5-2019

A Comparison of Standard Denoising Methods for Peptide Identification

Skylar Carpenter
East Tennessee State University

A Comparison of Standard Denoising Methods for Peptide Identification

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

Skylar Carpenter

May 2019

Christina Nicole Lewis, Ph.D., Chair

JeanMarie Hendrickson, Ph.D.

Jeff Knisley, Ph.D.

Keywords: Protein identification, peptides, MCMC, tandem mass spectrometry, denoising spectra, baseline removal, binning, wavelets

ABSTRACT

A Comparison of Standard Denoising Methods for Peptide Identification

by

Skylar Carpenter

Peptide identification using tandem mass spectrometry depends on matching the observed spectrum with the theoretical spectrum. The raw data from tandem mass spectrometry, however, is often not optimal because it may contain noise or measurement errors. Denoising this data can improve alignment between observed and theoretical spectra and reduce the number of peaks. The method used by Lewis et. al (2018) uses a combined constant and moving threshold to denoise spectra. We compare the effects of using the standard preprocessing methods baseline removal, wavelet smoothing, and binning on spectra with Lewis et. als threshold method. We consider individual methods and combinations, using measures of distance from Lewis et. al's scoring function for comparison. Our findings showed that no single method provided better results than Lewis et. al's, but combining techniques with that of Lewis et. al's reduced the distance measurements and size of the data set for many peptides.

ACKNOWLEDGMENTS

I would like to thank Dr. Nicole Lewis for her continued support and guidance over the last two years, from my first day as a TA for her until now. Without her counsel and feedback, I would not be where I am today. I would also like to thank Dr. Jeff Knisley and Dr. JeanMarie Hendrickson for agreeing to be on my thesis committee. I am grateful to the entire mathematics department at ETSU for giving me the opportunity to pursue this degree. Lastly, I would like to thank my family and friends for their encouragement throughout this endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

A genome is the set of complete genetic material of an organism. This genetic material, most commonly DNA in the form of genes, is found in every cell of the organism and contains the genes that code for the production of various molecules that are involved in every process the cell will carry out in its lifetime. One class of such molecules that is of great importance is proteins.

Proteins are large molecules that play a variety of roles and functions in organisms, such as antibodies that protect from viruses and bacteria, enzymes that help carry out chemical reactions within cells, or messengers that transmit signals throughout the organism. Other proteins are part of the structure of the organism itself. The complete set of proteins produced by an organism is known as a proteome. Proteomics, the study of proteomes, is a growing field where researchers face a task more complex than genomics due to the fact that an organism's genome remains relatively constant, whereas the proteome may vary throughout the life of the organism and may be subject to modifications depending on environmental factors and needs [2]. Identifying these proteins helps increase our understanding of cellular processes, can be used to test for disease, and aids in the development of new pharmaceutical drugs.

There are many obstacles when performing protein identification. When the genome of the organism from which the protein comes has not been fully identified, it is more difficult to correctly identify the protein. Microbial proteins in particular are difficult to identify, since only a small percent (1-10%) of microbes can be cultured due to the inability to perfectly emulate their ideal environmental conditions in the lab [3]. Even if the microbe has been cultured and its genome has been sequenced, there is no way to predict what modifications its proteome may display. Thus, knowing the genome does not equate to knowing the proteome. Being able to accurately identify proteins is applicable to many fields: it is essential for understanding biological systems and the interactions that lead to cell signaling cascades; microbial organisms can be classified based on their proteomes; and proteins can serve as disease markers that simplify diagnoses.

Current methods of protein identification are limited in their usefulness and accuracy. Mass spectrometry is one of the most commonly used techniques, but a major issue that arises is that the data obtained is often incomplete and/or noisy, that is, it contains errors, contaminants, or some other abnormalities that make analyzing it in its raw form challenging. The noisiness of mass spectrometry data reduces its accuracy in protein identification; however, the usefulness and accuracy of mass spectrometry can be improved when the noise is reduced or removed. There are several proposed techniques for reducing the noise of mass spectrometry data, some more effective than others. A peptide identification method proposed by Lewis, Hitchcock, Dryden, and Rose (2018), which uses a Bayesian approach to identification, employs a method of denoising the data unique from those of more standard approaches [1].

1.1 Proposed Work

Lewis et. al's (2018) approach to peptide identification employs a Bayesian stochastic search that uses the prior knowledge of abundances of bond cleavages and the probability of specific amino acid sequences. The scoring function measures the closeness of each observed m/z value to a theoretical m/z value. A Markov chain

Monte Carlo (MCMC) scheme is used to simulate candidate peptides from the posterior distribution. The peptide with the largest posterior probability is chosen as the estimate for the true peptide. The method used by Lewis et. al uses a combined constant and moving threshold preprocessing technique to denoise and reduce mass spectrum data. We compare the results of this method with those of binning, wavelets, and baseline removal, which are standard denoising techniques. We consider each technique individually as well as in various combinations, which are described in Chapter 5. The objective of these techniques is to denoise and reduce the size of the data while producing the same or more accurate results. We use Lewis et. al's scoring function as our measure of closeness for the basis of comparisons. The data that we are using comes from the Pacific Northwest National Laboratory (PNNL), which can be publicly accessed online. The data were obtained from an LTQ Orbitrap yielding doubly charged tryptic peptides. For each of the 1,026 peptides in the dataset, there is a corresponding set of m/z (mass-to-charge) values and intensity values.

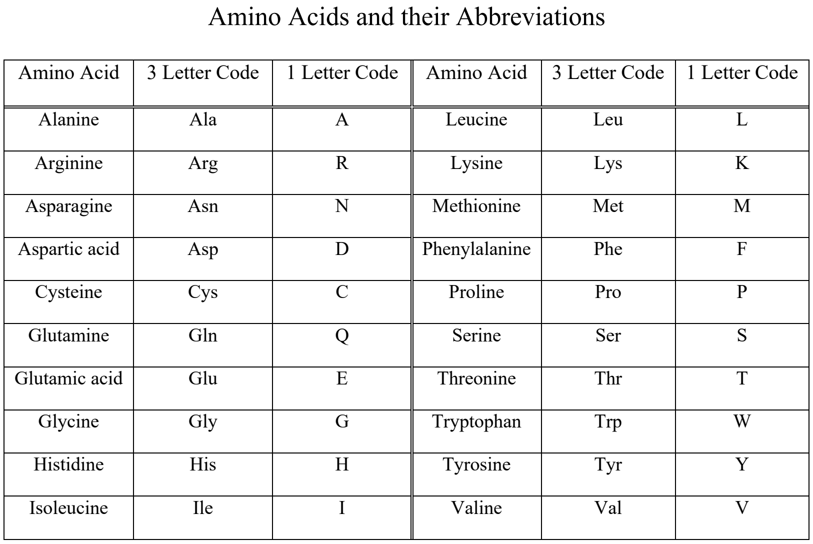
1.2 Overview of Thesis

The thesis is arranged as follows. Chapter 2 is an overview of proteins and peptides, their functions, and the methods currently used to identify them. Chapter 3 covers peptide fragmentation and the use of fragmentation to construct the theoretical spectrum. Chapters 4 describes the Bayesian method and the MCMC algorithm. Chapter 5 provides an overview of standard denoising techniques as well as the method used by Lewis et. al. Chapter 6 presents the results using these denoising methods. The thesis is concluded in Chapter 7.

2 PROTEINS AND PEPTIDES

Proteins are composed of a string of smaller molecules called amino acids. There are twenty different amino acids from which a protein can be made, and these strings may be thousands of amino acids long. Amino acids get their name from their structural similarities: each one is composed of an amine group (-NH_2) and a carboxyl group (-COOH) together with a side chain known as the R group that is specific to each of the 20 amino acids. The order in which the amino acids appear is known as the primary structure, and this determines not only the form that the protein will take but also what function it will have. Table 1 lists each of the amino acids.

Table 1: The 20 amino acids along with their one and three letter abbreviations



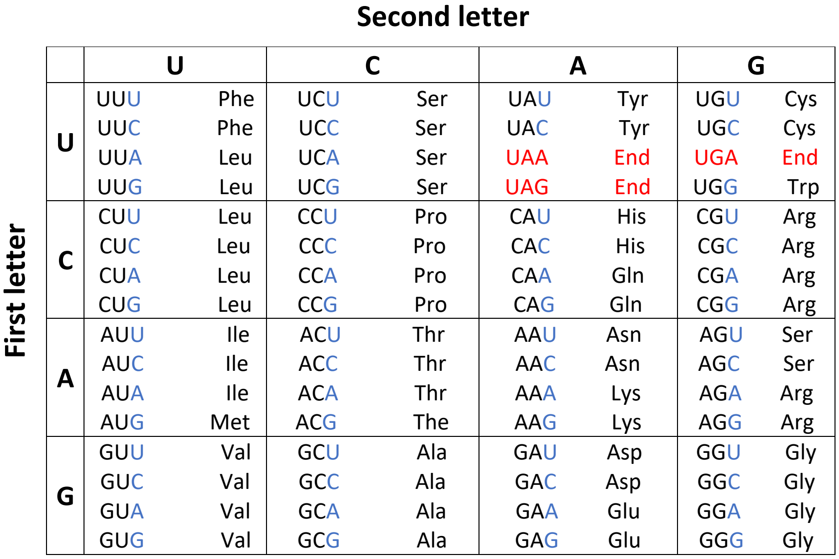
Amino Acids and their Abbreviations

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	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

Deoxyribonucleic acid, or DNA, carries all of the genetic information in every cell of our bodies. DNA is composed of nucleotides connected in the form of a double helix. Each nucleotide consists of a nucleobase, a sugar called deoxyribose, and a phosphate group. There are four nucleobases: cytosine (C), guanine (G), adenine (A), and thymine (T). The sugar group on one nucleotide bonds to the phosphate group of the next, forming a chain. There are two such chains composed of nucleotides held together by covalent bonds running antiparallel to one another. The two chains are connected by bonding between the nucleobases across from each other: A bonds with T, and G bonds with C. Each connected segment between the chains is called a base pair. In humans, there are approximately 3 billion of these base pairs [4]. The sequence of these nucleobases encodes genetic information for the production of each of the twenty amino acids. Genome sequencing is the process of identifying the order of all of the DNA nucleotides in the genome.

Through a process called transcription, DNA strands are used as a template to create ribonucleic acid (RNA) strands [5]. These RNA strands then undergo a process known as translation that converts the RNA to protein. More specifically, the RNA strand specifies the amino acid sequence for a protein. Each set of three nucleobases is known as a triplet codon. Since there are four nucleotides, there are 64 different triplet codon combinations. Triplet codons correspond to certain amino acids; for example, the amino acid alanine is coded by the triplet GCU. Since there are only 20 amino acids, there is some redundancy in the triplet codons. For example, GCA and GCG both code for alanine. Out of the 64 triplet combinations, 60 code for amino acids, one can code for the amino acid methionine or serve as a start codon that signals the beginning of translation (when at the beginning of the RNA strand), and three are known as stop codons that signal for translation to end. Table 2 shows each of the triplet combinations and their result.

Table 2: The 64 triplet codon combinations



Second letter

First letter

	U		**C**		**A**		**G**	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	End	UGA	End
	UUG	Leu	UCG	Ser	UAG	End	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	The	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Peptides and proteins are both made up of strings or chains of amino acids connected by amide (or peptide) bonds. Generally, a peptide is made up of a chain between 2 and 50 amino acids long, while proteins may have thousands. Fundamentally, then, a protein is just a very large peptide. Since proteins are longer, they tend to fold and twist into complex structures. Peptides, being much shorter, do not exhibit this behavior.

2.1 Protein in the Human Body

In the human genome consisting of an estimated 30,000 genes, each gene is responsible for the production of an average of three proteins [4]. Proteins are second only to water in the composition of the human body, with the average human being approximately 16-17% protein. There are many different types and classes of protein. An enzyme is a type of protein that works as a catalyst, which accelerates chemical reactions. Without these enzymes, many biological reactions necessary for survival would either not take place at all or would happen too slowly to sustain life. Proteins called antibodies are produced by B-lymphocyte cells of the immune system in response to foreign substances called pathogens. These pathogens are often from bacteria or viruses that have infected the body. These pathogens have another type of protein, called antigens, that the antibodies respond to by binding, which can directly incapacitate the foreign cell or triggers other cells in the body to attack. Different pathogens will have different antigens; only antibodies specific to that antigen will be able to bind to it. Thus, the human body must be able to produce many different antibodies in order to successfully defend itself [6].

Proteins are also involved in the structure and movement of the body. Keratin is a structural protein that can be found in skin, hair, and nails. Collagen is a protein that makes up fibers in muscles, tendons, ligaments, and bones. Actin and myosin are two proteins that assist in muscle contraction, allowing movement of the body. Many of the hormones that transport signals throughout the body are proteins. Insulin, the hormone that regulates the uptake of glucose into cells, is a protein produced in the pancreas. Vasopressin, also known as antidiuretic hormone (ADH), is a protein produced by the hypothalamus that regulates the amount of water in the blood. Yet another role of proteins in the human body is transportation. Channel proteins allow molecules such as water, sodium, and potassium to cross the cell membrane. Hemoglobin is a protein produced by red blood cells that carries oxygen from the lungs to body tissues and then carries carbon dioxide back to the lungs.

2.2 Peptide Identification

While an organism's genome generally remains constant, the proteome is subject to change from cell to cell depending on factors such as time, stresses, needs, and modifications. Gene expression depends on the cell type, so different cells will produce different proteins. Furthermore, alternative splicing of genes can allow the production of multiple proteins from a single gene, increasing the size of the proteome even more [7]. Post-translational modifications such as phosphorylation or methylation may alter or activate the function of a protein. Thus, knowing an organism's genome does not equate to knowing its proteome, which is much larger and more complex.

One of the most common tools in peptide identification is mass spectrometry (MS). While MS was originally used to measure atomic weights of elements, it has been increasingly used in biological sciences for the purpose of identifying peptides. With the development of new ionization methods, such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), mass spectrometry has become indispensable in studying proteins [8].

In MS, a sample of the purified peptide is first vaporized and ionized (charged by gaining or losing electrons) before being passed through the spectrometer, which sorts and separates the ions based on their mass and charge. The ions accelerate and deflect at different speeds and angles depending on their mass. The ions reach the detector in order of increasing mass (since lighter ions will travel faster), and the data is translated by a computer into a mass spectrum, which shows the relative abundance of ions according to their mass-to-charge (m/z) ratio. This spectrum can be used to predict the peptide by comparing it to previously identified peptides in databases.

A mass spectrometer has three main components: an ion source, a mass analyzer, and an ion detector, which are illustrated in Figure 1. These components may vary depending on the sample and the goal. In proteomics, common types of mass analyzers are time-of-flight (TOF), quadrupole, and ion trap.

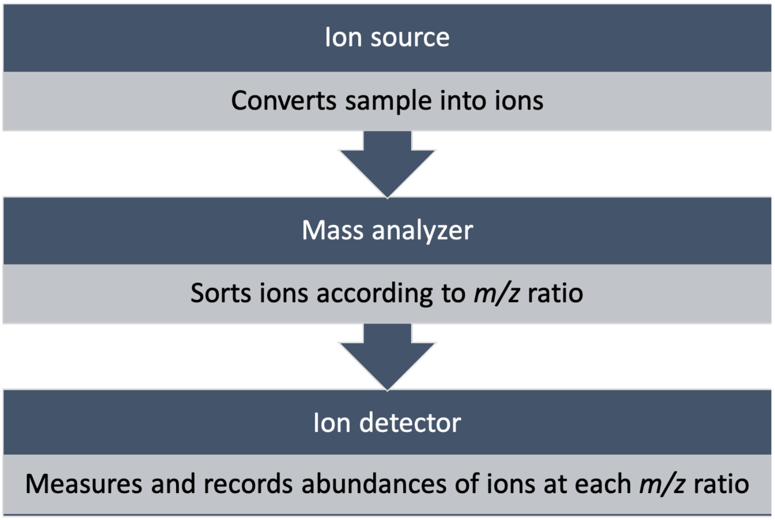

Ion source
Converts sample into ions
Mass analyzer
Sorts ions according to *m/z* ratio
Ion detector
Measures and records abundances of ions at each *m/z* ratio

Figure 1: Basic mass spectrometer components

The matrix-assisted laser desorption/ionization - time of flight mass spectrometer (MALDI-TOF MS) is one of the most common setups in proteomics. In MALDI-TOF

MS, the ion source is a laser beam that first ionizes a matrix consisting of crystallized molecules. The matrix then transfers part of its charge to the peptide sample, ionizing it without subjecting the fragile molecules directly to the laser. The ions are then separated by the mass analyzer according to their m/z ratios, which is estimated by measuring the time of flight of the ions. This is done by accelerating the ions via an electric field and passing them through an analyzer tube. The time of flight for the ions is directly proportional to the m/z ratio, thus faster moving ions have greater m/z ratios. Finally, the detector measures the number of ions at each m/z ratio. The m/z ratios along with the relative abundance is recorded for each ion, and this information is displayed in the form of the mass spectrum.

The use of magnetic fields in mass spectrometers can lead to degradation of the resolution and cause the instrument itself to lose calibration. An alternative approach is to replace the magnetic field with alternating quadrupolar electric fields, which is the setup used in the appropriately named quadrupole mass spectrometers [9]. The quadrupole consists of four cylindrical rods placed parallel to one another. Voltage is applied to these rods to create the electric field. The ions pass through the quadrupole at trajectories subject to the oscillations of the electric field; separation is achieved based on these oscillations, as only ions of certain m/z ratios will reach the detector, while the rest will collide with the rods.

Since peptides are typically large molecules, a common technique for obtaining peptide mass spectra is tandem mass spectrometry, or MS/MS. In its simplest form, MS/MS combines two mass spectrometers. The sample is passed through the first mass spectrometer, where it is ionized and separated based on the m/z ratio, and then ions of a given mass are selected to be fragmented into smaller particles before again being separated based on m/z ratio and finally reaching the detector. This fragmentation is often done by a process called collision-induced dissociation (CID), which involves introducing a neutral gas (such as argon) that collides repeatedly with the ions. This fragmentation results in a series of b and y ions. These fragmented ions then pass through the second mass spectrometer, producing the mass spectrum.

2.3 Current Identification Methods

Once the mass spectrum has been obtained, there are generally two approaches to identifying the peptide sequence. The first approach is to use a database of known peptides; the observed spectrum is compared against the database, where the peptide with the highest matching score will be selected as the true peptide [10]. There are numerous software applications available, each with their own algorithms for identification, such as Mascot and SEQUEST [11].

Since the peptide's mass is given by the mass spectrometer, Sequest uses this information to narrow down the set of possible peptide sequences to those with masses close to that of the observed peptide. For each of these candidate peptides, the software then creates a theoretical mass spectrum to compare with the observed peptide's mass spectrum. The candidate peptide sequence that best matches the observed data is chosen as the predicted peptide sequence. Similar to Sequest, the Mascot software uses the mass spectrometry data to identify peptides, but it also incorporates other search methods into its algorithm. Peptide mass fingerprinting, is a technique in which a protein is cleaved into peptides which are then passed through a mass spectrometer to determine their masses. These peptides are then compared to a database of known proteins and the peptides composing them, searching for the best match. Sequence query, in which the peptide mass data is combined with prior knowledge of amino acid sequences, is also used as part of Mascot's search algorithm.

De novo peptide sequencing is another approach to identifying the peptide sequence that does not require the use of a database, thus, it can be used for both known and novel peptides. The mass differences between fragment ions in the mass spectrum can be used to determine the amino acid residue for each ion; that is, the known masses for each amino acid can be used in conjunction with the data obtained from MS of the ions and their masses to determine the amino acid sequences of ions based on the mass differences between them. This process can be continued until the peptide is sequenced. Many different algorithms exist for automating this de novo sequencing, such as PEAKS and PepNovo. PEAKS uses the de novo sequencing approach alongside database searching and also compares the results of other search engines. The PepNovo software employs a probabilistic network to model the fragmentation and likelihood ratio hypothesis tests to select the best estimate for the peptide [12].

Both the database approach and de novo sequencing approaches can suffer if the data is noisy. When using the database approach with a spectrum containing too much noise, the observed spectrum may not align properly with the true spectrum, causing the peptide to be misidentified. With de novo sequencing, since the mass spectrum does not explicitly identify what type of ions correspond with each peak, noise peaks could be misidentified as true signal peaks. Additionally, not all of the

b or y ions may appear in the spectrum; this is particularly the case for the first b ion and last y ion due to their smaller masses [13]. Denoising the spectrum before identifying would result in the removal of some of the noise that is present, improving the alignment between the observed and true spectra. Additionally, denoising could reduce the dimension of the data set, i.e., decrease the number of m/z and intensity pairs, which would reduce the number of noise peaks that could potentially be misidentified as signal peaks. Thus, both approaches could benefit from denoising.

2.4 Applications of Proteomics

One major application of proteomics is pharmaceutical drug development. Proteins can play a large role in disease, and being able to block or inactivate these proteins is a form of treatment. Identifying these proteins and being able to model the 3D structure can help in the development of drugs to bond to these proteins and inactivate them.

Many proteins are involved in the production or activation of other proteins; these types of interactions are called protein-protein interactions. In systems biology, protein-protein interactions play a role in cell signaling cascades. Understanding these biological systems depends on being able to identify the proteins playing these roles.

Proteins can also serve as biomarkers, or indicators, for disease. If the proteins associated with a particular disease are known, then testing for the presence of those proteins can help in diagnosis. Testing for the presence of proteins is quicker and more efficient than blood tests.

In microbiology, proteomics can be used to identify antibiotic-resistant microbes and to classify microbes. Environmental technologists interested in the metabolic capabilities of uncultured microbes living in extreme conditions could use proteomics to better understand how these microbes thrive.

3 FRAGMENTATION

When using mass spectrometry to identify proteins, it is necessary to break the protein down into shorter peptides and examine them separately. Since proteins can be hundreds or thousands of amino acids long, it is notoriously difficult to identify intact proteins. Henceforth, we will be referring to peptides when discussing identification. During mass spectrometry, these peptides are fragmented into pairs of ions, with *b* and *y* ions being most common. These ions are detected by the mass spectrometer, and it is their intensities that are measured and recorded in the mass spectrum.

To find the *b* and *y* ions for a given peptide, one simply splits the peptide sequence into all possible sequence fragment combinations. Each *b* ion begins at the start of the peptide and ends at an amino acid with a free amine group ($-NH_2$). The *b* ions, then, have a charge on the N-terminus. The *y* ions are simply the complements of the *b* ions and have a charge on the C-terminus, which is the end with a free carboxyl group (-COOH). As an example, consider the peptide FNDAVIR. The *b* ions for this peptide would be F, FN, FND, FNDA, FNDAV, and FNDAVI. The *y* ions would be R, IR, VIR, AVIR, DAVIR, and NDAVIR. A visual representation of finding the *b* and *y* ions is shown in Figure 2.

