5-2017

Peptide Identification: Refining a Bayesian Stochastic Model

Theophilus Barnabas Kobina Acquah
East Tennessee State University

Peptide Identification: Refining a Bayesian Stochastic Model

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

Theophilus B.K. Acquah

May 2017

Christina Nicole Holder Lewis, Ph.D., Chair

Robert M. Price Jr., Ph.D.

JeanMarie Hendrickson, Ph.D.

Keywords: Bayesian, Protein Identification, MCMC, Peptides, Ions

ABSTRACT

Peptide Identification: Refining a Bayesian Stochastic Model

by

Theophilus B.K. Acquah

Notwithstanding the challenges associated with different methods of peptide identification, other methods have been explored over the years. The complexity, size and computational challenges of peptide-based data sets calls for more intrusion into this sphere. By relying on the prior information about the average relative abundances of bond cleavages and the prior probability of any specific amino acid sequence, we refine an already developed Bayesian approach in identifying peptides. The likelihood function is improved by adding additional ions to the model and its size is driven by two overall goodness of fit measures. In the face of the complexities associated with our posterior density, a Markov chain Monte Carlo algorithm coupled with simulated annealing is used to simulate candidate choices from the posterior distribution of the peptide sequence, where the peptide with the largest posterior density is estimated as the true peptide.

DEDICATION

I dedicate this work to my supportive parents (Very Rev. Isaac & Paulina Acquah), my siblings (Perpetua & husband Ernest, Isaac Jnr., Benedicta and Bonnie Acquah), my nieces (Angela & Angelica Ewuradwoa Agyei-Tuffour) and nephew (William Kofi Agyei-Tuffour). I also dedicate this work to all who have inspired, encouraged and supported me, especially Samuel Kakraba, my friends and church family.

ACKNOWLEDGMENTS

My uttermost gratitude goes to God Almighty for His unconditional love, strength, protection and wisdom that has brought me this far. I am greatly indebted to my thesis advisor, Dr. Christina Nicole Holder Lewis for her guidance and countless hours of encouragement, reading, ideas and patience throughout the entire process. I also appreciate my thesis committee members (Dr. Robert M. Price Jr. & Dr. JeanMarie Hendrickson) for being generous with their expertise and time.

I wish to acknowledge Dr. Robert Gardner, Dr. Teresa Haynes and all professors in the Department Of Mathematics & Statistics at the East Tennessee State University (ETSU) for their efforts in shaping my career, academic and research goals.

Finally, I am equally thankful to my family, my church family (Dr. Nyarambi Arnold & his wife) and all my friends who have deposited and contributed to my life in one way or the other.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The proclivity to the use of highly sensitive biological mass spectrometry has resulted in the surge of many proteomic strategies and the need for effective, efficient, and accurate methods for protein identification. The pursuit for cutting edge methods has ameliorated inaccuracy in protein identification and sustained high-throughput proteomics. It is our goal in this thesis to enrich and improve the identification of proteins through refining an already developed Bayesian approach to protein identification.

1.1 Brief Overview

The neologism 'omics' informally points to fields of study in biology such as proteomics, genomics, or metabolomics, ending in -omics [28]. The entry point for viewing the other 'omics' sciences is genomics [10]. Genomics is the study of the genetic composition of organisms, a science that deals with the discovery and identification of all the sequences in the complete genome of a particular organism. The fundamental principle of molecular biology is a motivating factor for explaining the basic flow of genetic information. The DNA, which is a 6 billion letter code, is what stores our genetic instructions; the information that programs all of our cell activities. The DNA is transcribed into a form known as ribonucleic acid (RNA), a nucleic acid present in all living cells, which acts as a messenger carrying instructions from DNA for controlling the synthesis of proteins. The entire set of RNA or the sum total of all RNA molecules expressed from the genes of an organism (also referred to as its transcriptome) is subject to some editing, metamorphosed into messenger-RNA (mRNA). The messenger-RNA (mRNA) is a large family of RNA molecules that carries information to the ribosome, the protein factory of the cell, which in-turn translates the message into a protein.

