Proteomics. Retrieved
from: <http://tamupcl.com/Services/ProteinIdentificationProteomics>
- [34] What is binning?. Retrieved from: https://docs.tibco.com/pub/spotfire/7.0.1/doc/html//bin/bin_what_is_binning.htm
- [35] Rami Cohen. Department of Electrical Engineering. Technion, Israel Institute of Technology. *Single Denoising Using Wavelets*. Nature 422: 198-207.
- [36] Kermit Murray. Louisiana State University. *Schematic representation of a tandem mass spectrometry experiment*. 13 May 2006.

VITA

FELIX OFFEI

Education: M.S. Mathematical Sciences,
East Tennessee State university, 2017

B.S. Mathematics
Kwame Nkrumah Univ. of Science and Technology, 2011

Professional Experience: Introduction to Prob. and Stat, Instructor
Department of Mathematics and Statistics,ETSU
Aug. 2016 - May 2017

Scholar Advisor
Envision Experience,VA
June. 2016 - July 2016

Mathematics and Statistics Tutor
Center for Academic Achievement,ETSU
Aug. 2015 - May 2016

Professional Development: Statistical and Mathematical:
SAS, R, SPSS, Minitab, GLPK

Programming Languages:
Visual Basic, Python

Database Server:
SQL Server, MySQL, PostgreSQL

Professional Development: Microsoft Office Suite:
MS Access , Word, Excel, PowerPoint, Publisher, Outlook

Graphic/Web design:
Adobe Creative Suite, Wordpress, Dreamweaver