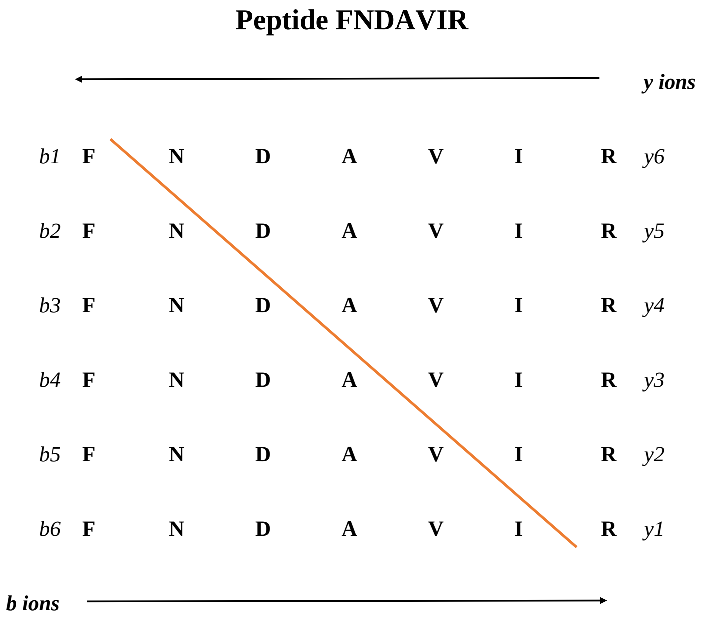
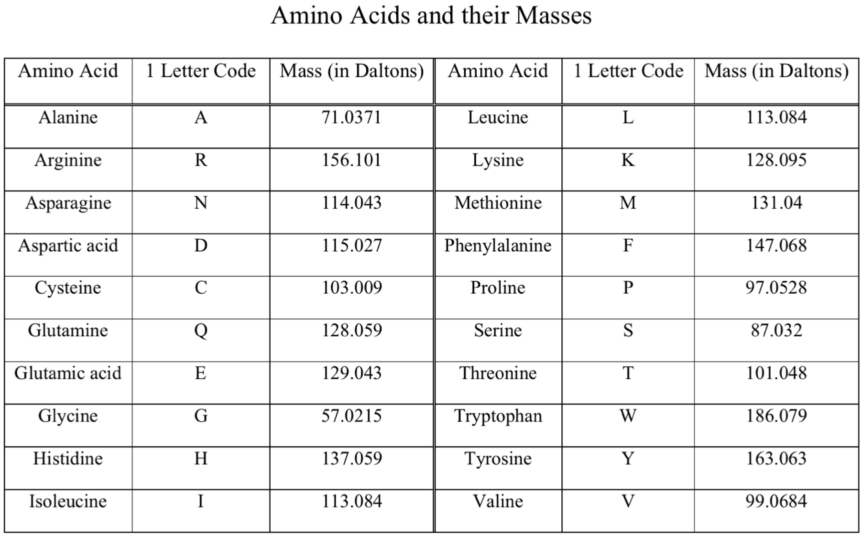


Figure 2: Illustration of the fragmentation of the b and y ions of the peptide FNDAVIR

Once the b and y ions have been identified, the mass for each ion must be determined. The mass for each ion is the sum of the masses for each amino acid in the sequence plus an offset value. We can represent this by $\sum_{i=1}^{k} m(p_i) + \epsilon$, where $m(p_i)$ is the mass of the ith amino acid, k is the total number of amino acids, i.e., the length of the peptide, and ϵ is an offset value. Table 3 shows the masses for each of the twenty amino acids.

Table 3: The 20 amino acids along with their one letter abbreviations and masses



Amino Acids and their Masses

Amino Acid	1 Letter Code	Mass (in Daltons)	Amino Acid	1 Letter Code	Mass (in Daltons)
Alanine	A	71.0371	Leucine	L	113.084
Arginine	R	156.101	Lysine	K	128.095
Asparagine	N	114.043	Methionine	M	131.04
Aspartic acid	D	115.027	Phenylalanine	F	147.068
Cysteine	C	103.009	Proline	P	97.0528
Glutamine	Q	128.059	Serine	S	87.032
Glutamic acid	E	129.043	Threonine	T	101.048
Glycine	G	57.0215	Tryptophan	W	186.079
Histidine	H	137.059	Tyrosine	Y	163.063
Isoleucine	I	113.084	Valine	V	99.0684

The offset values correspond to the peaks and are unique to the different ion types created by fragmentation. Earlier work in mass spectrometry had found that different mass spectrometers would produce different spectra for a given peptide because of differences in the offset values between spectrometers. A previous study by Dancik et al. (1999) resulted in an offset frequency function that allows one to define the ion types produced by a given mass spectrometer [14]. These offset values are given in Table 4.

Table 4: Offset values for each ion type, where M denotes $\sum_{i=1}^{k} m(p_i)$

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)

Once again, consider the peptide FNDAVIR. Using the offset frequency function and the mass of the amino acids, we find that the first b ion, F, has a mass of 147.068 + 0.85 = 147.918 Daltons (Da), while our first y ion, R, has a mass of 156.101 + 18.85 = 174.951 Da. Continuing with this process, one can find the masses for the rest of the b ions FN, FND, FNDA, FNDAV, and FNDAVI to be 261.961, 376.988, 448.025, 547.093, and 660.177 Da, respectively. Similarly, we can find the masses of the y ions IR, VIR, AVIR, DAVIR, and NDAVIR to be 288.035, 387.103, 458.14, 573.167, and 687.21 Da, respectively. Our theoretical spectrum for the peptide FNDAVIR is the set of masses: 147.918, 261.961, 376.988, 448.025, 547.093, 660.177, 174.951, 288.035, 387.103, 458.14, 573.167, and 687.21 Da. Each of these values can be seen on the (m/z) axis of theoretical spectrum of FNDAVIR in Figure 3. Note that the heights of the peaks are not meaningful here.

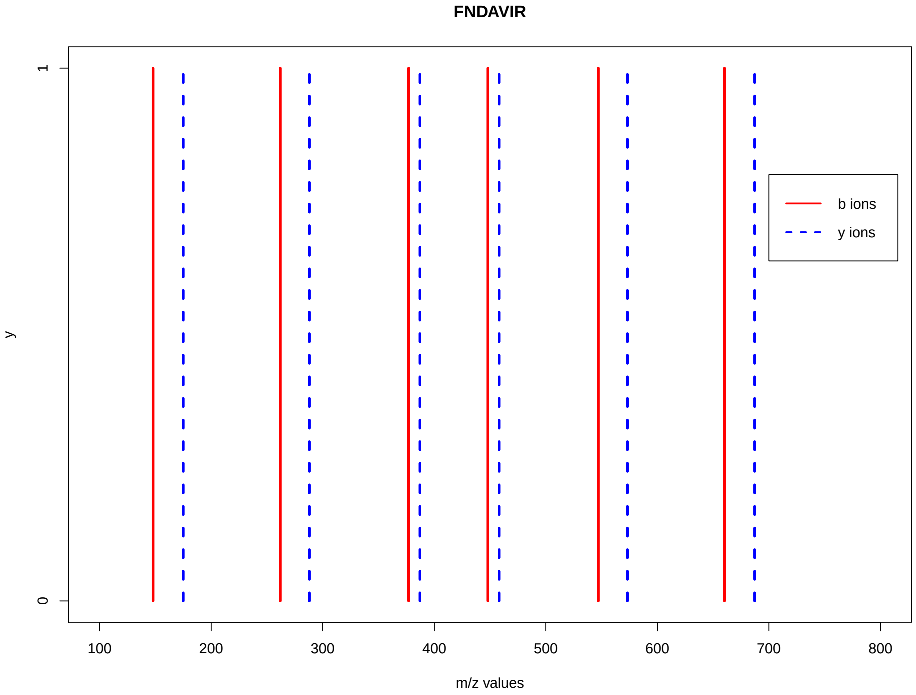


Figure 3: The theoretical spectrum for peptide FNDAVIR

The total mass of the peptide can be used to eliminate candidate peptides that do not fall within a tolerance range. The total mass of the peptide can be found by summing the individual masses for the amino acids in the peptide plus the mass of water, which is 18.010565 Daltons. We can represent this by $\sum_{i=1}^{k} m(p_i)+$ mass of H_20, where k is the number of amino acids in the peptide sequence and $m(p_i)$ is the mass for each amino acid. For data that are doubly charged, i.e., data that have acquired a second proton in the form of a hydrogen molecule, this total mass becomes $\sum_{i=1}^{k} m(p_i)+$ mass of H_2O + H, where H is 1.00794 Da. Thus, the total mass of the peptide FNDAVIR is 834.447 Da, assuming that the ion is doubly charged.

4 BAYESIAN MODEL

The approach presented by Lewis et al. (2018) uses a Bayesian model to identify the true peptide based on the observed spectrum, with an MCMC algorithm used to simulate candidate peptide sequences from an approximate posterior distribution [1]. Since mass spectrometry results in data that generally contains noise from the instrument (Poisson noise), electrical system (Johnson noise), and matrix ions (chemical noise), the observed spectrum is first thresholded to prevent this noise from hindering the peptide identification [15]. Peaks that have intensities below a certain threshold will be considered noise and will be disregarded, while those above the threshold with m/z values corresponding to the b and y ions for the candidate peptide are defined as signal peaks. Since the peaks at the beginning and end of the spectrum are not always captured by the mass spectrometer, a weighted average of constant and moving thresholds is used to avoid removing true peaks misidentified as noise; this threshold will be described in greater detail in Chapter 5.

A scoring function, composed of two overall goodness of fit measures, is used to give a measure of how well the observed spectrum and theoretical spectrum agree. One overall goodness of fit measure penalizes the candidate peptide when its theoretical spectrum does not align well with the observed spectrum. Since noise is still expected even after thresholding, another overall goodness of fit measure is incorporated that penalizes a candidate peptide when the observed spectrum shows many noise peaks that do not correspond to any m/z values of the candidate's theoretical spectrum.

The scoring function used by Lewis et al. (2018) is

$$L(X|\theta, \eta, \kappa_1, \kappa_2) \propto \kappa_1^{2p} exp(-\kappa_1 S_1) \kappa_2^{t-s} exp(-\kappa_2 S_2)$$

where the parameter vector $\theta = (\tau_1^b, ..., \tau_p^b, \lambda_1^b, ..., \lambda_p^b, \tau_1^y, ..., \tau_p^y, \lambda_1^y, ..., \lambda_p^y)$, X contains the observed set of m/z values for a particular spectrum, and η represents the string of amino acids for the candidate peptide. The other parameters are defined as

- s is the combined number of b and y ions for the candidate peptide
- p is the number of b ions (or equivalently, the number of y ions)
- t is the number of peaks in a given candidate peptide
- τ_i^b and τ_i^y are the m/z values for the b and y ions of the candidate peptide
- λ_i^b and $\lambda_i^y \in \{0, 1\}$ are indicator functions that signify whether the ith ion has a corresponding observed peak, where $i = 1, ..., p$
- κ_1 and κ_2 represent weights, which play the role of concentration parameters that control how tightly concentrated the observed peaks are around their corresponding true peaks.

Here, $\lambda_i^b = 1$ denotes the presence and $\lambda_i^b = 0$ denotes the absence of a b ion at position i; this notation similarly applies to the y ions.

The two goodness of fit measures in this function are

$$S_1 = \sum_{i=1}^{p} \left(\lambda_i^b \min_{j \in S} d(x_j, \tau_i^b) + \lambda_i^y \min_{j \in S} d(x_j, \tau_i^y) \right)$$

and

$$S_2 = \sum_{j \in \mathcal{N}} \min_{i,k} |x_j - \tau_i^k|$$

where $d(x_j, \tau_i^k) = \min(|x_j - \tau_i^k|, \delta)$.

S_1 measures the closeness of the nearest observed peak to each b or y ion of the candidate peptide, while S_2 measures the closeness of the nearest candidate peak to each observed peak. A value of $\delta = 3$ was chosen by Lewis et al. (2018), as it is believed that no true signal peak will lie beyond 3 Daltons from the observed peak. The S_1 and S_2 values defined here will be used in our comparisons. Since S_1 and S_2 are measured in Da., they are not true distance measurements; however, since we are using them as a method of gauging the closeness of peaks, we refer to them as "measures of distance."

Prior knowledge of abundances of bond cleavages and the probability of specific amino acid sequences as estimated by Huang et al. (2004) are used as part of the scoring function [16]. The cleavage prior used by Lewis et. al is derived from these probabilities. The sequence prior specifies a prior distribution for a particular sequence of amino acids in a peptide, where for each pair of amino acids in the candidate peptide, the number of occurrences of that pair in the set of all known peptides from the same species is used to find the empirical probability. κ_1 and κ_2, which are concentration parameters, are assumed to have independent Gamma(a_1, b_1) and Gamma(a_2, b_2) distributions, respectively, which are independent of the other parameters. For more detail on the priors, refer to Lewis et. al (2018) [1].

The posterior density can be written as

$$\pi(\eta, \lambda, \kappa_1, \kappa_2) \propto L(X|\theta, \eta, \kappa_1, \kappa_2) \times \pi(\lambda) \times \pi(\eta, \tau) \times \pi(\kappa_1, \kappa_2),$$

where λ, η, and κ_1, κ_2 are assumed independent. A Markov chain Monte Carlo (MCMC) scheme is used to simulate candidate peptides from the posterior distribution since the posterior density is only known up to a constant and its actual form is complicated. Of all the peptides visited by the search algorithm, the one with the largest posterior probability is chosen as our best estimate for the true peptide.

4.1 Markov chain Monte Carlo algorithm

A starting point for the MCMC algorithm can be found by randomly adding or removing amino acids until a candidate peptide is found with mass within a tolerance of 0.5 Da of the mass of the true peptide, which is known to us from the mass spectrometry data. Requiring the mass to be within this tolerance drastically reduces the parameter space. However, the starting peptide may still be very different that the truth, particularly if the candidate chosen is long.

Once the starting peptide has been chosen, the log of the scoring function value for the candidate peptide is calculated. The current peptide is denoted η_{curr}. The β and γ vectors are pre-determined at the beginning of the algorithm and remain constant throughout, and a vector λ_{curr} is generated using these.

The steps for the MCMC algorithm are as follows.

1. A new peptide is created by randomly replacing one, two, or three amino acids of the current peptide with one, two, or three amino acids. Thus, the next candidate peptide may have length 1 less, 1 more, 2 less, 2 more, or equal to that of the current peptide, but still with total mass within 0.5 Da of the true peptide.

2. Generate a vector λ_{new} using the β and γ vectors.

3. Generate κ_1 and κ_2 from their full conditional distribution: gamma distributions with the shape parameter $\alpha_1 = a_1 + s$ and scale parameter $\beta_1 = S_1 + b_1$ and shape parameter $\alpha_2 = a_2 + (t - s)$ and scale parameter $\beta_2 = S_2 + b_2$, respectively. Note that the values of S_1 and S_2 are based on the current peptide.

4. Compute the unnormalized posterior probability for both the new and current peptide, computed based on the new and current λ vectors, respectively. Denote these as ζ_1 and ζ_2, respectively.

5. Generate $U \sim U(0, 1)$. If

$$U < \left(\frac{\zeta_1}{\zeta_2} \times \frac{q(\lambda_{curr}|\lambda_{new})}{q(\lambda_{new}|\lambda_{curr})} \times \frac{q(\eta_{curr}|\eta_{new})}{q(\eta_{new}|\eta_{curr})} \right),$$

   then the new peptide becomes the current peptide, and λ_{new} becomes λ_{curr}. Otherwise, the current peptide remains the same and λ_{curr} is unchanged.

6. Return to step 1.

Steps 1 through 6 are repeated for a large number of iterations. To ensure irreducibility, a new peptide independent of the current state will be generated every 500 steps. The peptide with the largest posterior density is chosen as the estimate of the true peptide.

5 PREPROCESSING METHODS

The raw data obtained from tandem mass spectrometry is not optimal to use because it may be influenced by issues such as noise (both chemical and electrical), instrument distortion, or m/z measurement errors. Several preprocessing techniques have been proposed for tandem mass spectrometry data, with the goal of reducing the noise without affecting the alignment in such a way that identification is hindered. These methods can also be applied to other types of mass spectrometry data.

The first preprocessing technique we investigated was removing the baseline from the data. It is typical of MS/MS data for the intensity values at the lower end of the spectrum to be amplified as a result of chemical noise from the ion matrix [17]. We used the R package msProcess, which has a function called msDetrend that can compute and remove the baseline from mass spectrometry data. In order to subtract the baseline from the spectra, it must first be estimated by fitting a curve locally to the intensity data using polynomial regression [18]. The function uses locally estimated scatterplot smoothing (LOESS) to estimate this baseline, which performs the following steps:

1. The spectrum is first divided into small segments, and in each of these segments, a quantile is computed. We use the default value of 0.1.

2. A predictor is selected for each of the segments based on the following criteria:

   - If the intensity value for some point is less than the quantile for that segment, then the intensity for the corresponding point of the predictor is simply the intensity of the point.

- If the intensity value for a point is greater than or equal to the quantile for that segment, then the intensity of the corresponding point of the predictor is equal to the quantile.

3. Once the predictor for each segment has been obtained, the baseline is calculated by applying polynomial regression to the predictors.

We then subtract the intensities of our fitted curve from the corresponding intensities of our observed spectrum, removing some of the non-signal chemical noise present in mass spectrometry data and resulting in a cleaner spectrum [19].

Another technique that can be applied to mass spectrometry data is wavelet shrinkage denoising. Wavelets can be applied to data to preserve true signals while removing noise. There are many types of wavelet transforms that can be thought of as filters that can be applied to the data [20]. We again use the R package msProcess, using the function msDenoise which by default uses a discrete wavelet transform. This discrete wavelet transform decomposes our signal (the mass spectrometry intensity values) and removes noise by performing 3 main steps. First, the forward discrete wavelet transform that decomposes our signal is calculated; this transform can be written as

$$c(j,k) = \sum_t f(t)\psi^*_{j,k}(t)$$

where $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, the $c(j,k)$ are wavelet coefficients, $f(t)$ is an intensity value, $j \in \mathbb{N}$ is the scale step, and $k \in \mathbb{N}$ is the shift step [21].

The coefficients of this transform that are large in magnitude correspond to signal, while those that are small are most likely noise. These coefficients are then shrunk towards 0, which results in the removal of low-amplitude noise. Lastly, an inverse wavelet transform reproduces our desired signal without the noise. The inverse discrete wavelet transform is

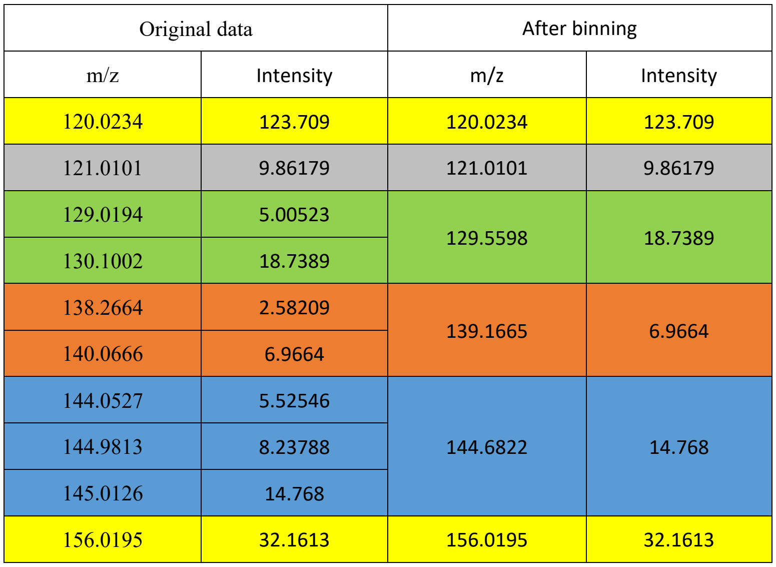
$$f(t) = \sum_{k} \sum_{j} c(j,k)\psi_{j,k}(t).$$

Binning is another commonly used preprocessing technique that can reduce the number of observations in a data set, making the data easier to analyze and possibly reducing the amount of noise [23]. Binning refers to the creation of bins for the data set which will contain values falling within the interval corresponding to that bin. These intervals are based on a window value; if this window is too large, the reduced data may no longer contain some signal peaks, while if it is too small, the intended effect of reducing the data set will not be as great.

Binning as applied to mass spectrometry involves first grouping adjacent m/z values into bins, and then choosing an intensity and m/z value to represent that bin. For our binning approach, the window width was calculated based on the average distances between m/z values, rounded up to the nearest 0.5 Da. Once the bins had been determined, the next step was to determine what portion of our data would be binned. Since peaks with large intensities are most likely true signal peaks, we separated these from the rest of the data to ensure they remained untouched. This separation was based on a percentile of the observed intensities for the given peptide; observations with intensities above this percentile were not binned, while those below it were. The percentile that we chose to use was 0.6 for short peptides and 0.1 for long peptides, based on previous binning work by Offei (2017) [22]. We define short peptides as those with total weights less than 1100 Da, and long peptides as those with weights greater than 1100 Da.

A single bin will consist of N m/z ratios and their corresponding intensity values and has the structure $[(m/z_1, I_1), (m/z_2, I_2), (m/z_3, I_3), ..., (m/z_N, I_N)]$, where each pair represents a peak within that bin. From this bin of N pairs, a single $(m/z, I)$ will be calculated to represent the bin. Various approaches can be used to calculate this pair: for the intensity value, an aggregate function such as the sum or the maximum of the N original intensities can be used, while the m/z may be determined as the mean or median of the original values or the value associated with the largest intensity. We chose to use the mean m/z value and the maximum intensity value as the representative pair for each bin. After binning was complete, we combined the binned observations with those that we had earlier separated. Table 5 gives an example of our binning method using a portion of the data from peptide FNDAVIR.

Table 5: Example of binning, using a window width of 2.5 and a 60th percentile of 21.50788



Original data		After binning	
m/z	Intensity	m/z	Intensity
120.0234	123.709	120.0234	123.709
121.0101	9.86179	121.0101	9.86179
129.0194	5.00523	129.5598	18.7389
130.1002	18.7389		
138.2664	2.58209	139.1665	6.9664
140.0666	6.9664		
144.0527	5.52546	144.6822	14.768
144.9813	8.23788		
145.0126	14.768		
156.0195	32.1613	156.0195	32.1613

Lewis et. al's method finds both a constant and moving threshold, from which a weighted average is calculated and used. This threshold is denoted by $T = (T_1, T_2, ..., T_{q^*})$, where q^* is equal to the number of m/z values present. For the constant threshold, the 75^{th} percentile of the observed intensity values is computed so that only the highest intensities are retained and becomes a component of t, a constant vector of the 75^{th} percentile with length q^*. However, since the mass spectrometer does not always capture the peaks at the beginning and end of the spectrum, using a constant threshold alone could remove true signal peaks misidentified as noise. Thus, a moving threshold is calculated as well, where for any fixed m/z value x^*, a subsection of the

m/z values is selected using a window width of 50 Da around x^*. Within this window, the 75$^{\text{th}}$ percentile of the observed intensity values is found. Each of these percentiles becomes a component of $t' = (t'_1, t'_2, ..., t'_{q^*})$. A weighted average of the constant and moving thresholds is then found using a sequence of weights, and when applied to the spectrum, observed m/z values with intensities above the corresponding threshold value in T are retained, and those below are removed.

6 RESULTS

The goal of this thesis is to compare these denoising methods to the method employed by Lewis et. al. To do so, we look at each of the denoising methods separately, as well as in various combinations. To gauge the effectiveness of each technique, we calculate the measures of closeness S_1 and S_2 (both with units of Da.) proposed by Lewis et. al.

For various peptides, we do the following:

1. Calculate the total weight.

2. Find the locations of the b and y ions and calculate the true λ vector based on these locations, where $\lambda_i^b = 1$ denotes the presence and $\lambda_i^b = 0$ denotes the absence of a b ion at position i; this notation similarly applies to the y ions.

3. For each of the following methods, find the true λ vector and record the dimension of the data set and the S_1 and S_2 values.