1.2 Role of Proteins

Proteins are responsible for an endless number of tasks of cellular life, including inner organization and cell shape, product manufacture, waste cleanup and routine maintenance. Cells employ proteins to undertake tasks because the reproductive machinery is equipped to produce proteins. The chemically nimble nature of proteins accounts for the set up of the reproductive machinery. Proteins are substantially the most functionally sophisticated and structurally complex molecules known, from a chemical perspective. Through a covalent peptide bond, a protein molecule consists of a long chain of these amino acids, each linked to its neighbor. The 20 types of amino acids in proteins each has a different chemical property. Many thousands of different proteins are known, each with its own distinct amino acid sequence. Proteins can act as a catalyst, a fundamental task that can increase the rate of virtually all the chemical reactions within cells. Even though RNAs are able to catalyze some reactions, most biological reactions are catalyzed by proteins. Proteins can also act as scaffolds, transporters, signals or fuel in watery or greasy environments, and can alternate between hydrophobic and hydrophilic situations. There are several classes of proteins and many proteins may actually fit into more than one of these categories. In order to access the versatility of proteins, we can break down these roles to give a glimpse of how they function within and between cells. The foremost, and conceptually simplest, class of protein are structural proteins, which executes mechanical roles in organisms. Another class of proteins worth considering are enzymes, which are biological catalysts that reduces the activation energy of a reaction that takes place in a biological system. Electron-transport proteins are another class of proteins, connected clearly in electron transport by facilitating the passage of electrons from one molecule to another. Other classes of proteins worth mentioning are storage and transport proteins, hormones, receptors and nucleic-acid-binding proteins. From regulating bodily activities, providing structure to cells, acting as storage sites for amino acids and innumerable others, it is inevitable that proteins play a vital role in most biological processes.

1.3 The Proteomics Field and Protein Identification in Proteome Analysis

Proteome describes the entire complement of proteins expressed by a genome, cell, tissue or organism at a certain time. The term "proteome" is a blend of a "protein" and a "genome." Proteomics is therefore the study of the entire compendium of proteins, which are expressed by a genome, particularly their structure, functions, abundances, variations and modifications [11]. The objective of proteomics is to explain how the structure and function of proteins facilitate what they perform, interact with, and how they contribute to life processes. A usual biological sample may contain a plethora of different forms of protein. The vast diversity of proteins is acknowledged to the extent of tens of thousands of genes that results in hundreds of thousands of mRNAs, which successively generate potentially a million or more forms of proteins along with post-translational modifications. The information required to make functional molecules called proteins are contained and carried out by most genes and other molecules are produced by a few genes that help proteins to be assembled by cells. Consisting of two major steps, transcription and translation also known as gene expression, the migration from gene to protein is a complex process tightly controlled within each cell. With an understanding of the structure and function of proteins, scientists in proteomics start with the protein and work backwards to find the exact gene responsible for its production.

Reasons can be advanced for the tremendous progress in the study and proclivity towards proteins compared to mRNA or DNA. Being driven by both genetic and environmental factors, proteins do not only show risk or disposition but also a measure of actual biological and disease status [3]. Proteins are therefore a useful source of potential biomarkers. Proteomic tests can therefore be carried out on serum and urine, which are effortlessly and easily accessible, unlike mRNA or DNA [11, 17]. A sub-discipline of proteomics is clinical proteomics, which entails the application of proteomic technologies on clinical specimens like urine, blood and serum [18]. Cancer is a model disease for applying such technologies with the aim of identifying unique biomarkers and biosignatures for prognosis, diagnosis and therapeutic prediction. The rapid progress in cancer genomics over the years is limited to furnishing us with a glimpse of what may be determined by the genetic code. However, the need to access in real time what is happening in a patient leads to finding revealing and significant proteins that provides understanding into the biological processes of cancer development. The underlying factor is that genes are only the "recipes" of the cell, while the proteins encoded by the genes are fundamentally the functional and useful players driving both normal and disease physiology [18]. Eliciting and provoking considerable protein-focused research for the search of biomarkers is the accessibility of cancer-related proteins to help diagnose variety of cancers, diseases, or viruses with the aim of early detection. Further development in the work of proteomics is aiding clinicians in the early diagnosis of colorectal or colon cancer, the third leading type of cancer in males and fourth in females in the US. In essence, proteomics may soon help clinicians in the early and rapid incisionless treatment of colorectal cancer. The SimpliPro Colon test by the Applied Proteomics, Inc, is one such test that provides results for the risk of colorectal cancer and advanced adenoma [3]. The test measures and analyzes 11 protein markers connected with a risk of colorectal cancer and advanced adenoma [3]. With proteomics having the ability to cross-examine a variety of biospecimens for their protein contents and accurately measuring their concentration, the efficacy of most drugs is hinged on targeting proteins and some being proteins themselves [11, 18]. It is worth noting that protein drugs are receiving overwhelming acceptance in therapeutics. A drug's efficiency is connected to the level of its binding with the proteins in the blood serum, such that the less bound a drug is, the more efficiently it can diffuse through the cell membrane (Meyer and Guttman, 1968; Koch-Wester and Sellers,1976) [8]. Having an understanding of the structure and functioning of abnormal proteins, identifying and isolating them will help researchers locate the precise piece of mutated gene causing protein malfunction, and remedy the defective DNA [19]. The process of unlocking the previously opaque functioning of our cells and cellular machinery is not very distant, owing to the constant inspection of our genomic DNA and its protein products [19]. The advancement in proteomics is leading to the stage where scientists ultimately hope to make personalized or precision medicine tailor made for an individual, enhancing its effectiveness with less side effects [11].

Protein extraction, separation, identification and characterization forms major steps in proteome analysis. Protein identification is a very useful procedure in determining and discovering the identity of a protein-protein interaction, characterizing proteomes to recount biological processes and to discover disease-related biomarkers, pharmaceutical targets and protein functions. An obvious and fundamental way to identify a protein is to go through the excruciating process of finding its sequence of amino acids. The process involves but is not limited to employing restriction enzymes, running gels and finding masses. Properly sequencing microbial genome is not only necessary for producing accurate reference genomes for microbial identification but also other comparative genomic studies. However, the challenge surfaces in protein identification precisely in microbial samples, when an organism's genome has not been sequenced [11]. Being the foundation of the biosphere, judging from both an environmental and evolutionary viewpoint, it is estimated that microbial species accounts for about 60% of the Earth's biomass [4]. The search for antibiotics is on the surge with microbiologists discovering new ways to explore the extensive universe of undiscovered microbes [6]. Hugely inhibiting scientific knowledge of microbial life and the hunt for new antibiotics is the fact that only 1%—10% of microbes in the ecosystem can be cultured and an estimated 85% —99% of bacteria and archaea are unable to be grown in the lab [6, 11]. The minimal progress made in the field of environmental proteomics is not only restricted to the countless unidentified microbes but also the evidence of post translational modifications displayed by microbes that are cultured [11]. In what has come to be known as 'the mysterious dark matter of the microbial world', researchers are yearning for alternative ways to investigate the array of uncultured organisms [6]. Correctly identifying these microbes via protein identification is of enormous significance specifically in ecological samples such as soil and water samples (Schulze, 2004) [11]. The move will go a long way to aid in the advancement of cultivation methods and uncover breathtaking amounts of microbial diversity in samples cutting across soil to permafrost, marine sponges, hydrothermal vents and the crevices of the human body [6]. As discussed previously, proteins are known for carrying out the tasks of life and keeping us healthy through guiding our bodies activities and defending us against infection. Having the ability of controlling cell functions, defects in the instructions for making a protein can prevent the cell from functioning properly. Proteins in their mutant forms or enhanced proteolytic degradation of mutants proteins is a common molecular pathological mechanism causing genetic diseases such as cystic fibrosis, alpha-1-antitrypsin deficiency, phenylketonuria, mitochondrial acyl-COA dehydrogenase, and hemophilia [22]. Correctly identifying proteins will therefore generally help in the advancement of clinical proteomics such as determining whether an organism has a genetic disease.

The scope and complexity of any proteome is so vast that current technologies are not sufficient to provide complete detection and quantification of the proteins present [30], a reason necessitating more methods in the area of protein identification. Further corroborating this assertion, is the fact that the human genome has about $20,300$ protein-encoding genes and the total number of proteins in human cells is estimated to be $0.25 - 1 \times 10^6$ [30]. A broad range of methods coupled with tools are easily accessible to complement various proteomic approaches and ensure, as much as possible, proteins within any particular experiment are identified correctly. The limitation with current methods of protein identification, emanating from inadequate number of genome sequences, incomplete ion sequences, and noisy data, are also hindering, more critically, the accuracy and effectiveness of protein identification [11]. Researchers are relentlessly on the hunt for the most-up-to-date methods to aid in the maximization of true protein identification and curtail inaccurate identifications. In this thesis, we seek to refine an already developed Bayesian approach which aims to enhance and improve the identification of proteins.

1.4 Overview of Thesis

The arrangement of the thesis is as follows. Chapter 2 highlights some technological advancement made in proteomics by way of obtaining the proteomic profile of a sample. The fundamental concepts of peptide fragmentation is discussed in Chapter 3, leading us to the construction of the theoretical spectrum. Chapter 4 highlights our refined Bayesian model. Inherent in this is our likelihood function in section 4.2 and the priors we use in our model. Section 4.8 defines our posterior distribution. Chapter 4 ends with discussions on the Markov Chain Monte Carlo algorithm and how we aided our search of the true peptide with simulated annealing. Chapter 5 gives results for our work. Chapter 6 concludes the thesis with a discussion.