    - No denoising
    - Lewis et.al's method
    - Baseline removal
    - Wavelet shrinkage denoising
    - Binning
    - Baseline removal + binning
    - Wavelet shrinkage + binning

- Baseline removal + wavelet shrinkage + binning
- Baseline removal + Lewis et. al's method
- Wavelet shrinkage + Lewis et. al's method
- Binning + Lewis et. al's method
- Baseline removal + binning + Lewis et. al's method
- Baseline removal + wavelet shrinkage + binning + Lewis et. al's method

4. Compare the observed spectrum before denoising, after Lewis's approach, and after the combination with the lowest S_1 and S_2 values.

6.1 Short peptides

We begin by examining peptides that have a total weight less than 1100 Da, which we classify as short based on the work of Offei (2017) [22].

6.1.1 Example 1

Consider the peptide FNDAVIR, which has a total weight of 834.4476 Da and consists of 316 pairs of m/z and intensity values. Table 6 shows the results for each of the denoising methods. Here, we can see that none of the individual methods reduces the S_1 and S_2 values as much as Lewis et. al's method. However, some combinations of the standard denoising methods and Lewis resulted in lower distance values and smaller sizes, with Baseline removal + Lewis having the lowest S_1 and S_2 values and a smaller dimension of 58 pairs.

Table 6: Results for peptide FNDAVIR

Method	S_1	S_2	Dimension
None	0.1823758	2.847290	314
Lewis et. al's threshold	0.1092727	2.591605	75
Baseline removal	0.1823758	2.823239	259
Wavelet	0.1823758	2.841320	296
Binning	0.1823758	2.828167	272
Baseline removal + binning	0.1823758	2.804278	226
Wavelet + binning	0.1823758	2.817053	247
Baseline removal + wavelet + binning	0.1092727	2.780744	197
Baseline removal + Lewis	0.1092727	2.487374	58
Wavelet + Lewis	0.1162830	2.477295	63
Binning + Lewis	0.1092727	2.498054	59
Baseline removal + binning + Lewis	0.1092727	2.511678	50
Wavelet + binning + Lewis	0.1162830	2.462221	52
Baseline removal + wavelet + binning + Lewis	0.1162830	2.466346	46

In Figure 4, we compare the spectra before denoising is applied and after each of our combinations. We can see that the denoising method of Baseline removal + Lewis, which resulted in lower distance values, results in a cleaner spectrum without the loss of true signal peaks. Furthermore, we can see that some methods, such as Baseline removal + wavelet + binning + Lewis, remove not only noise but some of the true signal peaks as well.

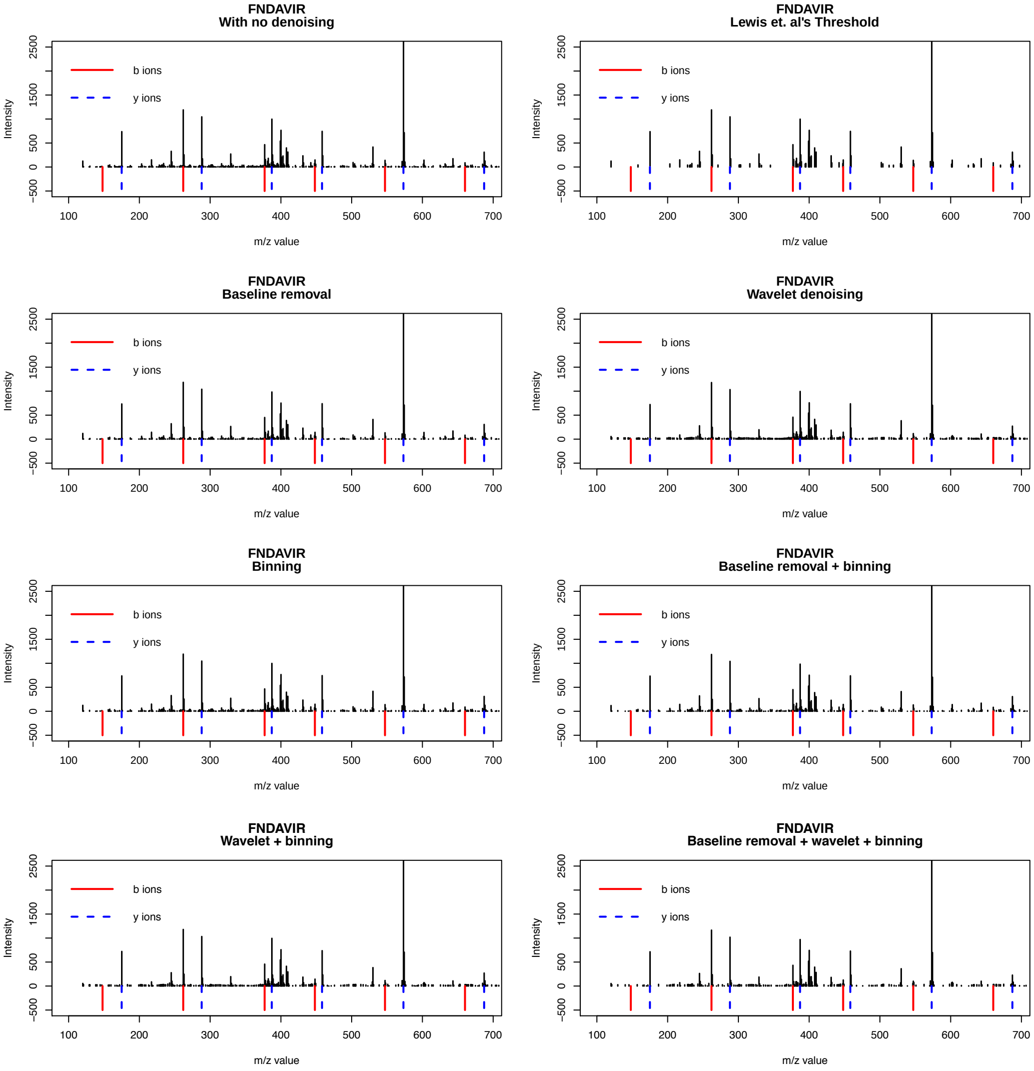


Figure 4: Observed spectra for peptide FNDAVIR (continued on next page)

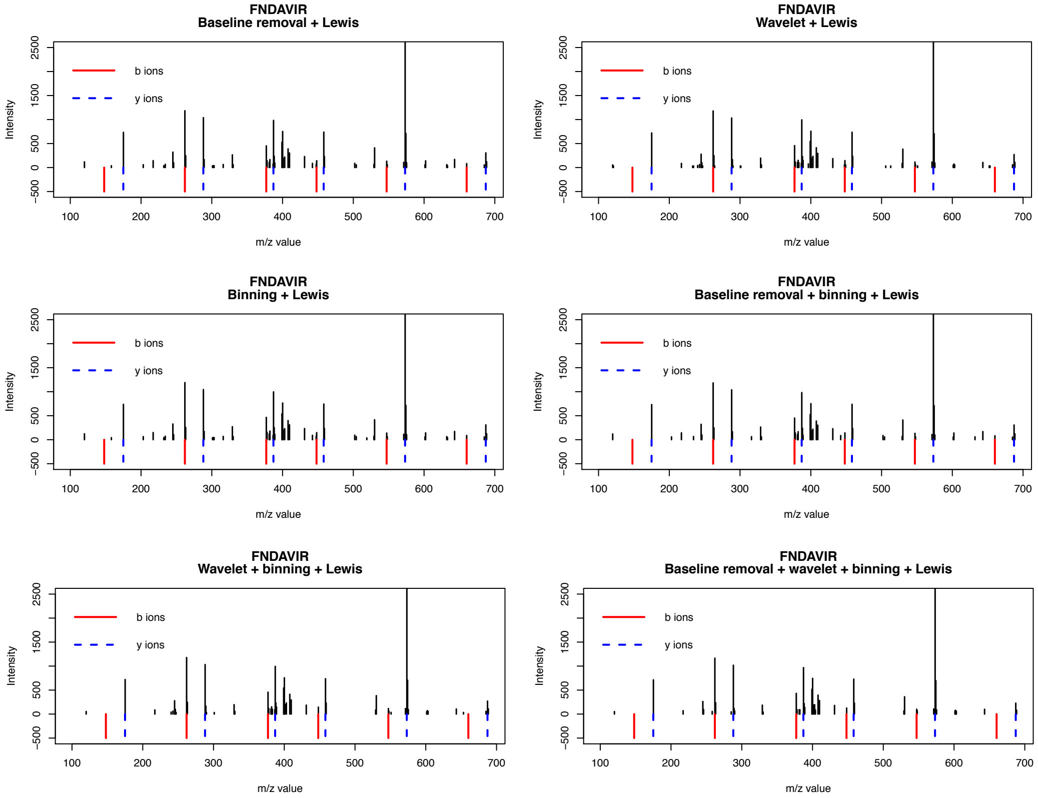


Figure 4: Observed spectra for peptide FNDAVIR (continued)

Table 7 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 7: Comparison of the distances before and after denoising for the peptide FNDAVIR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline removal + Lewis	Difference after baseline removal + Lewis
147.918	146.9315	0.98651	158.0906	10.17261	158.0906	10.17261
174.951	175.1306	0.17958	175.1306	0.17958	175.1306	0.17958
261.961	261.991	0.03	261.991	0.03	261.991	0.03
288.035	288.1577	0.12271	288.1577	0.12271	288.1577	0.12271
376.988	377.0584	0.07041	377.0584	0.07041	377.0584	0.07041
387.1034	387.2724	0.16897	387.2724	0.16897	387.2724	0.16897
448.0251	448.2078	0.18272	448.2078	0.18272	448.2078	0.18272
458.1405	458.2951	0.1546	458.2951	0.1546	458.2951	0.1546
547.0935	547.1872	0.09369	547.1872	0.09369	547.1872	0.09369
573.1675	573.2686	0.10112	573.2686	0.10112	573.2686	0.10112
660.1775	660.2167	0.03917	660.2167	0.03917	660.2167	0.03917
687.2105	687.2695	0.05903	687.2695	0.05903	687.2695	0.05903

6.1.2 Example 2

Next, we look at the peptide LLDNLLTK. This peptide has a total weight of 929.5662 Da and consists of 262 pairs of m/z and intensity values. Table 8 shows us that, while none of the individual methods gave lower distance values, several of the combinations with the Lewis et. al method did. In this example, we see that the usage of all three denoising techniques with the Lewis threshold resulted in the lowest S_1 and S_2 values and has the least number of observations.

Table 8: Results for peptide LLDNLLTK

Method	S_1	S_2	Dimension
None	0.1447354	2.814817	262
Lewis et. al's threshold	0.1447354	2.644831	66
Baseline removal	0.1447354	2.808918	217
Wavelet	0.1447354	2.816024	256
Binning	0.1447354	2.799330	225
Baseline removal + binning	0.1447354	2.802123	192
Wavelet + binning	0.2399673	2.812353	207
Baseline removal + wavelet + binning	0.1894565	2.782881	164
Baseline removal + Lewis	0.1447354	2.578049	55
Wavelet + Lewis	0.1435090	2.588175	56
Binning + Lewis	0.1447354	2.606179	58
Baseline removal + binning + Lewis	0.1522500	2.638443	50
Wavelet + binning + Lewis	0.1435090	2.501406	46
Baseline removal + wavelet + binning + Lewis	0.1435090	2.448134	41

The observed spectra for some of the denoising combinations applied to LLDNLLTK are shown in Figure 5. We can see that before any denoising, the first b ion is missing, which is unsurprising. Comparing the results from Lewis and the other methods, we see that in some cases, such as when using all three methods and Lewis et.al's together, too much is removed and many of the b and y ions are missing. Baseline removal + Lewis, however, does not appear to be missing any more of the b and y ions than Lewis et. al's method, yet it resulted in a lower S_2 values and a smaller dimension.

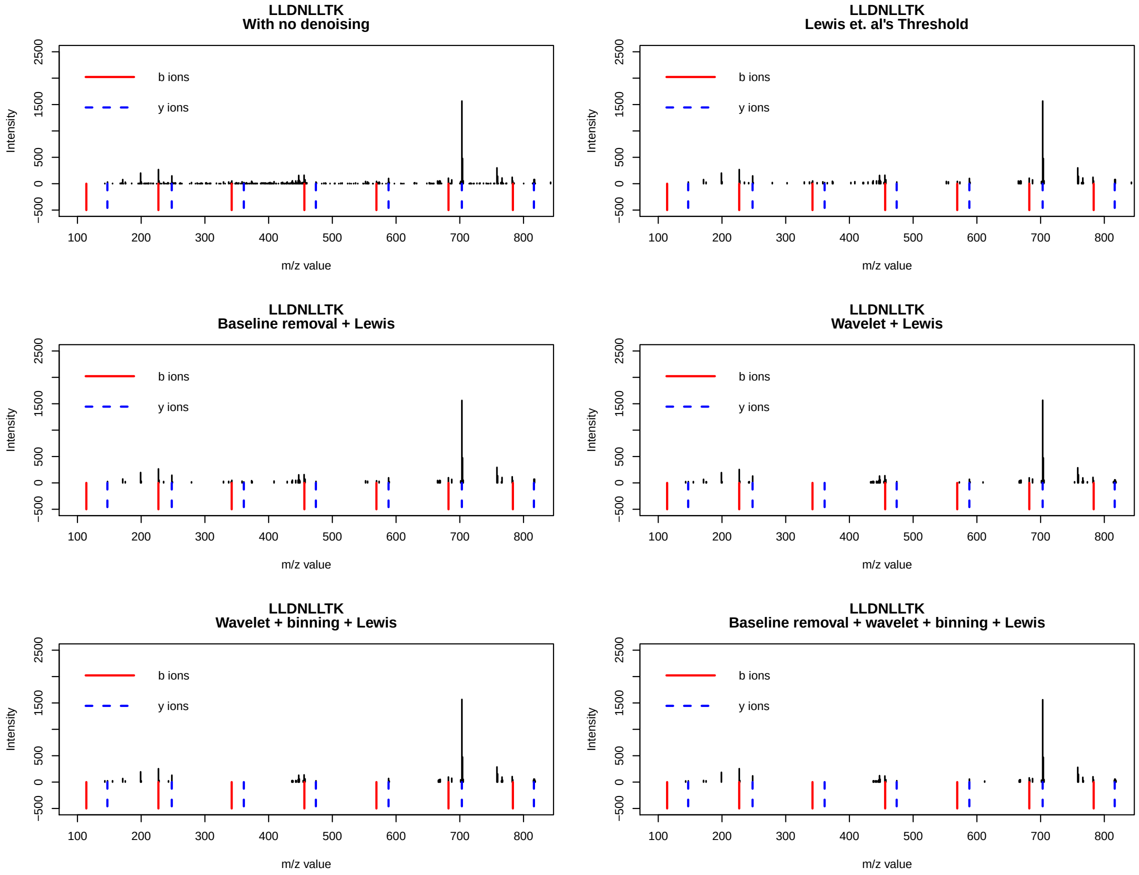


Figure 5: Observed spectra for peptide LLDNLLTK

Table 9 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 9: Comparison of the distances before and after denoising for the peptide LLDNLLTK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline removal + Lewis	Difference after baseline removal + Lewis
113.934	142.9998	29.06576	147.1624	33.22835	147.1624	33.22835
146.945	147.1624	0.21735	147.1624	0.21735	147.1624	0.21735
227.018	227.1014	0.08341	227.1014	0.08341	227.1014	0.08341
247.993	248.159	0.166	248.159	0.166	248.159	0.166
342.045	342.2608	0.21583	342.2608	0.21583	342.2608	0.21583
361.077	361.1316	0.05456	361.1316	0.05456	361.1316	0.05456
456.088	455.6391	0.44887	455.6391	0.44887	455.6391	0.44887
474.161	474.2799	0.11888	474.2799	0.11888	474.2799	0.11888
569.172	569.3481	0.17608	569.3481	0.17608	569.3481	0.17608
588.204	588.2653	0.06126	588.2653	0.06126	588.2653	0.06126
682.256	682.1585	0.09755	682.1585	0.09755	682.1585	0.09755
703.231	703.2893	0.05825	703.2893	0.05825	703.2893	0.05825
783.304	783.1446	0.15935	783.1446	0.15935	783.1446	0.15935
816.315	816.3392	0.02417	816.3392	0.02417	816.3392	0.02417

6.1.3 Example 3

Peptide TGMSNVSK is our next example, with a total weight of 823.3982 Da and 258 pairs of m/z and intensity values. Once again, we see in Table 10 that the only methods that result in smaller distance values than Lewis's threshold are those that combine it with standard denoising techniques. Baseline removal + Lewis or Binning + Lewis threshold produce the best results here, with the same S_1, a slightly smaller S_2, and a smaller dimension compared to Lewis et. al's threshold alone.

Table 10: Results for peptide TGMSNVSK

Method	S_1	S_2	Dimension
None	0.08102615	2.808701	258
Lewis et. al's threshold	0.08102615	2.583948	60
Baseline removal	0.08102615	2.798126	206
Wavelet	0.08102615	2.815459	243
Binning	0.08102615	2.781371	225
Baseline removal + binning	0.08102615	2.762959	173
Wavelet + binning	0.08102615	2.778984	193
Baseline removal + wavelet + binning	0.10714769	2.796000	154
Baseline removal + Lewis	0.08102615	2.564812	53
Wavelet + Lewis	0.09006545	2.600395	59
Binning + Lewis	0.08102615	2.564812	53
Baseline removal + binning + Lewis	0.08739083	2.561415	43
Wavelet + binning + Lewis	0.09006545	2.550767	50
Baseline removal + wavelet + binning + Lewis	0.09006545	2.562090	39

In the comparison of the observed spectra in Figure 6, we can clearly see that all of the b and y ions are still present in the spectra where Baseline + Lewis and Binning + Lewis have been applied, but each has slightly less noise compared to Lewis et. al alone. Additionally, we can see that some of the others methods result in the removal of true signal peaks.

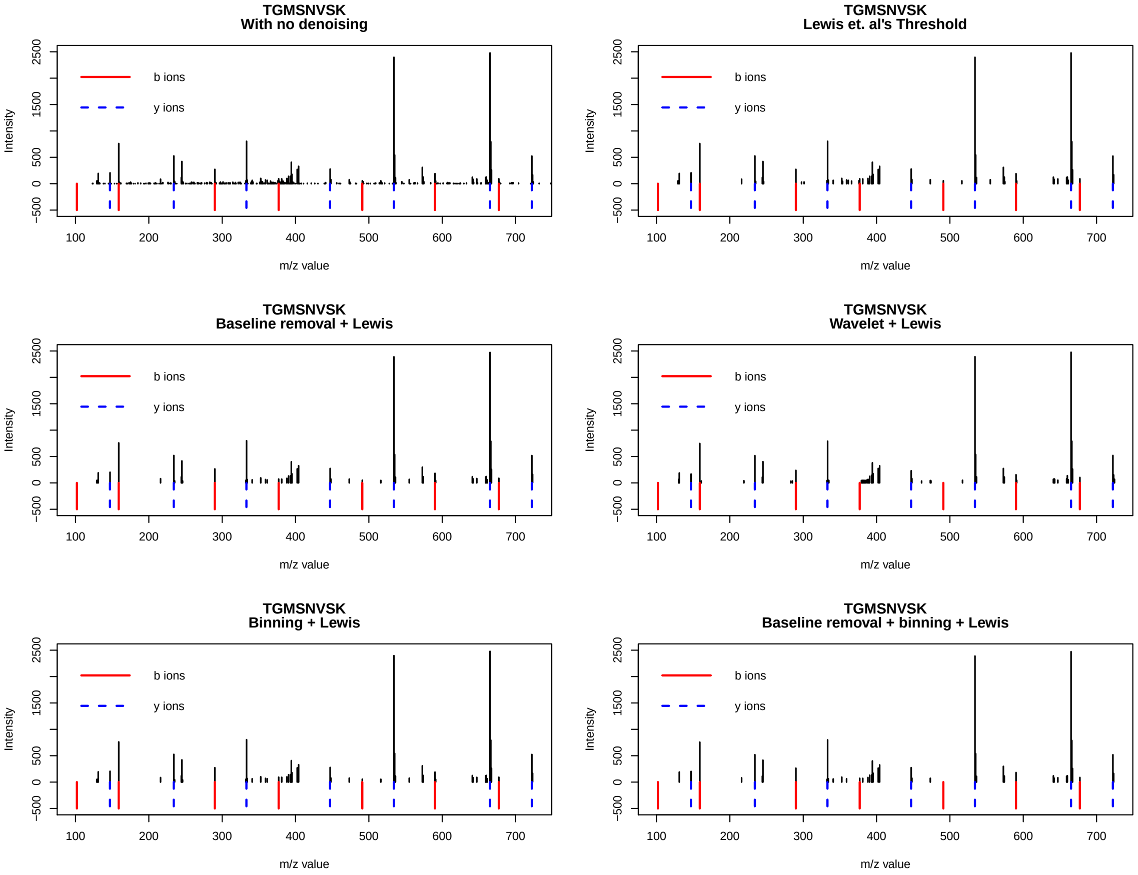


Figure 6: Observed spectra for peptide TGMSNVSK

Table 11 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 11: Comparison of the distances before and after denoising for the peptide TGMSNVSK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline removal + Lewis	Difference after baseline removal + Lewis
101.898	123.0595	21.16152	129.0267	27.1287	129.0267	27.1287
146.945	147.0966	0.15163	147.0966	0.15163	147.0966	0.15163
158.9195	158.9794	0.0599	158.9794	0.0599	158.9794	0.0599
233.977	234.0808	0.10381	234.0808	0.10381	234.0808	0.10381
289.9595	290.0093	0.04984	290.0093	0.04984	290.0093	0.04984
333.0454	333.2553	0.20991	333.2553	0.20991	333.2553	0.20991
376.9915	377.0495	0.05797	377.0495	0.05797	377.0495	0.05797
447.0884	447.2143	0.12589	447.2143	0.12589	447.2143	0.12589
491.0345	491.0392	0.00465	491.0392	0.00465	491.0392	0.00465
534.1204	534.1837	0.06332	534.1837	0.06332	534.1837	0.06332
590.1029	590.1507	0.0478	590.1507	0.0478	590.1507	0.0478
665.1604	665.1866	0.02618	665.1866	0.02618	665.1866	0.02618
677.1349	677.2228	0.08788	677.2228	0.08788	677.2228	0.08788
722.1819	722.2465	0.06456	722.2465	0.06456	722.2465	0.06456

6.1.4 Example 4

The peptide PFVDGGVIK has a total weight of 931.5247 Da and has 272 pairs of m/z and intensity values. Our various methods provide mixed results with this peptide; while some of the techniques reduce the S_1 and S_2 values, they also result in the incorrect classification of some of the b and y ions as noise, removing them from the spectra, as can be seen in Table 12.

Table 12: Results for peptide PFVDGGVIK

Method	S_1	S_2	Dimension
None	0.1595407	2.864543	272
Lewis et. al's threshold	0.2653985	2.569789	64
Baseline removal	0.1595407	2.840482	224
Wavelet	0.1595407	2.859647	263
Binning	0.1595407	2.850775	234
Baseline removal + binning	0.1595407	2.817913	194
Wavelet + binning	0.2290232	2.843611	213
Baseline removal + wavelet + binning	0.2715236	2.814970	166
Baseline removal + Lewis	0.2653985	2.524302	59
Wavelet + Lewis	0.0772050	2.536550	61
Binning + Lewis	0.2653985	2.523027	59
Baseline removal + binning + Lewis	0.2653985	2.435165	48
Wavelet + binning + Lewis	0.0772050	2.477601	52
Baseline removal + wavelet + binning + Lewis	0.0772050	2.272788	41

In the spectra in Figure 7, we see that several of the b and y ions are missing or have very low intensities to begin with. Both Lewis's approach and our combined methods result in the removal of some of these. The combinations involving wavelet, however, removes a greater number of the true signal peaks, which would negatively impact the identification for this peptide.

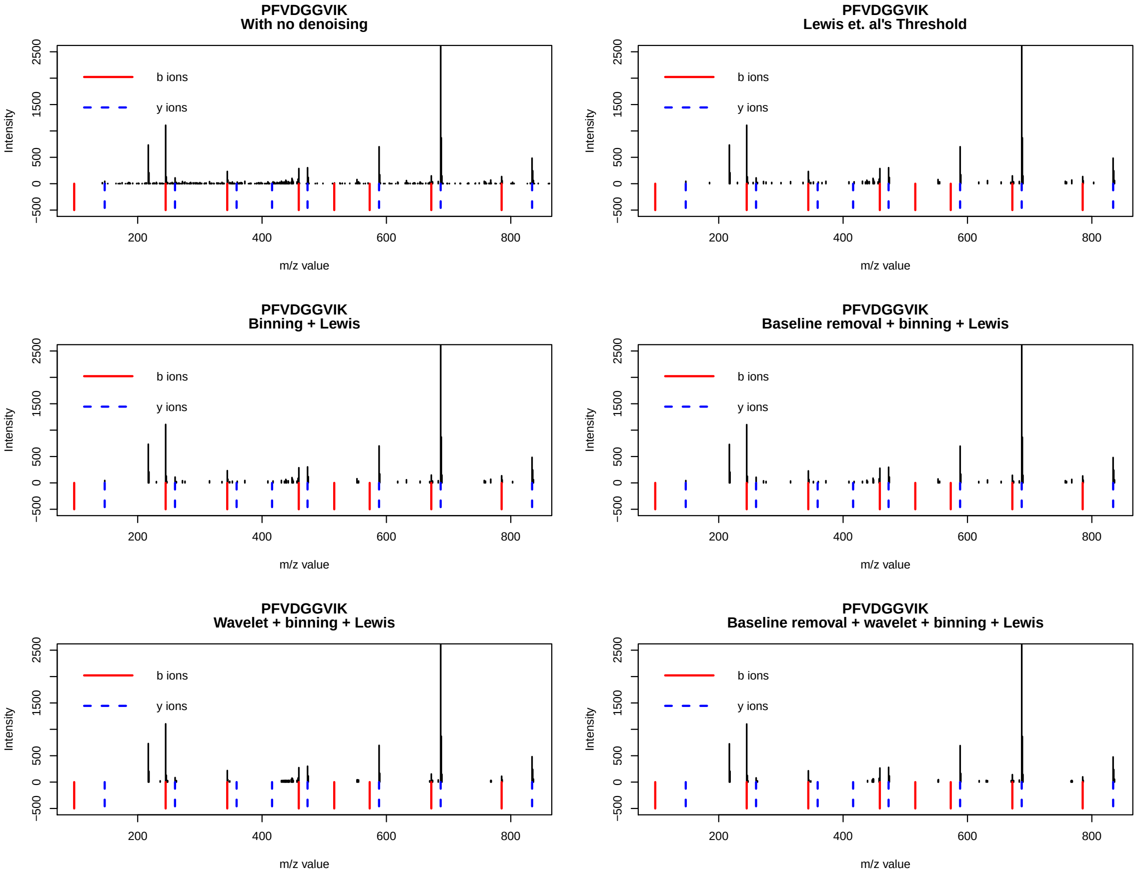


Figure 7: Observed spectra for peptide PFVDGGVIK

Table 13 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 13: Comparison of the distances before and after denoising for the peptide PFVDGGVIK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
97.9028	143.1296	45.22682	147.0574	49.15456	147.0574	49.15456
146.945	147.0574	0.11236	147.0574	0.11236	147.0574	0.11236
244.9708	245.0069	0.03607	245.0069	0.03607	245.0069	0.03607
260.029	260.1677	0.13869	260.1677	0.13869	260.1677	0.13869
344.0392	344.1313	0.09209	344.1313	0.09209	344.1313	0.09209
359.0974	360.0727	0.97529	360.0727	0.97529	360.0727	0.97529
416.1189	416.1399	0.02096	417.7094	1.59048	417.7094	1.59048
459.0662	459.1938	0.12762	459.1938	0.12762	459.1938	0.12762
473.1404	473.2045	0.06413	473.2045	0.06413	473.2045	0.06413
516.0877	516.4406	0.35291	552.664	36.5763	552.664	36.5763
573.1092	570.5151	2.59412	588.2405	15.13134	588.2405	15.13134
588.1674	588.2405	0.07314	588.2405	0.07314	588.2405	0.07314
672.1776	672.2109	0.03328	672.2109	0.03328	672.2109	0.03328
687.2358	687.277	0.04124	687.277	0.04124	687.277	0.04124
785.2616	785.3333	0.07165	785.3333	0.07165	785.3333	0.07165
834.3038	834.2097	0.09414	834.2097	0.09414	834.2097	0.09414

6.1.5 Example 5

Our next example is peptide LSDYGVQLR, which has a total weight of 1050.558 and has 469 pairs. We see in Table 14 that most of the techniques resulted in the same S_1 value, however, the combination of baseline removal, binning, and Lewis et. al's approach resulted in a lower S_2 and a smaller data set.

Table 14: Results for peptide LSDYGVQLR

Method	S_1	S_2	Dimension
None	0.1071133	2.870772	469
Lewis et. al's threshold	0.1071133	2.660098	112
Baseline removal	0.1071133	2.850258	389
Wavelet	0.1071133	2.869604	426
Binning	0.1071133	2.852467	404
Baseline removal + binning	0.1071133	2.832798	335
Wavelet + binning	0.1071133	2.847582	353
Baseline removal + wavelet + binning	0.1071133	2.829989	288
Baseline removal + Lewis	0.1071133	2.614872	96
Wavelet + Lewis	0.1071133	2.661710	103
Binning + Lewis	0.1071133	2.641433	102
Baseline removal + binning + Lewis	0.1071133	2.610940	84
Wavelet + binning + Lewis	0.1129107	2.613383	91
Baseline removal + wavelet + binning + Lewis	0.1106154	2.596628	73

Comparison of the spectra in Figure 8 shows that all of the b and y ions present in Lewis et al's threshold are still present in all of the other models except for the last two. Baseline + binning + Lewis, which resulted in the lowest S_1 and S_2 values, has considerably less noise than when applying Lewis et. al's method alone.

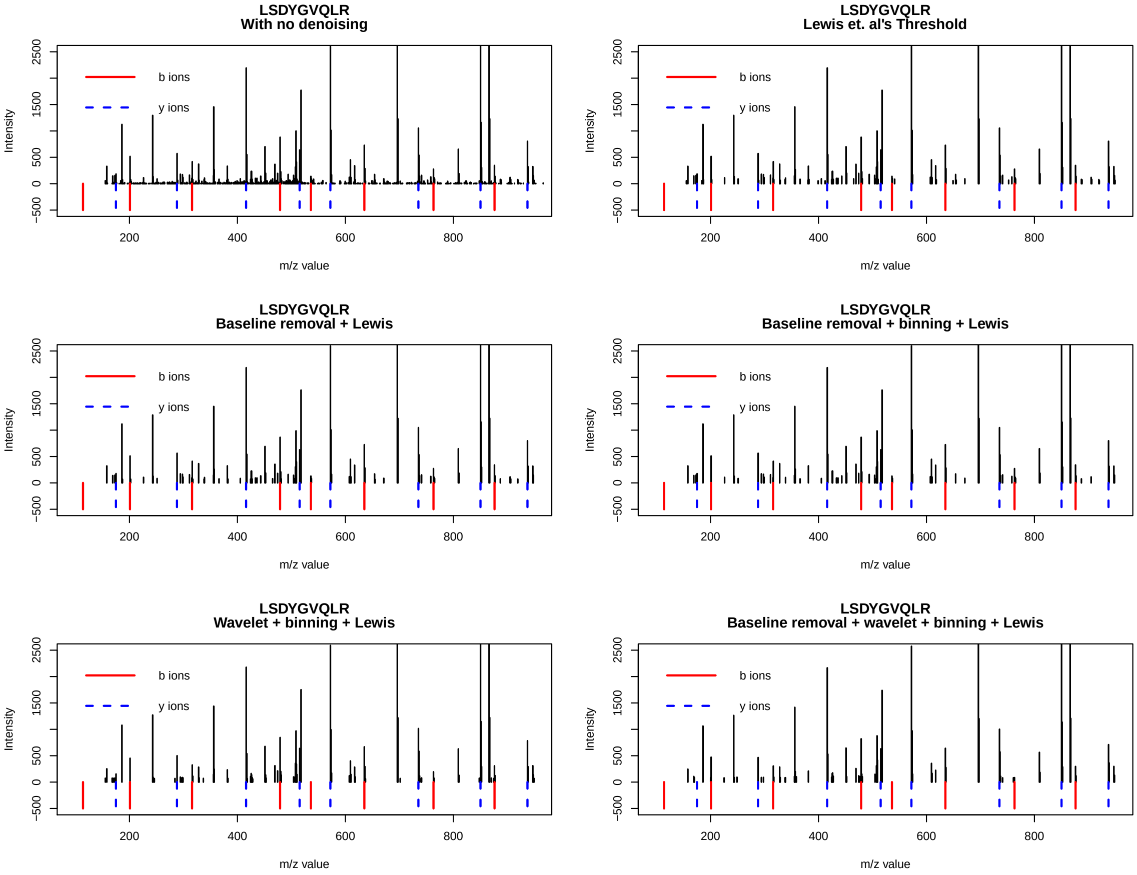


Figure 8: Observed spectra for peptide LSDYGVQLR

Table 15 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that only one of the distances increased compared to Lewis et. al's method; values in bold indicate an increase.

Table 15: Comparison of the distances before and after denoising for the peptide LSDYGVQLR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
113.934	155.255	41.32097	155.255	41.32097	158.0459	**44.11185**
174.951	175.0938	0.14275	175.0938	0.14275	175.0938	0.14275
200.966	201.0907	0.1247	201.0907	0.1247	201.0907	0.1247
288.035	288.1805	0.14554	288.1805	0.14554	288.1805	0.14554
315.993	316.2869	0.29387	316.2869	0.29387	316.2869	0.29387
416.094	416.2673	0.17333	416.2673	0.17333	416.2673	0.17333
479.056	479.0506	0.0054	479.0506	0.0054	479.0506	0.0054
515.1624	515.353	0.19063	515.353	0.19063	515.353	0.19063
536.0775	536.1034	0.02595	536.1034	0.02595	536.1034	0.02595
572.1839	572.2805	0.09656	572.2805	0.09656	572.2805	0.09656
635.1459	635.0522	0.09371	635.0522	0.09371	635.0522	0.09371
735.2469	735.3122	0.06529	735.3122	0.06529	735.3122	0.06529
763.2049	763.1658	0.03907	763.1658	0.03907	763.1658	0.03907
850.2739	850.3229	0.04904	850.3229	0.04904	850.3229	0.04904
876.2889	876.1458	0.14309	876.1458	0.14309	876.1458	0.14309
937.3059	937.3237	0.01777	937.3237	0.01777	937.3237	0.01777

6.1.6 Example 6

Next, we look at the peptide SILSELVR, which has a total weight of 916.5479 Da and contains 212 pairs of m/z values and intensities. In Table 16, we see that three of our combinations resulted in the same or lower values for S_1 and S_2 compared to Lewis et. al's method.

Table 16: Results for peptide SILSELVR

Method	S_1	S_2	Dimension
None	0.1546125	2.769107	212
Lewis et. al's threshold	0.1550491	2.403455	50
Baseline removal	0.2776392	2.743566	168
Wavelet	0.1546125	2.771986	205
Binning	0.2273754	2.740031	180
Baseline removal + binning	0.2776392	2.711459	144
Wavelet + binning	0.2233842	2.745436	165
Baseline removal + wavelet + binning	0.4219806	2.777968	117
Baseline removal + Lewis	0.1550491	2.351828	43
Wavelet + Lewis	0.1747111	2.519895	49
Binning + Lewis	0.1550491	2.351828	43
Baseline removal + binning + Lewis	0.1598420	2.281463	36
Wavelet + binning + Lewis	0.1747111	2.512967	40
Baseline removal + wavelet + binning + Lewis	0.1347457	2.378247	28

Looking at the spectra in Figure 9, we see that, as in other examples, combining all three denoising methods with Lewis et. al's results in the removal of true signal peaks, while using Baseline + Lewis or Binning + Lewis does not.

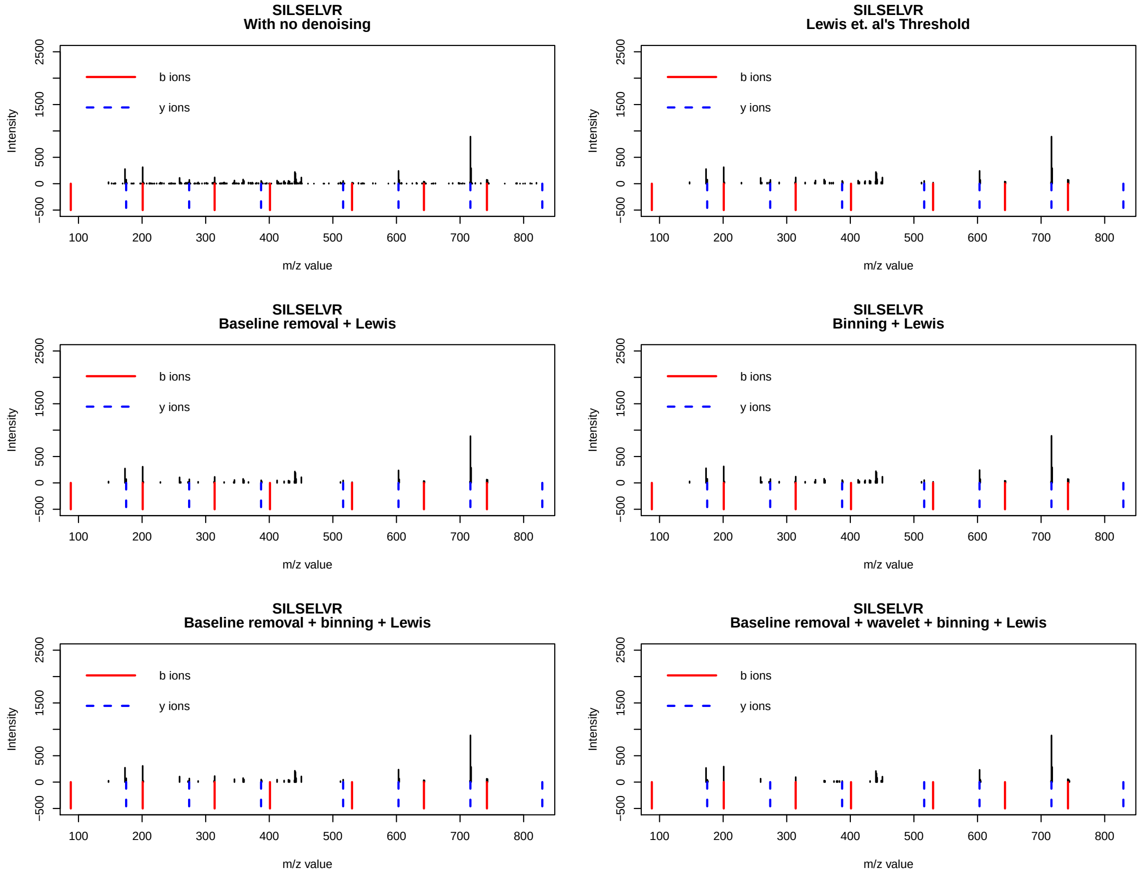


Figure 9: Observed spectra for peptide SILSELVR

Table 17 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 17: Comparison of the distances before and after denoising for the peptide SILSELVR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
87.882	147.2322	59.35024	147.2322	59.35024	147.2322	59.35024
174.951	175.1029	0.15191	175.1029	0.15191	175.1029	0.15191
200.966	200.9217	0.04431	200.9217	0.04431	200.9217	0.04431
274.0194	274.2922	0.27278	274.2922	0.27278	274.2922	0.27278
314.05	314.1382	0.08818	314.1382	0.08818	314.1382	0.08818
387.1034	387.2689	0.16549	387.2689	0.16549	387.2689	0.16549
401.082	400.9322	0.14981	412.2889	11.20691	412.2889	11.20691
516.1464	516.1724	0.02602	516.1724	0.02602	516.1724	0.02602
530.125	530.2321	0.10712	530.2321	0.10712	530.2321	0.10712
603.1784	603.2726	0.09418	603.2726	0.09418	603.2726	0.09418
643.209	642.8526	0.3564	642.8526	0.3564	642.8526	0.3564
716.2624	716.3303	0.06792	716.3303	0.06792	716.3303	0.06792
742.2774	741.9462	0.33123	741.9462	0.33123	741.9462	0.33123
829.3464	820.1333	9.2131	743.2372	86.10916	743.2372	86.10916

6.1.7 Example 7

Peptide QVMELLQ has a total weight of 840.4366 and has 325 pairs of m/z values and intensities. Our results in Table 18 show that none of the methods performed better than Lewis et. al's in terms of lowering both the S_1 and S_2 values, however, the combination of Wavelet + binning + Lewis does lower S_2 by 0.111267 at the cost of increasing S_1 by 0.0058834. This method also reduces the dimension of the data set from 80 pairs to 63.

Table 18: Results for peptide QVMELLQ

Method	S_1	S_2	Dimension
None	0.1153408	2.851626	325
Lewis et. al's threshold	0.1004036	2.723167	80
Baseline removal	0.2233075	2.828620	253
Wavelet	0.1153408	2.844425	310
Binning	0.1693242	2.843022	274
Baseline removal + binning	0.2233075	2.816608	216
Wavelet + binning	0.1693242	2.820343	252
Baseline removal + wavelet + binning	0.2233075	2.798706	195
Baseline removal + Lewis	0.1882627	2.737825	63
Wavelet + Lewis	0.1062870	2.672906	73
Binning + Lewis	0.1004036	2.727763	68
Baseline removal + binning + Lewis	0.1882627	2.734340	55
Wavelet + binning + Lewis	0.1062870	2.611190	63
Baseline removal + wavelet + binning + Lewis	0.2029320	2.650117	47

The spectra seen in Figure 10 show that the method of Wavelet + binning + Lewis does indeed clean the spectrum more than Lewis et. al's threshold alone, however, it also removes one of the *b* ions with a very small intensity.

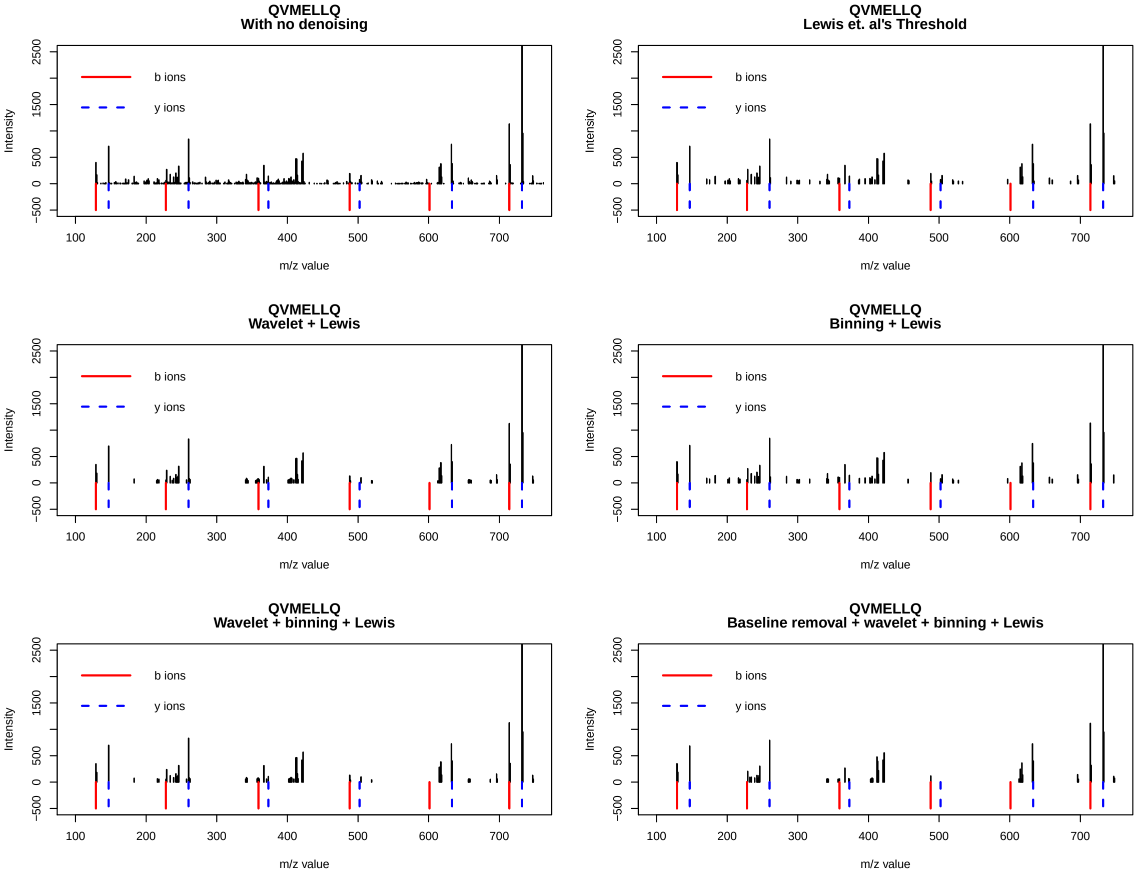


Figure 10: Observed spectra for peptide QVMELLQ

Table 19 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that two of the distances increased compared to Lewis et. al's method; values in bold indicate increases.

Table 19: Comparison of the distances before and after denoising for the peptide QVMELLQ

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after wavelet + binning + Lewis	Difference after wavelet + binning + Lewis
128.909	128.9447	0.03569	128.9447	0.03569	128.9447	0.03569
146.909	146.9926	0.08363	146.9926	0.08363	146.9926	0.08363
227.9774	228.1265	0.14911	228.1265	0.14911	228.1265	0.14911
259.993	260.1338	0.14082	260.1338	0.14082	260.1338	0.14082
359.0174	359.0739	0.05648	359.0739	0.05648	359.0739	0.05648
373.077	373.0233	0.05365	373.0233	0.05365	373.0233	0.05365
488.0604	488.2929	0.23248	488.2929	0.23248	488.2929	0.23248
502.12	502.0784	0.04157	502.0784	0.04157	504.2458	**2.12582**
601.1444	600.8647	0.27965	597.0975	4.04687	614.9736	**13.82917**
633.16	633.3168	0.15683	633.3168	0.15683	633.3168	0.15683
714.2284	714.108	0.12043	714.108	0.12043	714.108	0.12043
732.2284	732.2622	0.03375	732.2622	0.03375	732.2622	0.03375

6.1.8 Example 8

The next peptide we consider is peptide IGENINIR, which has a weight of 928.5213 Da and whose data set contains 279 pairs. In Table 20, we see that while Baseline + Lewis and Binning + Lewis provide results similar to that of Lewis et. al alone, methods involving Wavelet greatly reduce the S_1 and S_2 values.