2 TECHNOLOGY OF PROTEOMICS AND DEVELOPMENTS IN PROTEIN IDENTIFICATION METHODS

The advancement of various techniques with increase sensitivity resolution, high-throughput applications and capability to analyze complex samples has led to the transition from protein chemistry to the modern proteomic field. Several technologies have given means to obtaining the proteomic profile of a sample as well as providing an understanding of cell physiology by way of depicting the molecular biodescriptors of gene expression, the proteins [11, 15].

2.1 Two-Dimensional Gel Electrophoresis

In the 1970s, the 2D gel electrophoresis (2DE) was one of the oldest technologies developed. Forming the premise of protein separation in a lot of proteome studies, the resolving capacity of 2D gel electrophoresis for protein separation is demonstrable and unparalleled. As a primary application, the 2DE can be used for protein expression profiling. Protein expression of any two different samples are examined and compared both quantitatively and qualitatively [21, 11]. The 2DE also has the capacity to resolve proteins that have been through some sort of post-translational modifications and different forms of proteins arising from proteolytic processing [21]. In spite of setting the gauge for protein separation procedures, the 2DE has not rid itself of several critisms. Even though large number of proteins can be visualized and separated, the collective and unaddressed procedure creates identification difficulties for any individual protein.

2.2 Mass-Spectrometry (MS)-Based Proteomics

Paving way for greater productivity in proteomics, and the formalization of the proteomics field, is the combining effect of techniques for large-scale protein separation (2DE) with high fidelity, rigorous mass spectrometry-based methods to the characterization and analysis of the separated proteins [15]. Dawning as the basic tool for protein identification and the linchpin of proteomics, mass spectrometry has become an indispensable tool to correlate proteins to their genes. With the overwhelming impact made in mass spectrometry, it is unusual, based on the mass analysis of protein-peptide for the identity of hundreds of proteins, to be unfolded in a sole proteome project [15]. Mass spectrometry has also seen improvements in terms of speed, accuracy and sample weight range over the years, making them complaisant to greater applications in proteomics and other areas of life sciences [16]. On a broad assessment, mass spectrometry and proteomics can be used for mining of proteomes, sequencing of proteins, identification of the type locations of post-translational modifications (PTMs), protein profiling and identification of protein networks. In terms of sequencing, structural information relating to peptide masses or amino acid sequences obtained from mass spectrometry is useful in the identification of protein by searching through protein databases and nucleotide [21]. The three important stages involved in the harvesting of protein information by mass spectrometry are sample preparation, sample ionization and mass analysis [21].

A good sample preparation for mass-spectrometry is useful for optimization of a sample and critical for good data. One common approach in most proteomics by which complex protein mixtures is resolved is by using a 1 —or —2D polyacrylamide gel electrophoresis (PAGE) [21]. The clear goal is to refine the sample and analyze it by mass spectrometry by first extracting the protein or its constituent peptides from the gel. In order to enhance the efficiency of extraction from the gel, a protein is often "in-gel" digested with a protease and now predominantly applied to both 1 —or —2D gels [21]. Also, very instrumental in the process of sample preparation of the peptide is sample purification, which is often required before being analyzed by mass spectrometry. The reverse-phase chromatography, available in a range of forms, is one such approach of peptide purification together with others such as Framingham, ZipTips (Millipore), mass or by high-pressure liquid chromatography (HPLC) [21].

Intrinsic in the analysis of peptides and proteins, and of all biological samples, is the ability to have molecules that are dry and charged. This is attainable by converting them to gas-phase ions from charged molecules [21, 29]. Due to the challenge imposed by hard ionization techniques, such as whole degradation of samples, matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), are the two "soft ionization" methods that have overcome the limitations without remarkable loss of sample coherence [21]. ESI creates fine mists of charged droplets by a potential difference placed between a capillary and the inlet to the mass spectrometer, through a liquid sample flow from a micro capillary tube into the outlet of the mass spectrometer [21, 29]. The solvent evaporates through the application of either drying gas or heat, ultimately resulting in the formation of desolvated ions. Peculiar with ESI is the creation of highly charged ions devoid of fragmentation [29]. Stemming from depth of work into the usage of lasers for ionization of biomolecules, MALDI is a pulse ionization approach where sample is incorporated into a matrix and analyte mixture and eventually subjected to radiation by laser. MALDI also comes with benefits of having the full procedure and analysis being automated, as well as samples often used directly without any purification after in-gel digestion [21, 29].