Table 20: Results for peptide IGENINIR

Method	S_1	S_2	Dimension
None	0.1301923	2.834307	279
Lewis et. al's threshold	0.1848092	2.623748	65
Baseline removal	0.1301923	2.790122	223
Wavelet	0.1301923	2.840984	266
Binning	0.1301923	2.805840	240
Baseline removal + binning	0.1301923	2.771797	192
Wavelet + binning	0.1349454	2.814088	221
Baseline removal + wavelet + binning	0.1292751	2.802088	174
Baseline removal + Lewis	0.1848092	2.638960	54
Wavelet + Lewis	0.1125417	2.574574	55
Binning + Lewis	0.1848092	2.624860	58
Baseline removal + binning + Lewis	0.1848092	2.688988	46
Wavelet + binning + Lewis	0.1125417	2.554732	45
Baseline removal + wavelet + binning + Lewis	0.1125417	2.532187	40

The spectra in Figure 11 indicate that, while the Wavelet methods are decreasing the distance values, they are also removing one of the *b* ions with a very small intensity that was still present when using the other denoising methods.

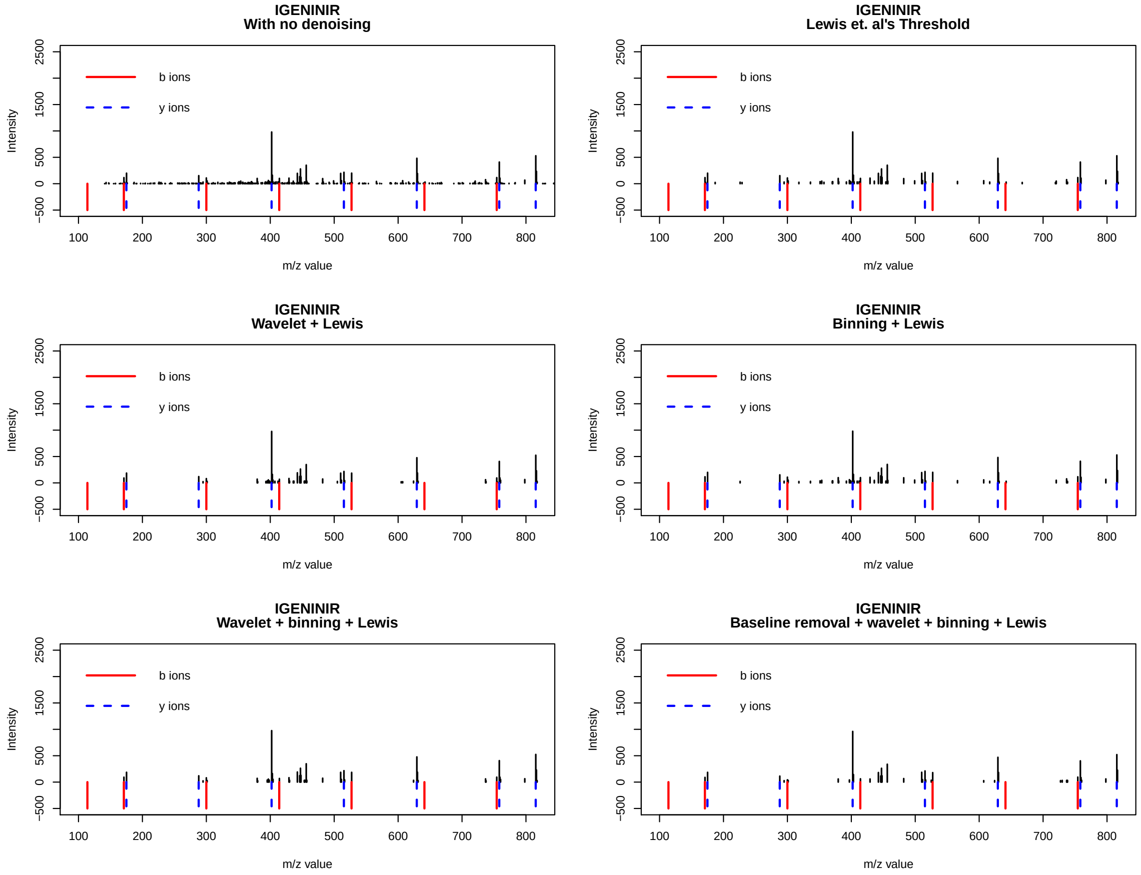


Figure 11: Observed spectra for peptide IGENINIR

Table 21 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that one of the distances increased compared to Lewis et. al's method; values in bold indicate increases.

Table 21: Comparison of the distances before and after denoising for the peptide IGENINIR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + wavelet + binning + Lewis	Difference after baseline + wavelet + binning + Lewis
113.934	140.9316	26.99759	171.0836	57.14959	171.0836	57.14959
170.9555	171.0836	0.12809	171.0836	0.12809	171.0836	0.12809
174.951	175.15	0.19896	175.15	0.19896	175.15	0.19896
288.035	288.1366	0.1016	288.1366	0.1016	288.1366	0.1016
299.9985	299.9937	0.00479	299.9937	0.00479	299.9937	0.00479
402.078	402.1953	0.11734	402.1953	0.11734	402.1953	0.11734
414.0415	414.2643	0.22275	414.2643	0.22275	414.2643	0.22275
515.162	515.2747	0.11272	515.2747	0.11272	515.2747	0.11272
527.1255	527.3021	0.17656	527.3021	0.17656	527.3021	0.17656
629.205	629.3256	0.12062	629.3256	0.12062	629.3256	0.12062
641.1685	641.5105	0.342	642.2205	1.05202	630.2726	**10.89592**
754.2525	754.2968	0.04431	754.2968	0.04431	754.2968	0.04431
758.248	758.2889	0.04094	758.2889	0.04094	758.2889	0.04094
815.2695	815.3513	0.08182	815.3513	0.08182	815.3513	0.08182

6.1.9 Example 9

The next peptide that we look at is peptide FLDQVNAK. This peptide has a true weight of 934.4994 Da and contains 346 pairs of m/z and intensity values. Table 22 gives the results for each of our denoising methods, and we can see that the different methods provide mixed results. Binning + Lewis and Baseline removal + Lewis provide similar, yet slightly better results than Lewis et. al alone. Once again, methods including the Wavelet denoising give lower S_1 values, with Wavelet + binning + Lewis having the lowest S_1 and S_2 overall.

Table 22: Results for peptide FLDQVNAK

Method	S_1	S_2	Dimension
None	0.1269757	2.860879	346
Lewis et. al's threshold	0.1638800	2.701270	80
Baseline removal	0.1269757	2.848925	281
Wavelet	0.1269757	2.857883	339
Binning	0.1269757	2.846172	307
Baseline removal + binning	0.1269757	2.831049	248
Wavelet + binning	0.1638800	2.831063	272
Baseline removal + wavelet + binning	0.1718121	2.849969	214
Baseline removal + Lewis	0.1638800	2.711272	69
Wavelet + Lewis	0.1193264	2.712125	78
Binning + Lewis	0.1638800	2.700029	73
Baseline removal + binning + Lewis	0.1638800	2.694255	60
Wavelet + binning + Lewis	0.1193264	2.661620	68
Baseline removal + wavelet + binning + Lewis	0.1064470	2.701399	52

The spectra in Figure 12 show that, even before denoising is applied, several of the b and y ions have very low intensities. The other plots indicate that some of these peaks were misidentified as noise and thus removed, with those methods involving wavelets removing the most.

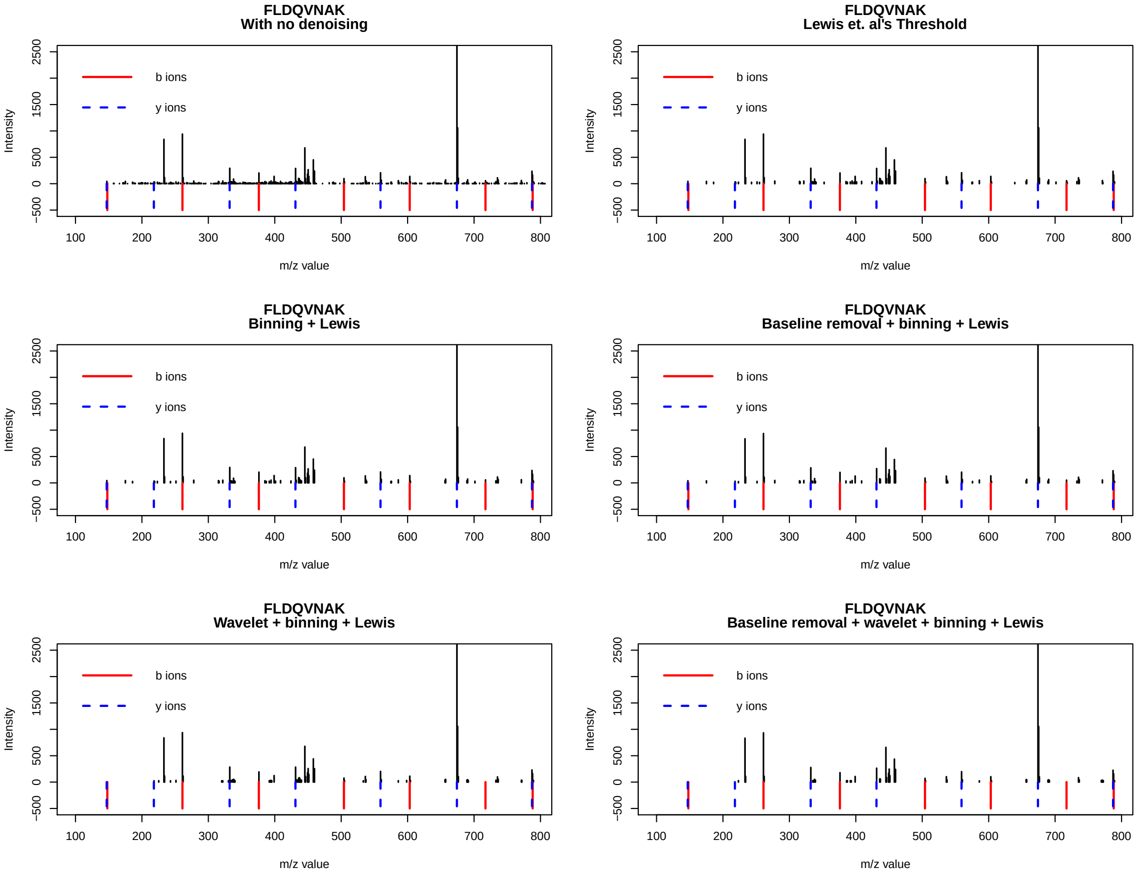


Figure 12: Observed spectra for peptide FLDQVNAK

Table 23 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 23: Comparison of the distances before and after denoising for the peptide FLDQVNAK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
146.945	147.1856	0.24055	147.1856	0.24055	147.1856	0.24055
147.918	148.1338	0.21579	147.1856	0.73245	147.1856	0.73245
217.9821	218.2302	0.24812	218.2302	0.24812	218.2302	0.24812
261.002	260.9587	0.04326	260.9587	0.04326	260.9587	0.04326
332.0251	332.2493	0.22417	332.2493	0.22417	332.2493	0.22417
376.029	376.1601	0.13106	376.1601	0.13106	376.1601	0.13106
431.0935	431.3058	0.21232	431.3058	0.21232	431.3058	0.21232
504.088	504.1824	0.0944	504.1824	0.0944	504.1824	0.0944
559.1525	559.202	0.04947	559.202	0.04947	559.202	0.04947
603.1564	603.2235	0.06711	603.2235	0.06711	603.2235	0.06711
674.1795	674.2315	0.05201	674.2315	0.05201	674.2315	0.05201
717.1994	717.2081	0.00873	717.2081	0.00873	717.2081	0.00873
787.2635	787.1916	0.07191	787.1916	0.07191	787.1916	0.07191
788.2365	788.1177	0.11876	788.1177	0.11876	788.1177	0.11876

6.1.10 Example 10

For our last short peptide example, we consider the peptide GYEFINDIK, which has a total weight of 1098.547 and consists of 456 pairs of m/z and intensity values. Table 24 shows that, while all of the methods give the same S_1 value, the combination of Baseline removal + wavelet + binning + Lewis gives a much lower S_2 compared to Lewis et. al alone, while also reducing the dimension from 103 to 60.

Table 24: Results for peptide GYEFINDIK

Method	S_1	S_2	Dimension
None	0.09810538	2.863754	456
Lewis et. al's threshold	0.09810538	2.599123	103
Baseline removal	0.09810538	2.846636	370
Wavelet	0.09810538	2.862328	442
Binning	0.09810538	2.848123	397
Baseline removal + binning	0.09810538	2.832943	319
Wavelet + binning	0.09810538	2.845493	365
Baseline removal + wavelet + binning	0.09810538	2.834742	285
Baseline removal + Lewis	0.09810538	2.652561	88
Wavelet + Lewis	0.09810538	2.643418	94
Binning + Lewis	0.09810538	2.611138	91
Baseline removal + binning + Lewis	0.09810538	2.623355	76
Wavelet + binning + Lewis	0.09810538	2.565328	74
Baseline removal + wavelet + binning + Lewis	0.09810538	2.481490	60

Figure 13 contains the spectra for several of the denoising methods, and we can see that the first y ion is absent from the beginning. The last plot for the spectra resulting from the combination of Baseline removal + wavelet + binning + Lewis contains the least amount of noise while retaining all of the signal peaks initially present.

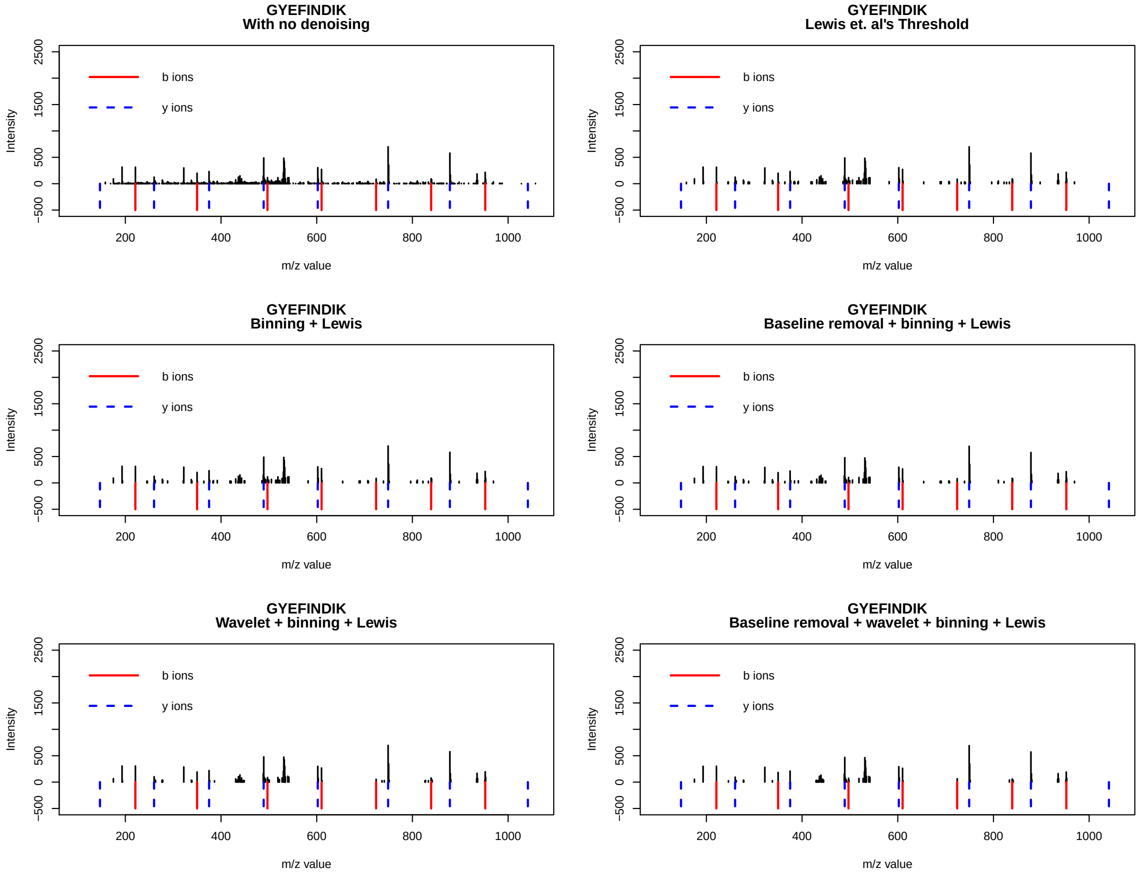


Figure 13: Observed spectra for peptide GYEFINDIK

Table 25 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that three of the distances increased compared to Lewis et. al's method, however, all of these were for ions that already had large distances; values in bold indicate increases.

Table 25: Comparison of the distances before and after denoising for the peptide GYEFINDIK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + wavelet + binning + Lewis	Difference after baseline + wavelet + binning + Lewis
57.8715	158.1386	100.2671	158.1386	100.2671	174.9857	**117.11423**
146.945	158.1386	11.1936	158.1386	11.1936	174.9857	**28.04073**
220.9345	221.0412	0.10673	221.0412	0.10673	221.0412	0.10673
260.029	260.2161	0.18709	260.2161	0.18709	260.2161	0.18709
349.9775	349.9164	0.06112	349.9164	0.06112	349.9164	0.06112
375.056	374.9333	0.12271	374.9333	0.12271	374.9333	0.12271
489.099	489.2649	0.16589	489.2649	0.16589	489.2649	0.16589
497.0455	497.1787	0.13315	497.1787	0.13315	497.1787	0.13315
602.183	602.2354	0.05241	602.2354	0.05241	602.2354	0.05241
610.1295	610.0545	0.075	610.0545	0.075	610.0545	0.075
724.1725	724.2306	0.05815	724.2306	0.05815	724.2306	0.05815
749.251	749.242	0.009	749.242	0.009	749.242	0.009
839.1995	839.0591	0.14042	839.0591	0.14042	839.0591	0.14042
878.294	878.3164	0.02235	878.3164	0.02235	878.3164	0.02235
952.2835	952.1422	0.14135	952.1422	0.14135	952.1422	0.14135
1041.357	1027.7791	13.5779	969.3653	71.9917	953.1313	**88.22565**

6.2 Long peptides

We now consider peptides that have a total weight greater than 1100 Da, which we classify as long based on the work of Offei (2017) [22]. For these peptides, we use the 0.10 percentile when binning.

6.2.1 Example 1

The first long peptide we examine is AFNEALPLTGVVLTK, which has a total weight of 1572.9 Da and contains 837 pairs of m/z and intensity values. The results in Table 26 show that several of the combinations result in smaller distance measures for this peptide.

Table 26: Results for peptide AFNEALPLTGVVLTK

Method	S_1	S_2	Dimension
None	0.2356724	2.861463	837
Lewis et. al's threshold	0.2320350	2.722083	190
Baseline removal	0.2356724	2.837751	659
Wavelet	0.2356724	2.858323	819
Binning	0.2356724	2.860950	834
Baseline removal + binning	0.2356724	2.837495	658
Wavelet + binning	0.2577242	2.861618	802
Baseline removal + wavelet + binning	0.2605260	2.836140	618
Baseline removal + Lewis	0.1887583	2.677443	149
Wavelet + Lewis	0.1805247	2.679161	152
Binning + Lewis	0.2320350	2.721112	187
Baseline removal + binning + Lewis	0.1887583	2.677443	149
Wavelet + binning + Lewis	0.1805247	2.674408	150
Baseline removal + wavelet + binning + Lewis	0.1833856	2.596614	126

However, we can see in Figure 14 that some of these methods are removing more than just noise from the spectrum; the methods Wavelet + binning + Lewis and Baseline + wavelet + binning + Lewis have removed many of the b and y ions from the spectra. Binning + Lewis does not appear to remove any more than Lewis et. al alone, and it has a slightly smaller S_2 as seen in Table .

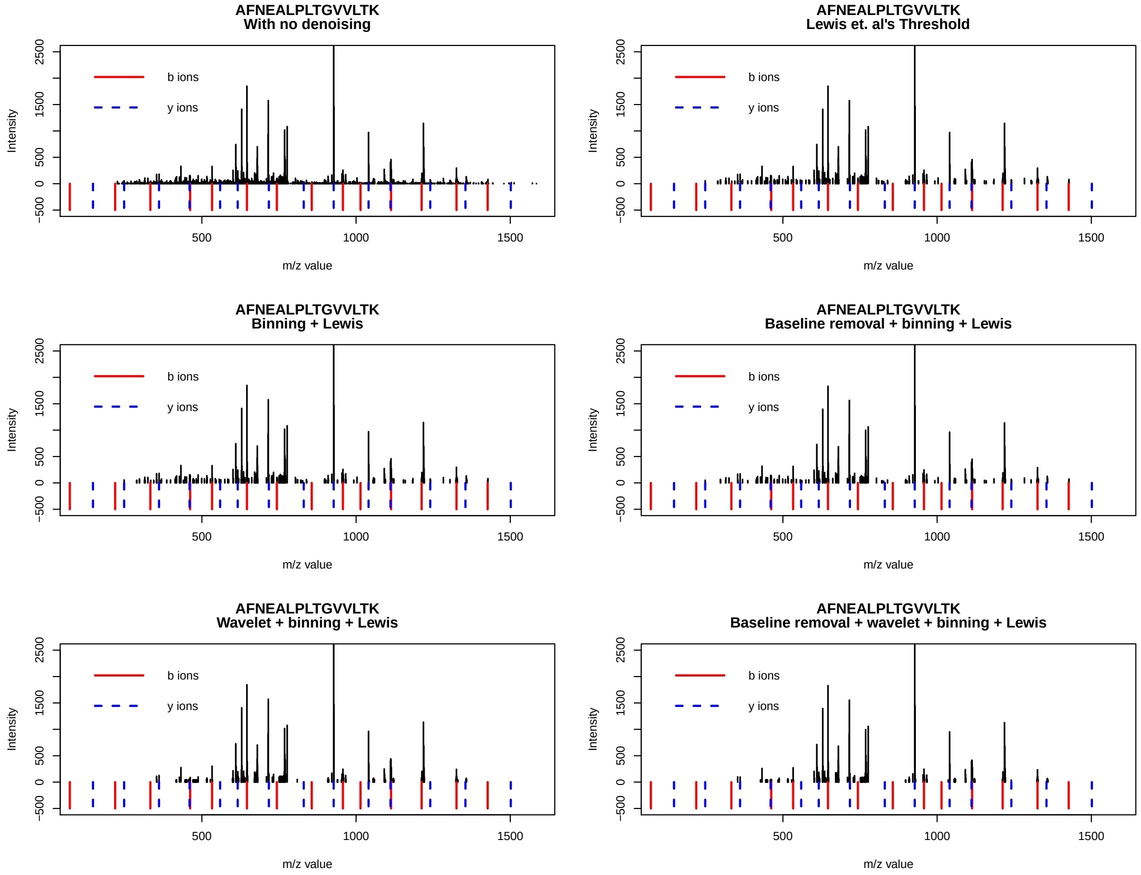


Figure 14: Observed spectra for peptide AFNEALPLTGVVLTK

Table 27 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 27: Comparison of the distances before and after denoising for the peptide AFNEALPLTGVVLTK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline removal + Lewis	Difference after baseline removal + Lewis
71.8871	226.0699	154.18283	248.0106	176.12346	248.0106	176.12346
146.945	226.0699	79.12493	248.0106	101.06556	248.0106	101.06556
218.9551	226.0699	7.11483	248.0106	29.05546	248.0106	29.05546
247.993	248.0106	0.01756	248.0106	0.01756	248.0106	0.01756
332.9981	333.2209	0.22279	334.1173	1.11924	334.1173	1.11924
361.077	361.3013	0.2243	361.3013	0.2243	361.3013	0.2243
460.1454	460.3021	0.15672	460.3021	0.15672	460.3021	0.15672
462.0411	462.2884	0.24729	462.2884	0.24729	462.2884	0.24729
533.0782	533.1494	0.07115	533.1494	0.07115	533.1494	0.07115
559.2138	559.2665	0.05274	557.6227	1.59112	557.6227	1.59112
616.2353	616.4191	0.18377	616.4191	0.18377	616.4191	0.18377
646.1622	646.1708	0.00858	646.1708	0.00858	646.1708	0.00858
717.2833	717.3715	0.08822	717.3715	0.08822	717.3715	0.08822
743.215	743.3388	0.12381	743.3388	0.12381	743.3388	0.12381
830.3673	830.7348	0.3675	826.9399	3.42742	826.9399	3.42742
856.299	855.869	0.43004	840.0734	16.22564	840.0734	16.22564
927.4201	927.4365	0.01636	927.4365	0.01636	927.4365	0.01636
957.347	957.4034	0.05638	957.4034	0.05638	957.4034	0.05638
1014.3685	1014.3723	0.0038	1003.0646	11.3039	1003.0646	11.3039
1040.5041	1040.4634	0.0407	1040.4634	0.0407	1040.4634	0.0407
1111.5412	1111.496	0.0452	1111.496	0.0452	1111.496	0.0452
1113.4369	1113.3696	0.0673	1113.3696	0.0673	1113.3696	0.0673
1212.5053	1212.438	0.0673	1212.438	0.0673	1212.438	0.0673
1240.5842	1240.3579	0.2263	1240.3579	0.2263	1240.3579	0.2263
1325.5893	1325.4674	0.1219	1325.4674	0.1219	1325.4674	0.1219
1354.6272	1355.9399	1.3127	1357.1927	2.5655	1357.1927	2.5655
1426.6373	1426.8048	0.1675	1426.8048	0.1675	1426.8048	0.1675
1501.6952	1503.2671	1.5719	1427.5935	74.1017	1427.5935	74.1017

6.2.2 Example 2

ENLMQVYQQAR is the next peptide that we look at. This peptide has a total weight of 1379.675 Da and contains 589 pairs. Table 28 shows that we see little change in the S_1 value across our different methods, but we can see a reduction in S_2, particularly with Baseline removal + binning + Lewis and Baseline removal + wavelet + binning + Lewis.

Table 28: Results for peptide ENLMQVYQQAR

Method	S_1	S_2	Dimension
None	0.08943833	2.878313	589
Lewis et. al's threshold	0.08662882	2.742305	129
Baseline removal	0.08662882	2.864216	477
Wavelet	0.08943833	2.892661	575
Binning	0.08943833	2.878099	588
Baseline removal + binning	0.08662882	2.862724	472
Wavelet + binning	0.08943833	2.891884	571
Baseline removal + wavelet + binning	0.08662882	2.877203	449
Baseline removal + Lewis	0.08662882	2.699101	106
Wavelet + Lewis	0.08662882	2.723046	117
Binning + Lewis	0.08662882	2.739983	128
Baseline removal + binning + Lewis	0.08662882	2.695682	105
Wavelet + binning + Lewis	0.08662882	2.723046	117
Baseline removal + wavelet + binning + Lewis	0.08972125	2.617224	89

The plots in Figure 15 show that the method Baseline removal + binning + Lewis denoises the spectrum without removing any more of the true signal peaks, whereas the method of Baseline removal + wavelet + binning + Lewis causes one of the b ions to be removed.

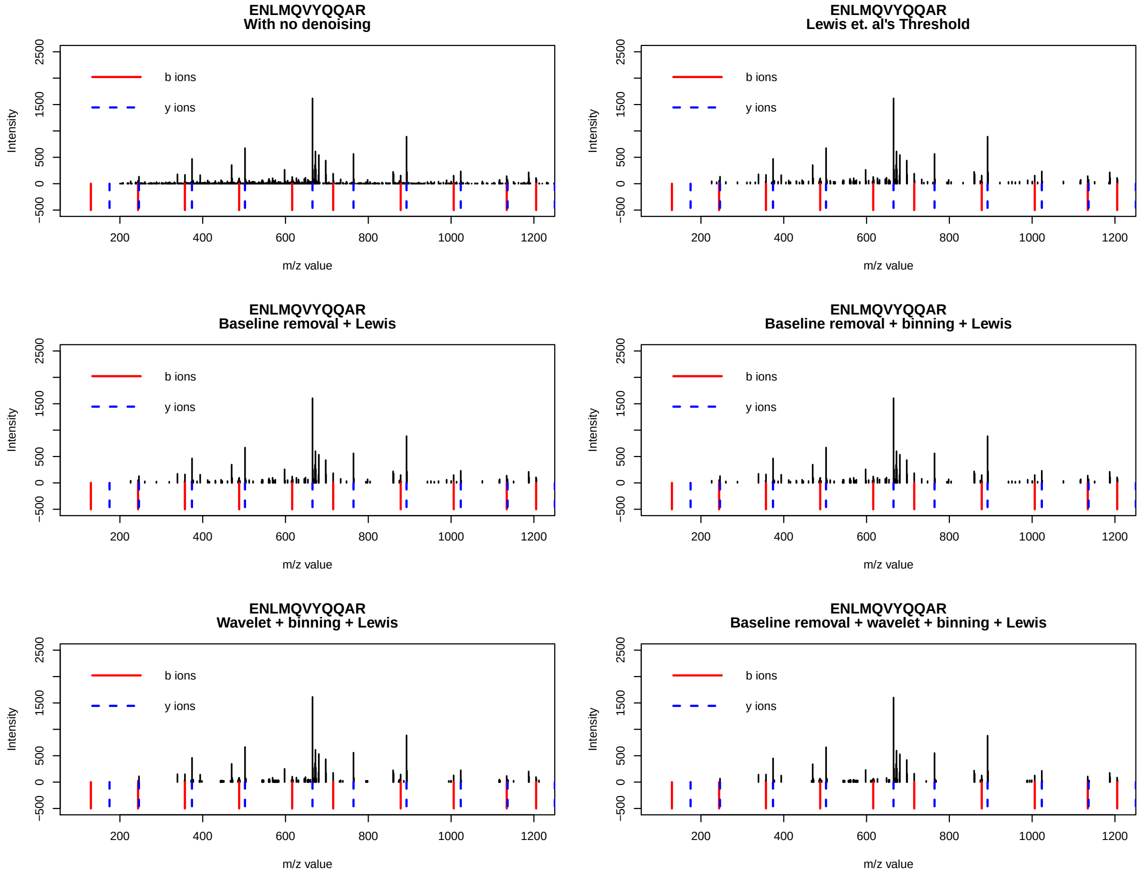


Figure 15: Observed spectra for peptide ENLMQVYQQAR

Table 29 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 29: Comparison of the distances before and after denoising for the peptide ENLMQVYQQAR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
129.893	200.0935	70.20052	225.9569	96.06391	225.9569	96.06391
174.951	200.0935	25.14252	225.9569	51.00591	225.9569	51.00591
243.936	243.8989	0.03715	243.8989	0.03715	243.8989	0.03715
245.9881	246.0997	0.11159	246.0997	0.11159	246.0997	0.11159
357.02	357.1619	0.14193	357.1619	0.14193	357.1619	0.14193
374.0471	374.253	0.20595	374.253	0.20595	374.253	0.20595
488.06	488.0701	0.01007	488.0701	0.01007	488.0701	0.01007
502.1061	502.2154	0.10935	502.2154	0.10935	502.2154	0.10935
616.119	616.1608	0.04183	616.1608	0.04183	616.1608	0.04183
665.1691	665.2491	0.07998	665.2491	0.07998	665.2491	0.07998
715.1874	715.1245	0.06295	715.1245	0.06295	715.1245	0.06295
764.2375	764.2294	0.00813	764.2294	0.00813	764.2294	0.00813
878.2504	878.1607	0.08969	878.1607	0.08969	878.1607	0.08969
892.2965	892.3641	0.06757	892.3641	0.06757	892.3641	0.06757
1006.3094	1006.1984	0.111	1006.1984	0.111	1006.1984	0.111
1023.3365	1023.3416	0.0051	1023.3416	0.0051	1023.3416	0.0051
1134.3684	1134.2614	0.107	1134.2614	0.107	1134.2614	0.107
1136.4205	1136.3928	0.0277	1136.3928	0.0277	1136.3928	0.0277
1205.4055	1205.1498	0.2557	1205.1498	0.2557	1205.1498	0.2557
1250.4635	1250.3263	0.1372	1206.1615	44.302	1206.1615	44.302

6.2.3 Example 3

Our next peptide of interest is peptide GYAGDTATTSEVK, with a total weight of 1299.605 Da and a size of 437 pairs. We see in Table 30 that while most of the methods result in the same S_1, some of the combinations produce a smaller S_2.

Table 30: Results for peptide GYAGDTATTSEVK

Method	S_1	S_2	Dimension
None	0.08671905	2.812878	437
Lewis et. al's threshold	0.08671905	2.681798	102
Baseline removal	0.08671905	2.782375	346
Wavelet	0.08671905	2.811624	418
Binning	0.08671905	2.811061	433
Baseline removal + binning	0.08671905	2.781704	345
Wavelet + binning	0.08671905	2.810763	412
Baseline removal + wavelet + binning	0.08671905	2.782621	324
Baseline removal + Lewis	0.08671905	2.618342	83
Wavelet + Lewis	0.07631833	2.596835	98
Binning + Lewis	0.08671905	2.681798	102
Baseline removal + binning + Lewis	0.08671905	2.618342	83
Wavelet + binning + Lewis	0.07631833	2.621794	95
Baseline removal + wavelet + binning + Lewis	0.12706684	2.489790	68

As in many of our other examples, we in Figure 16 that combining all three denoising methods with that of Lewis et. al results in the removal of some of the b and y ions as well. Baseline removal + binning + Lewis, however, does not appear to remove any more than Lewis et. al alone, yet results in a smaller S_2 value and data size.

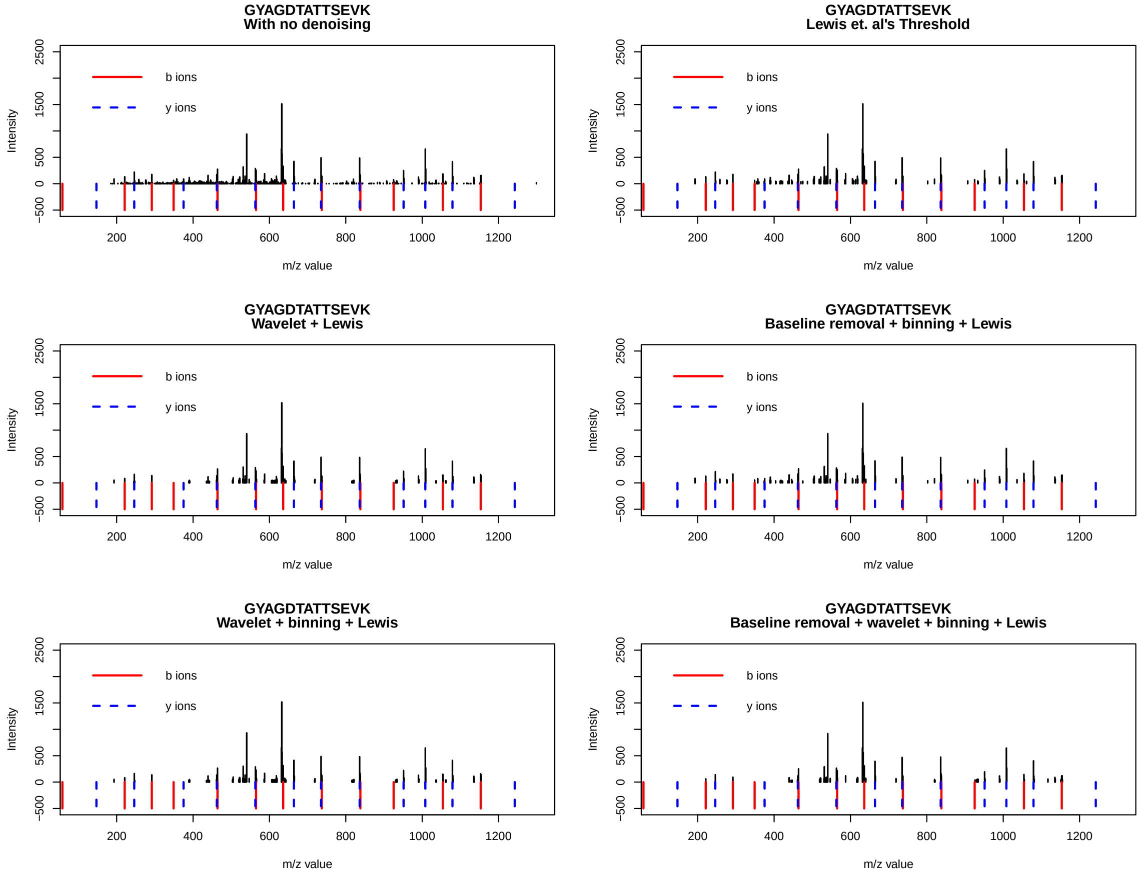


Figure 16: Observed spectra for peptide GYAGDTATTSEVK

Table 31 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 31: Comparison of the distances before and after denoising for the peptide GYAGDTATTSEVK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
57.8715	184.7585	126.88695	193.108	135.2365	193.108	135.2365
146.945	184.7585	37.81345	193.108	46.163	193.108	46.163
220.9345	221.0305	0.09599	221.0305	0.09599	221.0305	0.09599
246.0134	246.1802	0.16684	246.1802	0.16684	246.1802	0.16684
291.9716	292	0.0284	292	0.0284	292	0.0284
348.9931	349.0432	0.05005	349.0432	0.05005	349.0432	0.05005
375.0564	375.3398	0.28344	375.3398	0.28344	375.3398	0.28344
462.0884	462.1874	0.09904	462.1874	0.09904	462.1874	0.09904
464.0201	464.0333	0.01323	464.0333	0.01323	464.0333	0.01323
563.1364	563.2669	0.13051	563.2669	0.13051	563.2669	0.13051
565.0681	565.143	0.07491	565.143	0.07491	565.143	0.07491
636.1052	636.1932	0.08798	636.1932	0.08798	636.1932	0.08798
664.1844	664.2163	0.03191	664.2163	0.03191	664.2163	0.03191
735.2215	735.2527	0.03119	735.2527	0.03119	735.2527	0.03119
737.1532	737.2899	0.13672	737.2899	0.13672	737.2899	0.13672
836.2695	836.2596	0.00992	836.2596	0.00992	836.2596	0.00992
838.2012	838.163	0.03818	838.163	0.03818	838.163	0.03818
925.2332	925.1193	0.11388	925.1193	0.11388	925.1193	0.11388
951.2965	951.2425	0.05401	951.2425	0.05401	951.2425	0.05401
1008.318	1008.2764	0.0416	1008.2764	0.0416	1008.2764	0.0416
1054.2762	1054.2556	0.0206	1054.2556	0.0206	1054.2556	0.0206
1079.3551	1079.3062	0.0489	1079.3062	0.0489	1079.3062	0.0489
1153.3446	1153.0808	0.2638	1153.0808	0.2638	1153.0808	0.2638
1242.4181	1299.3097	56.8916	1154.2095	88.2086	1154.2095	88.2086

6.2.4 Example 4

Next, we examine peptide YLDLISNDESR, which has a total weight of 1324.638 Da and contains 538 pairs of m/z and intensity values. In Table 32 we see that while nearly all of the methods result in the same S_1, most of the combined approaches also give a smaller S_2.

Table 32: Results for peptide YLDLISNDESR

Method	S_1	S_2	Dimension
None	0.1089372	2.884250	538
Lewis et. al's threshold	0.1076044	2.816966	115
Baseline removal	0.1089372	2.877357	410
Wavelet	0.1089372	2.883385	534
Binning	0.1089372	2.884854	534
Baseline removal + binning	0.1089372	2.875772	405
Wavelet + binning	0.1089372	2.881782	527
Baseline removal + wavelet + binning	0.1089372	2.876680	392
Baseline removal + Lewis	0.1076044	2.785172	90
Wavelet + Lewis	0.1076044	2.739808	107
Binning + Lewis	0.1076044	2.816966	115
Baseline removal + binning + Lewis	0.1076044	2.782230	89
Wavelet + binning + Lewis	0.1076044	2.733961	105
Baseline removal + wavelet + binning + Lewis	0.1184800	2.737659	74

Based on Figure 17, it seems that Wavelet + binning + Lewis appears to work best for this peptide, as it denoises the spectrum the most without the removal of any true signal peaks.

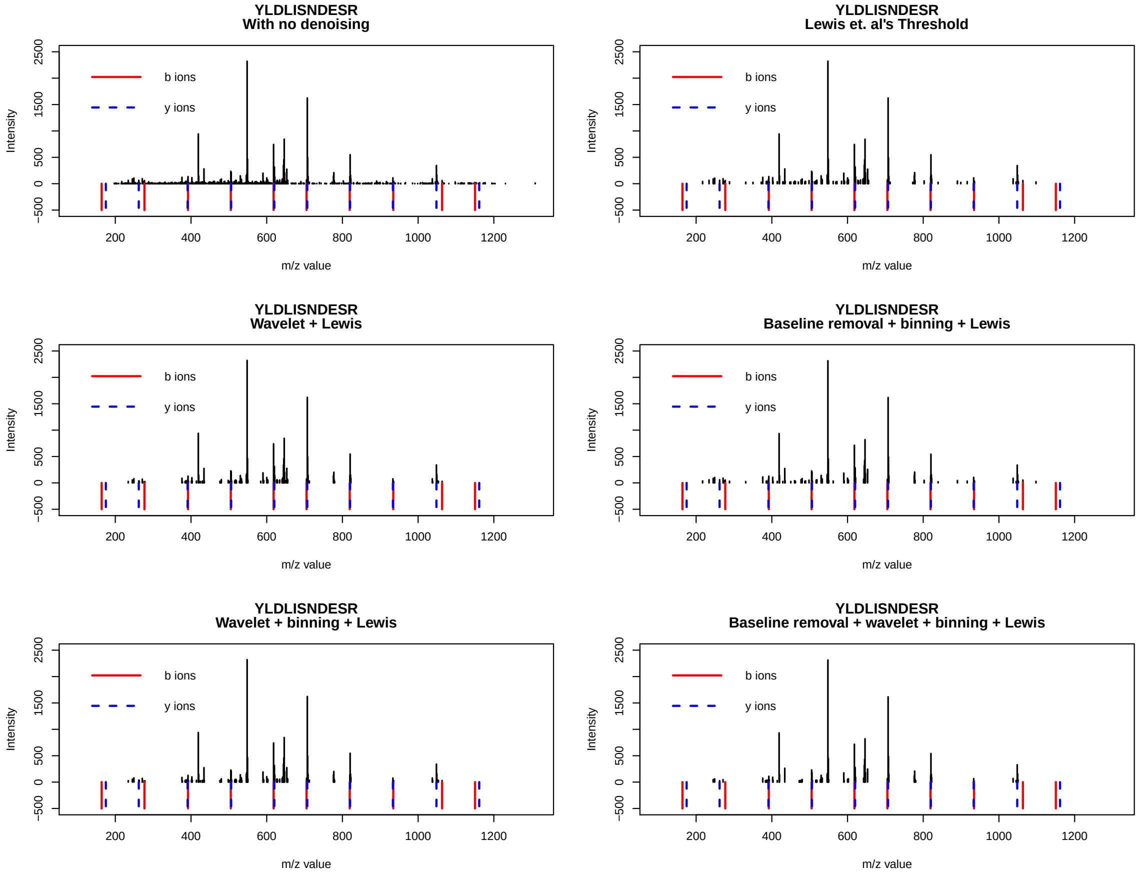


Figure 17: Observed spectra for peptide YLDLISNDESR

Table 33 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that four of the distances increased compared to Lewis et. al's method, however, all of these were for ions that already had large distances; values in bold indicate increases.

Table 33: Comparison of the distances before and after denoising for the peptide YLDLISNDESR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after wavelet + Lewis	Difference after wavelet + Lewis
163.913	197.1798	33.26678	217.1251	53.21211	234.126	**70.21298**
174.951	197.1798	22.22878	217.1251	42.17411	234.126	**59.17498**
261.983	262.0832	0.10016	262.0832	0.10016	262.0832	0.10016
276.997	277.0482	0.05125	277.0482	0.05125	277.0482	0.05125
391.026	391.1785	0.15253	391.1785	0.15253	391.1785	0.15253
392.024	392.1072	0.08324	392.1072	0.08324	392.1072	0.08324
505.108	505.1605	0.05252	505.1605	0.05252	505.1605	0.05252
506.053	506.1879	0.13493	506.1879	0.13493	506.1879	0.13493
618.192	618.115	0.07701	618.115	0.07701	618.115	0.07701
620.096	620.2951	0.1991	620.2951	0.1991	620.2951	0.1991
705.224	705.165	0.05896	705.165	0.05896	705.165	0.05896
707.128	707.2272	0.09923	707.2272	0.09923	707.2272	0.09923
819.267	819.7246	0.45761	819.7246	0.45761	819.7246	0.45761
820.212	820.3905	0.1785	820.3905	0.1785	820.3905	0.1785
933.296	933.2723	0.02372	933.2723	0.02372	933.2723	0.02372
934.294	934.253	0.04101	934.253	0.04101	934.253	0.04101
1048.323	1048.3232	0.0002	1048.3232	0.0002	1048.3232	0.0002
1063.337	1063.3253	0.0117	1063.3253	0.0117	1063.3253	0.0117
1150.369	1150.2228	0.1462	1097.8979	52.4711	1063.3253	**87.0437**
1161.407	1161.5	0.093	1097.8979	63.5091	1063.3253	**98.0817**

6.2.5 Example 5

We next look at the peptide MPPTEGETGGQVLGSK, with a total weight of 1587.767 Da and 566 pairs of m/z and intensity values. The results shown in Table 34 indicate that our methods give varying results for the S_1 and S_2 values, with Baseline removal + binning + Lewis having the lowest values.

Table 34: Results for peptide MPPTEGETGGQVLGSK

Method	S_1	S_2	Dimension
None	0.1117525	2.798592	566
Lewis et. al's threshold	0.1519754	2.537712	129
Baseline removal	0.1117525	2.772219	449
Wavelet	0.1117525	2.801225	555
Binning	0.1117525	2.797463	563
Baseline removal + binning	0.1117525	2.771676	448
Wavelet + binning	0.1117525	2.799859	551
Baseline removal + wavelet + binning	0.1117525	2.768980	424
Baseline removal + Lewis	0.1316536	2.479706	107
Wavelet + Lewis	0.1393517	2.500919	125
Binning + Lewis	0.1519754	2.533180	128
Baseline removal + binning + Lewis	0.1316536	2.473283	106
Wavelet + binning + Lewis	0.1393517	2.500919	125
Baseline removal + wavelet + binning + Lewis	0.1335939	2.413214	95

The spectra presented in Figure 18 show that several of the b and y ions are missing for all of our approaches, however, there are more removed when including the wavelet technique. Binning + Lewis appears to be the only combination that does not remove any more true signal peaks than using Lewis et. al's method.

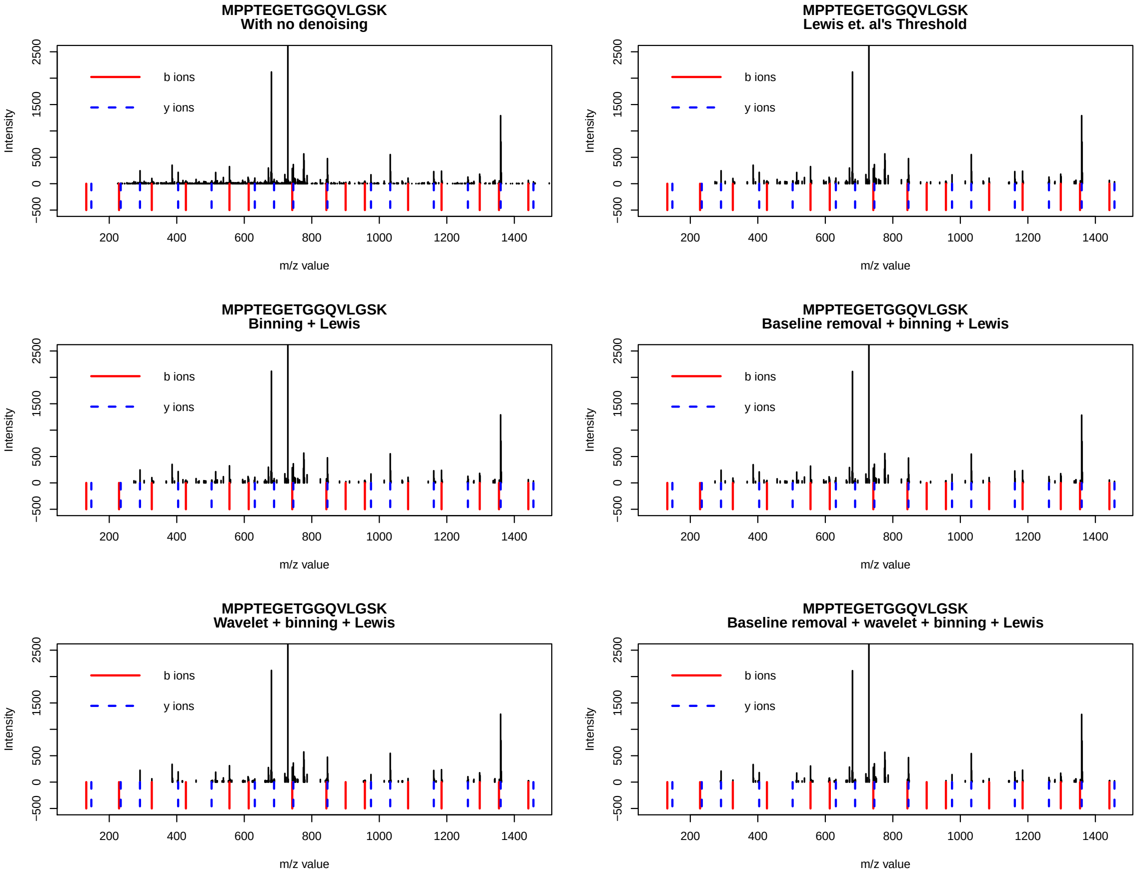


Figure 18: Observed spectra for peptide MPPTEGETGGQVLGSK

Table 35 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 35: Comparison of the distances before and after denoising for the peptide MPPTEGETGGQVLGSK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after binning + Lewis	Difference after binning + Lewis
131.89	225.0143	93.12433	273.2078	141.31782	273.2078	141.31782
146.945	225.0143	78.06933	273.2078	126.26282	273.2078	126.26282
228.9428	228.9616	0.01884	273.2078	44.26502	273.2078	44.26502
233.977	234.2344	0.25736	273.2078	39.23082	273.2078	39.23082
290.9985	291.207	0.20853	291.207	0.20853	291.207	0.20853
325.9956	326.1994	0.2038	326.1994	0.2038	326.1994	0.2038
404.0825	404.1811	0.09859	404.1811	0.09859	404.1811	0.09859
427.0436	427.0263	0.01735	427.0263	0.01735	427.0263	0.01735
503.1509	503.3522	0.20133	503.3522	0.20133	503.3522	0.20133
556.0866	556.1209	0.03431	556.1209	0.03431	556.1209	0.03431
613.1081	613.1916	0.08349	613.1916	0.08349	613.1916	0.08349
631.2099	631.2529	0.04303	631.2529	0.04303	631.2529	0.04303
688.2314	688.4293	0.19792	688.4293	0.19792	688.4293	0.19792
742.1511	742.1512	0.00008	742.1512	0.00008	742.1512	0.00008
745.2529	745.3229	0.07004	745.3229	0.07004	745.3229	0.07004
843.1991	843.2421	0.04303	843.2421	0.04303	843.2421	0.04303
846.3009	846.377	0.07611	846.377	0.07611	846.377	0.07611
900.2206	899.9946	0.22603	900.8806	0.66002	900.8806	0.66002
957.2421	957.3796	0.13748	957.3796	0.13748	957.3796	0.13748
975.3439	975.2871	0.05685	975.2871	0.05685	975.2871	0.05685
1032.3654	1032.4125	0.0471	1032.4125	0.0471	1032.4125	0.0471
1085.3011	1085.1804	0.1207	1085.1804	0.1207	1085.1804	0.1207
1161.4084	1161.2969	0.1115	1161.2969	0.1115	1161.2969	0.1115
1184.3695	1184.2944	0.0751	1184.2944	0.0751	1184.2944	0.0751
1262.4564	1262.4938	0.0374	1262.4938	0.0374	1262.4938	0.0374
1297.4535	1297.2452	0.2083	1297.2452	0.2083	1297.2452	0.2083
1354.475	1354.2428	0.2322	1355.3717	0.8967	1355.3717	0.8967
1359.5092	1359.3586	0.1506	1359.3586	0.1506	1359.3586	0.1506
1441.507	1441.4039	0.1031	1441.4039	0.1031	1441.4039	0.1031
1456.562	1456.6309	0.0689	1456.6309	0.0689	1456.6309	0.0689

6.2.6 Example 6

Next, we consider the peptide SGPLAGYPVVDLGVR. This peptide has a total weight of 1499.822 Da and consists of 889 pairs of m/z and intensity values. Table 36 shows that the S_1 values vary widely for this peptide for our different methods.

Table 36: Results for peptide SGPLAGYPVVDLGVR

Method	S1	S2	Dimension
None	0.2288104	2.870067	889
Lewis et. al's threshold	0.3380465	2.704235	200
Baseline removal	0.1728383	2.850140	707
Wavelet	0.2288104	2.867773	871
Binning	0.2288104	2.869464	885
Baseline removal + binning	0.1728383	2.850926	702
Wavelet + binning	0.2288104	2.866353	862
Baseline removal + wavelet + binning	0.1728383	2.845178	675
Baseline removal + Lewis	0.3082341	2.689345	168
Wavelet + Lewis	0.1394807	2.702404	175
Binning + Lewis	0.3380465	2.708585	199
Baseline removal + binning + Lewis	0.3082341	2.689345	168
Wavelet + binning + Lewis	0.1394807	2.702404	175
Baseline removal + wavelet + binning + Lewis	0.1362037	2.613653	144

The spectra in Figure 19 indicate than none of the combined methods are really better than Lewis et. al's, as they all remove true signal peaks. Baseline removal + binning + Lewis, which had slightly smaller S_1 and S_2 values, is only missing one more signal peak than Lewis et. al's method alone.

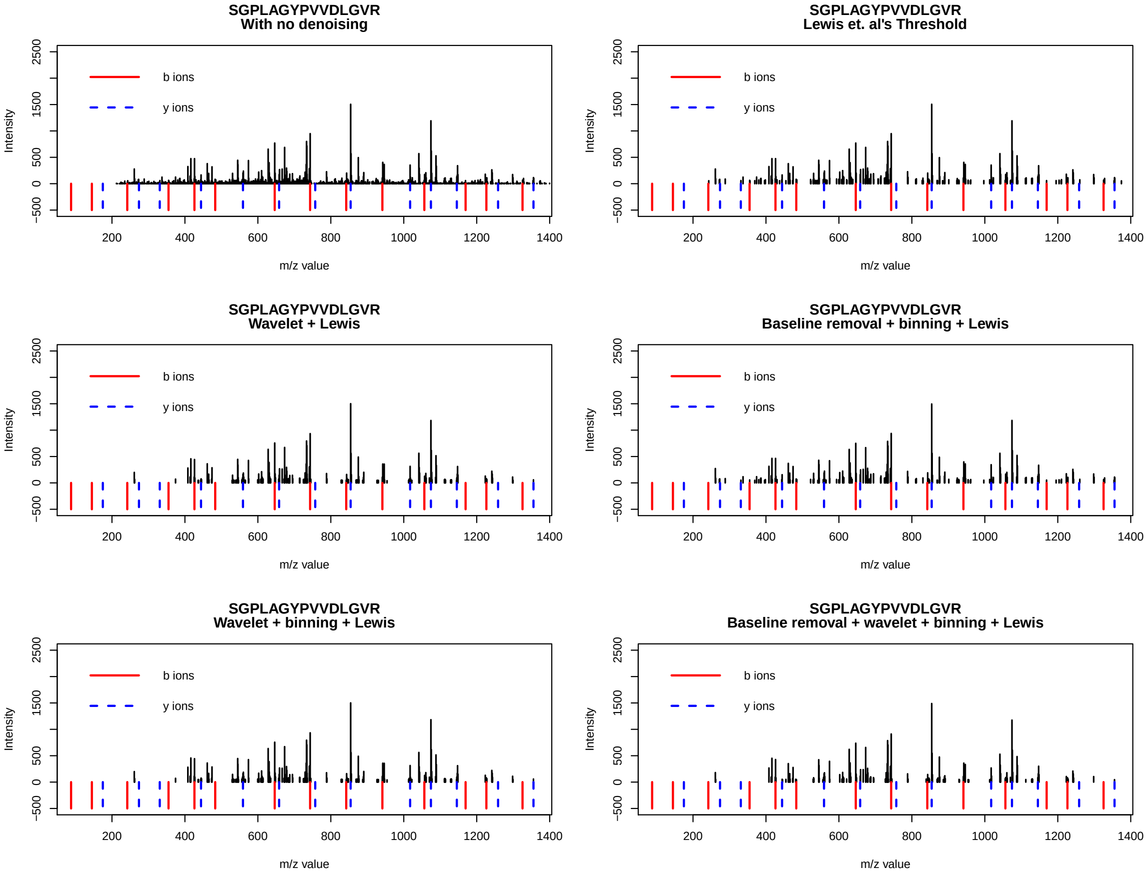


Figure 19: Observed spectra for peptide SGPLAGYPVVDLGVR

Table 37 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that five of the distances increased compared to Lewis et. al's method, however, four of these were for ions that already had large distances; values in bold indicate increases.

Table 37: Comparison of the distances before and after denoising for the peptide SGPLAGYPVVDLGVR

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
87.882	212.2774	124.39544	243.1531	155.27112	261.1791	**173.29714**
144.9035	212.2774	67.37394	243.1531	98.24962	261.1791	**116.27564**
174.951	212.2774	37.32644	243.1531	68.20212	261.1791	**86.22814**
241.9563	242.3252	0.36888	243.1531	1.19682	261.1791	**19.22284**
274.0194	274.2135	0.1941	273.265	0.75439	273.265	0.75439
331.0409	331.2965	0.25561	331.2965	0.25561	331.2965	0.25561
355.0403	355.2684	0.22813	355.2684	0.22813	355.2684	0.22813
426.0774	426.2693	0.19195	426.2693	0.19195	426.2693	0.19195
444.1249	444.3032	0.17826	444.3032	0.17826	444.3032	0.17826
483.0989	483.3888	0.28995	483.3888	0.28995	483.3888	0.28995
559.1519	559.2288	0.07692	559.2288	0.07692	559.2288	0.07692
646.1619	646.1638	0.00192	646.1638	0.00192	646.1638	0.00192
658.2203	658.2939	0.07358	658.2939	0.07358	658.2939	0.07358
743.2147	743.3928	0.17806	743.3928	0.17806	743.3928	0.17806
757.2887	758.8049	1.51617	743.3928	13.89594	743.3928	13.89594
842.2831	842.2435	0.03963	843.3152	1.03209	843.3152	1.03209
854.3415	854.3703	0.0288	854.3703	0.0288	854.3703	0.0288
941.3515	941.0935	0.25799	941.0935	0.25799	941.0935	0.25799
1017.4045	1017.3752	0.0293	1017.3752	0.0293	1017.3752	0.0293
1056.3785	1055.8073	0.5712	1055.8073	0.5712	1057.1526	0.7741
1074.426	1074.4003	0.0257	1074.4003	0.0257	1074.4003	0.0257
1145.4631	1145.3418	0.1213	1145.3418	0.1213	1145.3418	0.1213
1169.4625	1169.2534	0.2091	1169.2534	0.2091	1169.2534	0.2091
1226.484	1226.5149	0.0309	1226.5149	0.0309	1226.5149	0.0309
1258.5471	1258.9778	0.4307	1260.3969	1.8498	1260.3969	1.8498
1325.5524	1325.6619	0.1095	1325.6619	0.1095	1325.6619	0.1095
1355.5999	1355.5161	0.0838	1355.5161	0.0838	1355.5161	0.0838
1412.6214	1399.8624	12.759	1374.4906	38.1308	1356.3344	**56.287**

6.2.7 Example 7

Now, we consider the peptide IMNVLGEPVDMK, which has a total weight of 1345.687 Da and has 570 pairs of m/z and intensity values. The results in Table 38 indicate that the combinations involving Wavelet result in significantly smaller S_1 values, while those without give S_1 values closer to that of Lewis et. al and have slightly smaller S_2 values.

Table 38: Results for peptide IMNVLGEPVDMK

Method	S_1	S_2	Dimension
None	0.16964650	2.848481	570
Lewis et. al's threshold	0.27503375	2.672832	119
Baseline removal	0.19249050	2.832182	448
Wavelet	0.16964650	2.847091	565
Binning	0.16964650	2.847928	568
Baseline removal + binning	0.19249050	2.832182	448
Wavelet + binning	0.16964650	2.843941	554
Baseline removal + wavelet + binning	0.19249050	2.828861	433
Baseline removal + Lewis	0.27926467	2.644146	101
Wavelet + Lewis	0.06479625	2.752632	99
Binning + Lewis	0.27503375	2.669625	118
Baseline removal + binning + Lewis	0.27926467	2.644146	101
Wavelet + binning + Lewis	0.06479625	2.732018	92
Baseline removal + wavelet + binning + Lewis	0.06479625	2.707622	82

As expected based on our previous examples, Figure 20 shows that the methods involving wavelet technique that resulted in much smaller S_1 values result in the removal of several of our true signal peaks, whereas the method of Binning + Lewis does not remove any more than Lewis et. al and reduces the S_2 value slightly.

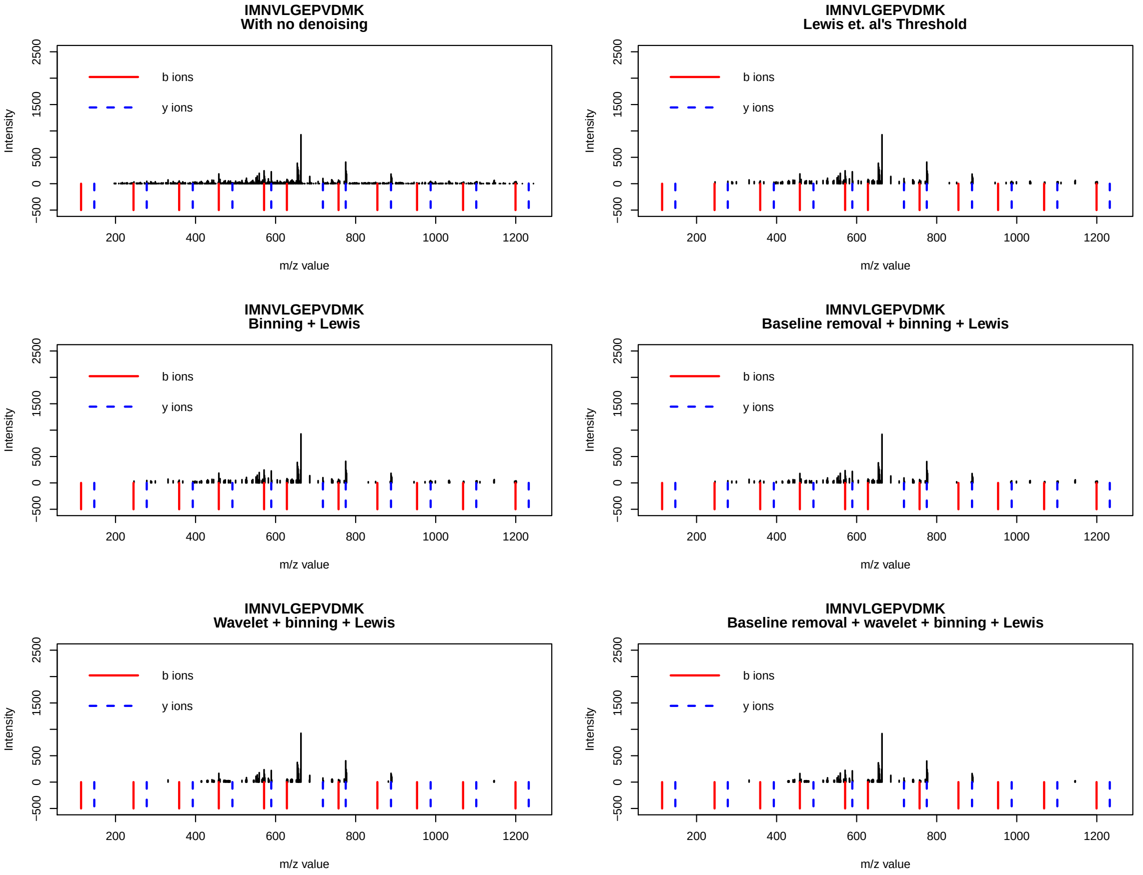


Figure 20: Observed spectra for peptide IMNVLGEPVDMK

Table 39 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 39: Comparison of the distances before and after denoising for the peptide IMNVLGEPVDMK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after binning + Lewis	Difference after binning + Lewis
113.934	196.9008	82.96679	246.1361	132.20205	246.1361	132.20205
146.945	196.9008	49.95579	246.1361	99.19105	246.1361	99.19105
244.974	245.2037	0.22967	246.1361	1.16205	246.1361	1.16205
277.985	278.1264	0.14143	278.1264	0.14143	278.1264	0.14143
359.017	359.1175	0.10055	359.1175	0.10055	359.1175	0.10055
393.012	393.2236	0.21157	393.2236	0.21157	393.2236	0.21157
458.0854	458.174	0.08861	458.174	0.08861	458.174	0.08861
492.0804	492.5035	0.42308	487.2874	4.79305	487.2874	4.79305
571.1694	571.158	0.01138	571.158	0.01138	571.158	0.01138
589.1332	589.1844	0.05119	589.1844	0.05119	589.1844	0.05119
628.1909	628.3601	0.16915	628.3601	0.16915	628.3601	0.16915
718.1762	718.2303	0.05409	718.2303	0.05409	718.2303	0.05409
757.2339	757.243	0.00908	757.243	0.00908	757.243	0.00908
775.1977	775.1703	0.02735	775.1703	0.02735	775.1703	0.02735
854.2867	853.4145	0.87215	849.8916	4.3951	849.8916	4.3951
888.2817	888.3892	0.10752	888.3892	0.10752	888.3892	0.10752
953.3551	953.2949	0.06024	945.8735	7.48156	945.8735	7.48156
987.3501	987.4028	0.05267	987.4028	0.05267	987.4028	0.05267
1068.3821	1068.1355	0.2466	1069.116	0.7339	1069.116	0.7339
1101.3931	1101.4226	0.0295	1100.8202	0.5729	1100.8202	0.5729
1199.4221	1199.0826	0.3395	1200.3292	0.9071	1200.3292	0.9071
1232.4331	1232.2655	0.1676	1201.3962	31.0369	1201.3962	31.0369

6.2.8 Example 8

Peptide LANELSDAADNK is our next example. This peptide has a true weight of 1260.607 Da and has 593 pairs of m/z and intensity values. In Table 40 we see that some of our combined methods result in lower S_1 and S_2 values than Lewis et. al's threshold.

Table 40: Results for peptide LANELSDAADNK

Method	S_1	S_2	Dimension
None	0.12190400	2.850635	593
Lewis et. al's threshold	0.09521222	2.603013	143
Baseline removal	0.12190400	2.819457	471
Wavelet	0.09776053	2.853144	566
Binning	0.12190400	2.849585	589
Baseline removal + binning	0.14690700	2.823181	470
Wavelet + binning	0.12407947	2.854947	560
Baseline removal + wavelet + binning	0.12190400	2.826879	436
Baseline removal + Lewis	0.09932562	2.576476	114
Wavelet + Lewis	0.09476600	2.590480	128
Binning + Lewis	0.09521222	2.596558	141
Baseline removal + binning + Lewis	0.09932562	2.576476	114
Wavelet + binning + Lewis	0.09476600	2.602966	126
Baseline removal + wavelet + binning + Lewis	0.10014214	2.545514	99

In the spectra in Figure 21, we see that only Binning + Lewis retains all of the b and y ions present when using Lewis et. al's threshold alone. The other methods seen here remove at least one additional b ion.

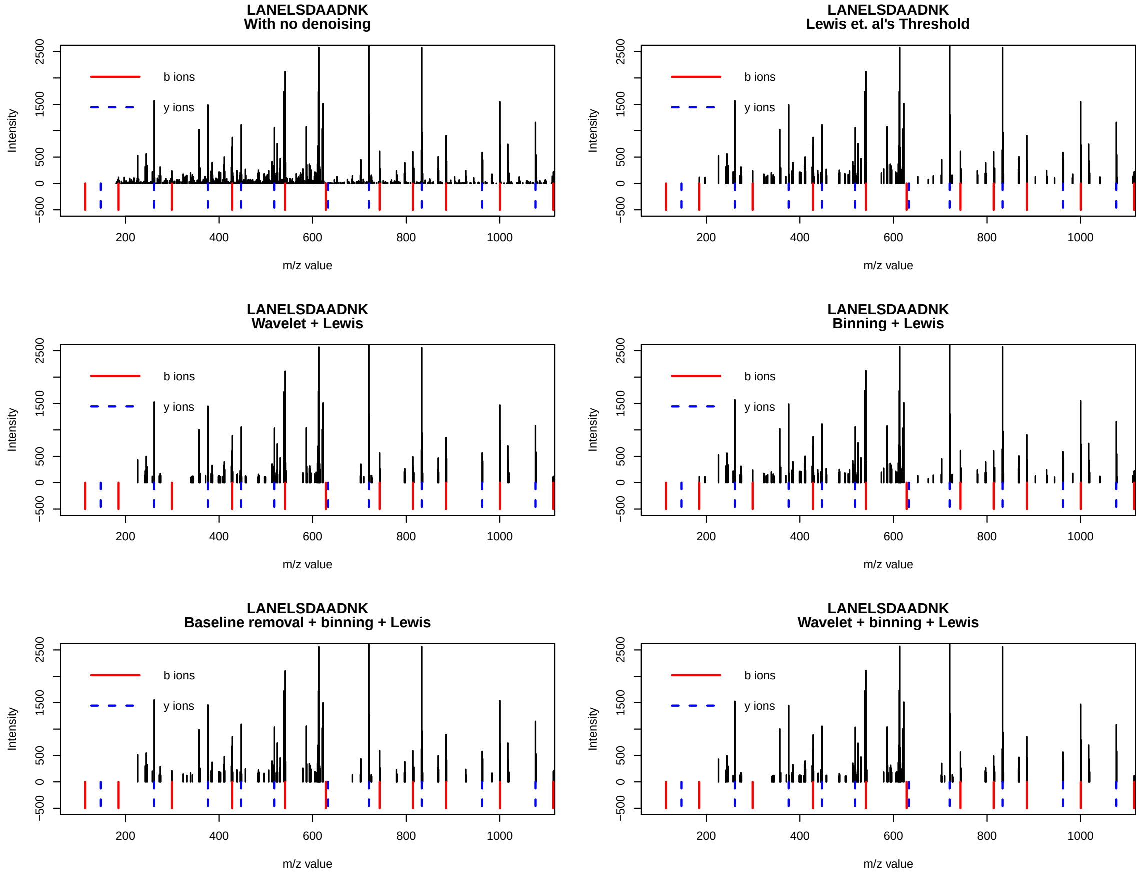


Figure 21: Observed spectra for peptide LANELSDAADNK

Table 41 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 41: Comparison of the distances before and after denoising for the peptide LANELSDAADNK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after binning + Lewis	Difference after binning + Lewis
113.934	180.1529	66.21886	185.0434	71.10941	185.0434	71.10941
146.945	180.1529	33.20786	185.0434	38.09841	185.0434	38.09841
184.9711	185.0434	0.07231	185.0434	0.07231	185.0434	0.07231
260.988	261.1651	0.17713	261.1651	0.17713	261.1651	0.17713
299.0141	299.1818	0.16772	299.1818	0.16772	299.1818	0.16772
376.015	376.1901	0.17506	376.1901	0.17506	376.1901	0.17506
428.0571	428.3161	0.259	428.3161	0.259	428.3161	0.259
447.0521	447.1982	0.14608	447.1982	0.14608	447.1982	0.14608
518.0892	518.2045	0.11533	518.2045	0.11533	518.2045	0.11533
541.1411	541.153	0.01192	541.153	0.01192	541.153	0.01192
628.1731	627.5925	0.58063	622.3089	5.8642	622.3089	5.8642
633.1162	633.2598	0.14363	622.3089	10.8073	622.3089	10.8073
720.1482	720.2249	0.07671	720.2249	0.07671	720.2249	0.07671
743.2001	743.2082	0.00809	743.2082	0.00809	743.2082	0.00809
814.2372	814.2045	0.03273	814.2045	0.03273	814.2045	0.03273
833.2322	833.2306	0.00155	833.2306	0.00155	833.2306	0.00155
885.2743	885.1655	0.10883	885.1655	0.10883	885.1655	0.10883
962.2752	962.2156	0.05956	962.2156	0.05956	962.2156	0.05956
1000.3013	1000.1318	0.1695	1000.1318	0.1695	1000.1318	0.1695
1076.3182	1076.2577	0.0605	1076.2577	0.0605	1076.2577	0.0605
1114.3443	1114.3638	0.0195	1114.3638	0.0195	1114.3638	0.0195
1147.3553	1147.303	0.0523	1147.303	0.0523	1147.303	0.0523

6.2.9 Example 9

For our next example, consider the peptide VLPAVAMLEER. This peptide has a total weight of 1227.677 Da and has 468 pairs of m/z and intensity values. With this peptide we see in Table 42 that Baseline removal + binning + Lewis appears to work the best, keeping the S_1 value the same while reducing the S_2 and the dimension.

Table 42: Results for peptide VLPAVAMLEER

Method	S_1	S_2	Dimension
None	0.1789447	2.863737	468
Lewis et. al's threshold	0.1697176	2.745889	98
Baseline removal	0.1803042	2.862331	364
Wavelet	0.1789447	2.861265	460
Binning	0.1789447	2.863506	463
Baseline removal + binning	0.1803042	2.861528	362
Wavelet + binning	0.1789447	2.864338	454
Baseline removal + wavelet + binning	0.2210534	2.872211	346
Baseline removal + Lewis	0.1697176	2.716188	86
Wavelet + Lewis	0.1235733	2.878539	90
Binning + Lewis	0.1697176	2.742713	97
Baseline removal + binning + Lewis	0.1697176	2.716188	86
Wavelet + binning + Lewis	0.1235733	2.876961	89
Baseline removal + wavelet + binning + Lewis	0.1186208	2.886919	68

In Figure 22, we see the spectra for some of the methods applied. The results when applying Baseline removal + binning + Lewis appear to contain slightly less noise than Lewis et. al alone, and this resulted in a slightly smalled S_2 value as seen above. The other methods, some of which gave smaller S_1 values, remove more of the true signal peaks that have low intensities.

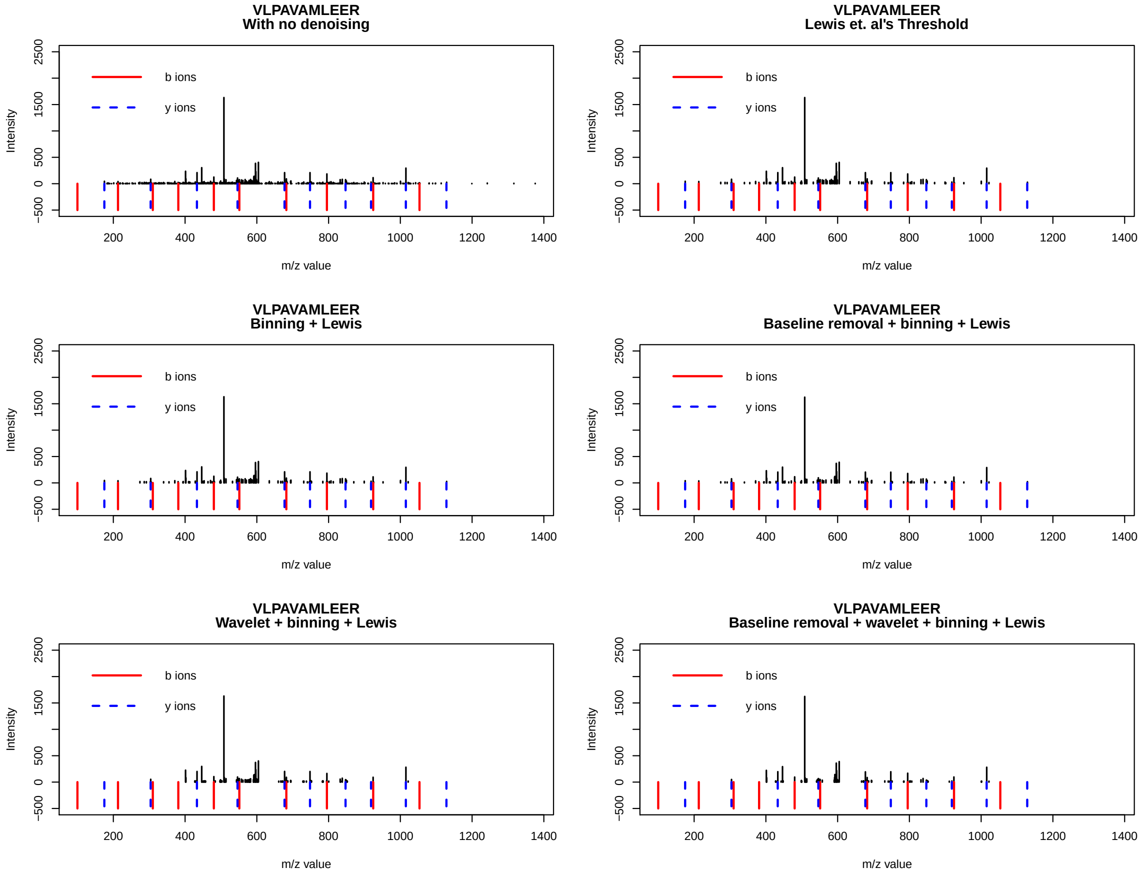


Figure 22: Observed spectra for peptide VLPAVAMLEER

Table 43 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 43: Comparison of the distances before and after denoising for the peptide VLPAVAMLEER

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after baseline + binning + Lewis	Difference after baseline + binning + Lewis
99.9184	175.1228	75.20443	175.1228	75.20443	175.1228	75.20443
174.951	175.1228	0.17183	175.1228	0.17183	175.1228	0.17183
213.0024	213.0143	0.01191	213.0143	0.01191	213.0143	0.01191
303.994	304.1819	0.18788	304.1819	0.18788	304.1819	0.18788
310.0552	309.2751	0.78005	305.1094	4.94582	305.1094	4.94582
381.0923	381.2859	0.19359	380.5169	0.57539	380.5169	0.57539
433.037	433.2485	0.21147	433.2485	0.21147	433.2485	0.21147
480.1607	480.2039	0.04316	480.2039	0.04316	480.2039	0.04316
546.121	546.1765	0.05545	546.1765	0.05545	546.1765	0.05545
551.1978	551.256	0.05818	551.256	0.05818	551.256	0.05818
677.161	677.3619	0.20088	677.3619	0.20088	677.3619	0.20088
682.2378	682.3727	0.13488	682.3727	0.13488	682.3727	0.13488
748.1981	748.1942	0.00395	748.1942	0.00395	748.1942	0.00395
795.3218	795.2596	0.06222	795.2596	0.06222	795.2596	0.06222
847.2665	847.3388	0.07231	847.3388	0.07231	847.3388	0.07231
918.3036	918.3628	0.05919	918.3628	0.05919	918.3628	0.05919
924.3648	924.0816	0.2832	924.0816	0.2832	924.0816	0.2832
1015.3564	1015.5257	0.1693	1015.5257	0.1693	1015.5257	0.1693
1053.4078	1053.5243	0.1165	1021.7844	31.6234	1021.7844	31.6234
1128.4404	1129.0244	0.584	1129.0244	0.584	1129.0244	0.584

6.2.10 Example 10

For our last example, we examine the peptide YLQDYGMGPETPLGEPK. This peptide has a total weight of 1894.89 Da and its data set consists of 776 pairs of m/z and intensity values. We notice in Table 44 that methods including baseline removal result in an increased S_1 value, while those including wavelet give a slightly smaller S_1.

Table 44: Results for peptide YLQDYGMGPETPLGEPK

Method	S_1	S_2	Dimension
None	0.1348874	2.872291	776
Lewis et. al's threshold	0.1397100	2.655205	185
Baseline removal	0.1348874	2.852457	621
Wavelet	0.1348874	2.874024	753
Binning	0.1348874	2.871606	772
Baseline removal + binning	0.1348874	2.850954	615
Wavelet + binning	0.1348874	2.873524	746
Baseline removal + wavelet + binning	0.1348874	2.853908	571
Baseline removal + Lewis	0.2058148	2.597392	149
Wavelet + Lewis	0.1384390	2.622072	160
Binning + Lewis	0.1397100	2.653064	184
Baseline removal + binning + Lewis	0.2058148	2.602419	147
Wavelet + binning + Lewis	0.1384390	2.622072	160
Baseline removal + wavelet + binning + Lewis	0.2109486	2.542300	123

The spectra shown in Figure 23 show that only the method of Binning + Lewis retains the same b and y ions that are present when using Lewis et. al's method alone. This method only gives a marginally smaller S_2 and only decreases the size of the data set by 1. Wavelet + binning + Lewis, which had smaller S_1 and S_2 values, removes some of the true signal peaks at the beginning and ends of the spectrum.

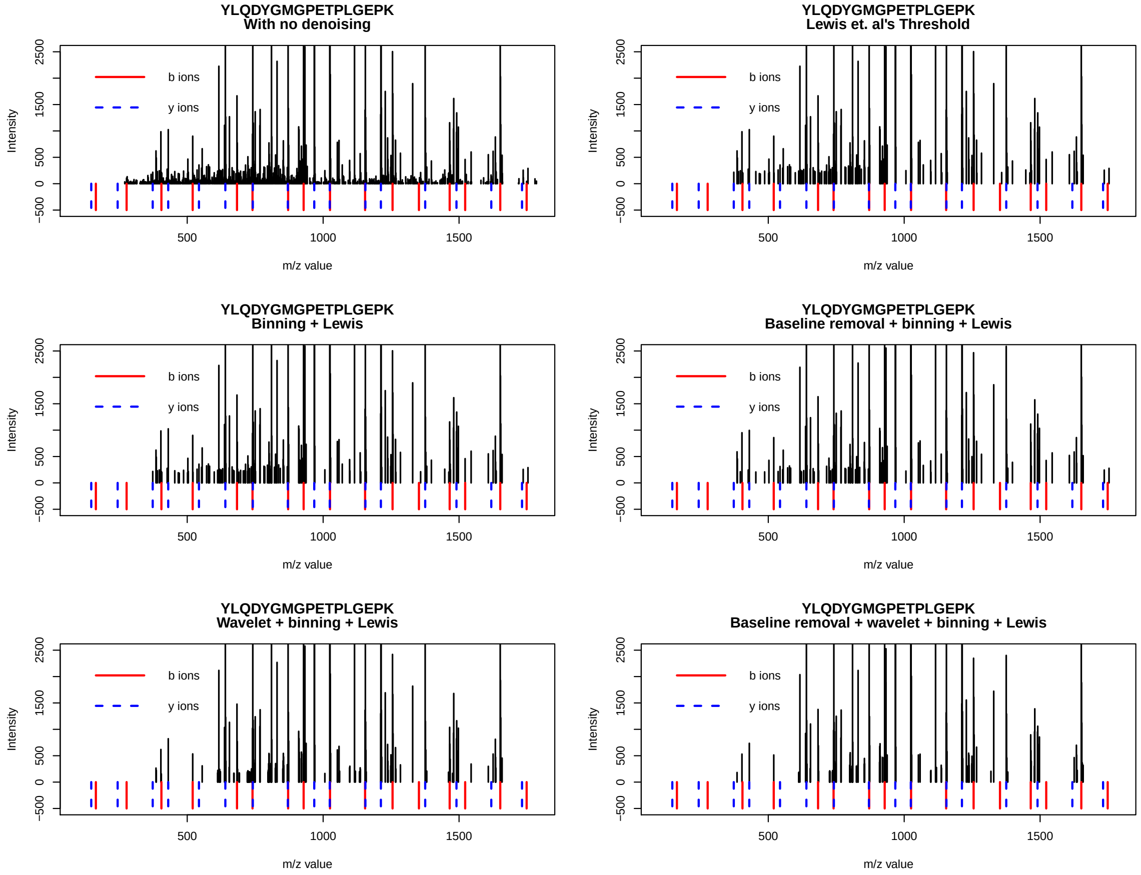


Figure 23: Observed spectra for peptide YLQDYGMGPETPLGEPK

Table 45 shows the locations of the b and y ions before any denoising is applied and the locations of the nearest m/z values after Lewis et. al's method and the method that appeared to work best based on our S_1 and S_2 values and the spectra plots have been applied. The table also shows the differences between the theoretical values and the observed values after applying these methods, representing the distances to the nearest m/z values for each of the b and y ions. Here, we see that none of the distances increased compared to Lewis et. al's method.

Table 45: Comparison of the distances before and after denoising for the peptide YLQDYGMGPETPLGEPK

Theoretical	Observed before denoising	Difference before denoising	Observed after Lewis et. al's threshold	Difference after Lewis et. al's threshold	Observed after binning + Lewis	Difference after binning + Lewis
146.945	270.2228	123.27784	373.1828	226.2378	373.1828	226.2378
163.913	270.2228	106.30984	373.1828	209.2698	373.1828	209.2698
243.9978	270.2228	26.22504	373.1828	129.185	373.1828	129.185
276.997	276.838	0.15902	373.1828	96.1858	373.1828	96.1858
373.0408	373.1828	0.142	373.1828	0.142	373.1828	0.142
405.056	404.6889	0.3671	404.6889	0.3671	404.6889	0.3671
430.0623	430.254	0.19173	430.254	0.19173	430.254	0.19173
520.083	520.1513	0.06825	520.1513	0.06825	520.1513	0.06825
543.1463	543.1995	0.05322	543.1995	0.05322	543.1995	0.05322
640.1991	640.2711	0.07196	640.2711	0.07196	640.2711	0.07196
683.146	683.0444	0.10163	683.0444	0.10163	683.0444	0.10163
740.1675	740.1331	0.03438	740.1331	0.03438	740.1331	0.03438
741.2471	741.2086	0.03854	741.2086	0.03854	741.2086	0.03854
870.2901	871.1085	0.81836	871.1085	0.81836	871.1085	0.81836
871.2075	871.1085	0.09904	871.1085	0.09904	871.1085	0.09904
928.229	928.0535	0.17553	928.0535	0.17553	928.0535	0.17553
967.3429	967.3562	0.0133	967.3562	0.0133	967.3562	0.0133
1024.3644	1024.3298	0.0346	1024.3298	0.0346	1024.3298	0.0346
1025.2818	1025.3848	0.103	1025.3848	0.103	1025.3848	0.103
1154.3248	1154.2417	0.0831	1154.2417	0.0831	1154.2417	0.0831
1155.4044	1155.3389	0.0655	1155.3389	0.0655	1155.3389	0.0655
1212.4259	1212.4493	0.0234	1212.4493	0.0234	1212.4493	0.0234
1255.3728	1255.2113	0.1615	1255.2113	0.1615	1255.2113	0.1615
1352.4256	1354.5903	2.1647	1358.6146	6.189	1358.6146	6.189
1375.4889	1375.3673	0.1216	1375.3673	0.1216	1375.3673	0.1216
1465.5096	1465.4159	0.0937	1465.4159	0.0937	1465.4159	0.0937
1490.5159	1490.3721	0.1438	1490.3721	0.1438	1490.3721	0.1438
1522.5311	1522.2805	0.2506	1522.2805	0.2506	1522.2805	0.2506
1618.5749	1618.6969	0.122	1625.316	6.7411	1625.316	6.7411
1651.5741	1651.4769	0.0972	1651.4769	0.0972	1651.4769	0.0972
1731.6589	1735.9524	4.2935	1735.9524	4.2935	1735.9524	4.2935
1748.6269	1748.619	0.0079	1753.5529	4.926	1753.5529	4.926

7 CONCLUSION

Since peptide identification is a field of increasing importance in fields ranging from biology to medicine, methods of improving the accuracy of identification using tandem mass spectrometry are becoming more important. We have presented various alternative methods for denoising tandem mass spectrometry data and compared their results to the method employed by Lewis et. al. We found that for 20 randomly chosen peptides, none of them were more effective than Lewis et. al's method when used individually. However, in many instances, we saw that using some combination of standard denoising methods with that of Lewis et. al gave better results than Lewis et. al alone. We compared the spectra that resulted from applying these methods to compare the presence and absence of the true signal peaks within them and saw that there typically was a method that did not remove more signal peaks than Lewis et. al's method, yet gave a cleaner spectrum with less noise and resulted in smaller distance values.

In future, we would want to apply these methods to a larger number of peptides in an attempt to find a "best" method to apply to any given peptide. In this thesis, we gauged the effectiveness of our various methods using the S_1 and S_2 values from Lewis et. al's scoring function, calculated based on the true λ vector. In future work, it may also be of interest to calculate these distance values when using randomly generated λ vectors akin to the technique used in step 2 of Lewis et. al's MCMC algorithm. Another area of additional study would be to explore different parameter values used in the denoising methods, such as the quantiles for baseline removal and binning or the percentile used in Lewis et. al's threshold.

BIBLIOGRAPHY

[1] Lewis, N. H., Hitchcock, D. B., Dryden, I. L., and Rose, J. R. (2018). Peptide refinement by using a stochastic search. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5), 1207-1236. doi:10.1111/rssc.12280

[2] Merrill, SA, and Mazza, AM, editors. (2006). National Research Council (US) Committee on Intellectual Property Rights in Genomic and Protein Research and Innovation; Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health. Washington (DC): National Academies Press (US); Genomics, Proteomics, and the Changing Research Environment. Available from: https://www.ncbi.nlm.nih.gov/books/NBK19861/

[3] Wade W. (2002). Unculturable bacteria–the uncharacterized organisms that cause oral infections. *Journal of the Royal Society of Medicine*, 95(2), 81-3.

[4] Human Genome Project Completion: Frequently Asked Questions. (2010, October 30). Retrieved February 4, 2019, from https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/

[5] Transcription, translation and replication. (n.d.). Retrieved January 14, 2019, from https://www.atdbio.com/content/14/Transcription-Translation-and-Replication

[6] The defense mechanisms of the adaptive immune system. (2016, August 04). Retrieved February 24, 2019, from https://www.ncbi.nlm.nih.gov/books/NBK279397/

[7] Park, J. W., and Graveley, B. R. (2007). Complex alternative splicing. *Advances in experimental medicine and biology*, 623, 50-63.

[8] Finehout, E. J. and Lee, K. H. (2004), An introduction to mass spectrometry applications in biological research. Biochem. Mol. Biol. Educ., 32: 93-100. doi:10.1002/bmb.2004.494032020331

[9] Mellon, F. A. (2003). Mass Spectrometry: Principles and Instrumentation. *Encyclopedia of Food Sciences and Nutrition*, 3739-3749.

[10] Paulo, J. A. (2013). Practical and efficient searching in proteomics: a cross engine comparison. Webmedcentral 4:WMCPLS0052. 10.9754/journal.wplus.2013.0052

[11] Tabb D. L. (2015). The SEQUEST family tree. *Journal of the American Society for Mass Spectrometry*, 26(11), 1814-9.

[12] Frank, A.M., and Pevzner, P.A. (2005). PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77 4, 964-73 .

[13] Medzihradszky, K. F., and Chalkley, R. J. (2015). Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass spectrometry reviews*, 34(1), 43-63.

[14] Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., and Pevzner, P.A. (1999). De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology: a journal of computational molecular cell biology*, 6 3-4, 327-42.

[15] Du, P., Kibbe W. A., and Lin S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, Volume 22, Issue 17, 1 September 2006, Pages 20592065, https://doi.org/10.1093/bioinformatics/btl355

[16] Huang, Y., Triscari, J. M., Pasa-Tolic, L., Anderson, G. A., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2004). Dissociation behavior of doubly-charged tryptic peptides: correlation of gas-phase cleavage abundance with Ramachandran plots. *Journal of the American Chemical Society*, 126(10), 3034-3035. doi:10.1021/ja038041t

[17] Armaanzas, R., Saeys, Y., Inza, I., Garca-Torres, M., Bielza, C., Peer, Y. V., and Larraaga, P. (2011). Peakbin Selection in Mass Spectrometry Data Using a Consensus Approach with Estimation of Distribution Algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3), 760-774. doi:10.1109/tcbb.2010.18

[18] Yang, C., He, Z., and Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC bioinformatics, 10, 4. doi:10.1186/1471-2105-10-4

[19] Stanford, T. E., Bagley, C. J., and Solomon, P. J. (2016). Informed baseline subtraction of proteomic mass spectrometry data aided by a novel sliding window algorithm. *Proteome science*, 14, 19. doi:10.1186/s12953-016-0107-8

[20] Lambers, J. (2006, May 9). PE281 Lecture 10 Notes. Retrieved February 17, 2019, from https://web.stanford.edu/class/energy281/WaveletAnalysis.pdf

[21] Hlaváĉ, V. (2015). Wavelets transformation. Czech Technical Univeristy in Prague. Lecture. Retrieved March 13, 2019, from http://people.ciirc.cvut.cz/hlavac/TeachPresEn/11ImageProc/14WaveletsEn.pdf

[22] Offei, F. (2017). Denoising Tandem Mass Spectrometry Data. East Tennessee State University Department of Mathematics and Statistics.

[23] Han, J., and Kamber, M. (2012). Data Mining: Concepts and techniques (3rd ed.). Haryana, India: Elsevier.

VITA

SKYLAR CARPENTER

Education:	M.S. Mathematical Sciences, East Tennessee State University, Johnson City, Tennessee 2019 B.S. Biology, Lincoln Memorial University Harrogate, Tennessee 2015
Professional Experience:	Graduate Assistant, ETSU, Johnson City, Tennessee, 2017–2019