Mass spectrometry as a systematic method does not only apply ionization in its process, but also mass analysis of the compounds. Attaining this mission is a mass analyzer, a part of the mass spectrometer that conveys ionized mass and disperses them based on the mass-to-charge (m/z) ratios and yields them to the detector [21]. This is eventually detected and converted to a digital output. Mass analyzers commonly used in mass spectrometry for the separation of ions are quadrupole mass analyzer, time of flight (TOF) mass analyzer, magnetic sector mass analyzer, electrostatic sector mass analyzer, quadrupole ion trap mass analyzers and ion cyclotron resonance. As one of the simplest mass analyzers operating without an electric or magnetic field, TOF instrument separates ions based on the kinetic energy and velocity of the ions [21]. The m/z ratio of an ion for a TOF analyzer is measured by accessing the time required to cut across the length of a flight tube. Incorporating into some TOF mass analyzers is an ion mirror, placed at the end of the flight, to reflect ions back to the detector through the flight tube. The ion mirror plays the role of increasing the length of the flight tube, as well as correcting for small energy difference among ions [21].

2.3 Mass Spectrometry Methods

Though mass spectrometers may have various components, four fundamental features common with most mass spectrometers are an ionization source, one or more mass analyzers, an ion mirror and a detector [21]. The diverse mass spectrometry platforms are evident of, and also a result of, the different set-ups of the mass spectrometer components, most of which trace their names from the ionization source and mass analyzer. The special design of these instruments go a long way to aid in measuring the masses of the constituents inherent in the samples in question. Some of the most common mass spectrometers are the triple quadrupole, more often used to obtain amino acid sequences, the quadrupole-TOF, which is a "hybrid" mass spectrometer emerging from combining different ionization sources with mass analyzers [21]. Others are the Fourier-transform ion cycloton resonance (FT-ICR) mass spectrometer, the matrix-assisted laser desorption and ionization-time of flight mass spectrometry (MALDI- TOF), surface-enhanced laser desorption and ionization mass spectrometry (SELDI-TOF), MALDI-QqTOF and ESI/TOF mass spectrometers. In this section, we discuss two relatively novel mass spectrometry methods, the matrix-assisted laser desorption and ionization-time of flight mass spectrometry (MALDI-TOF) and surface-enhanced laser desorption and ionization mass spectrometry (SELDI-TOF).

In MALDI-TOF, prior to loading the proteomic sample into the mass spectrometer, it is mixed with an energy absorbing matrix (EAM). The mixture is then allowed to crystallize on a metal plate, often a stainless steel plate, and consequently loaded into the instrument.

Coming from this paradigm is a modification and variation of the matrix-assisted laser desorption and ionization- time of flight mass spectrometry (MALDI- TOF), referred to as the surface-enhanced laser desorption and ionization mass spectrometry (SELDI-TOF) MALDI-TOF. In SELDI-TOF, there is extra chemistry on the targeted such that the mixture is spotted on a surface transformed with a chemical functionality such as binding affinity [11, 12]. This helps keep proteins from complex mixtures according to the specific properties of the proteins, as well as carefully and selectively binding to a subset of molecules from crude preparations [11, 12]. H50 (hydrophobic surface, similar to C6-C12 reverse phase chromatography), IMAC30 (metal-binding surface), and Q10 (strong anion exchanger) are some surfaces normally adopted, whiles surfaces can also be modified with antibodies, other proteins with proper binding properties, or even DNA [12]. In order to ensure that only oppositely charged proteins stick to the surface, some SELDI chips can be made using an electrically charged surface. The point of departure between SELDI-TOF and MALDI-TOF mass spectrometry ends after the sample is cleaved with a protease, and a matrix solution added. The procedure that follows is similar to the MALDI-TOF. A time-of-flight mass spectrometry (TOF-MS) is typically what is used to analyze samples spotted on a SELDI surface [12]. The metal plate is fixed in a vacuum chamber and the crystallized mixture obtained hit with pulses from a nitrogen laser [11]. The energy produced from the laser is consumed by the matrix crystallized molecules and transfers it to the proteins [11]. The proteins at this stage can be desorbed and ionized, creating ions in the gas phase. The process takes place in the presence of an electric field. The electric field accelerates the ions down into a flight tube. The ion then hits a detector that measures ions as they reach the end of the tube. Being sensitive to low-molecular-weight proteins, crude biological samples that are not apt for conventional MALDI-TOF, such as serum, can be added directly to the sample chip [12]. This has aided the progress made by SELDI-TOF mass spectrometry in cancer detection and diagnosis. Figure 1 show a simplied pictorial representation of the MALDI-TOF mass spectrometry.

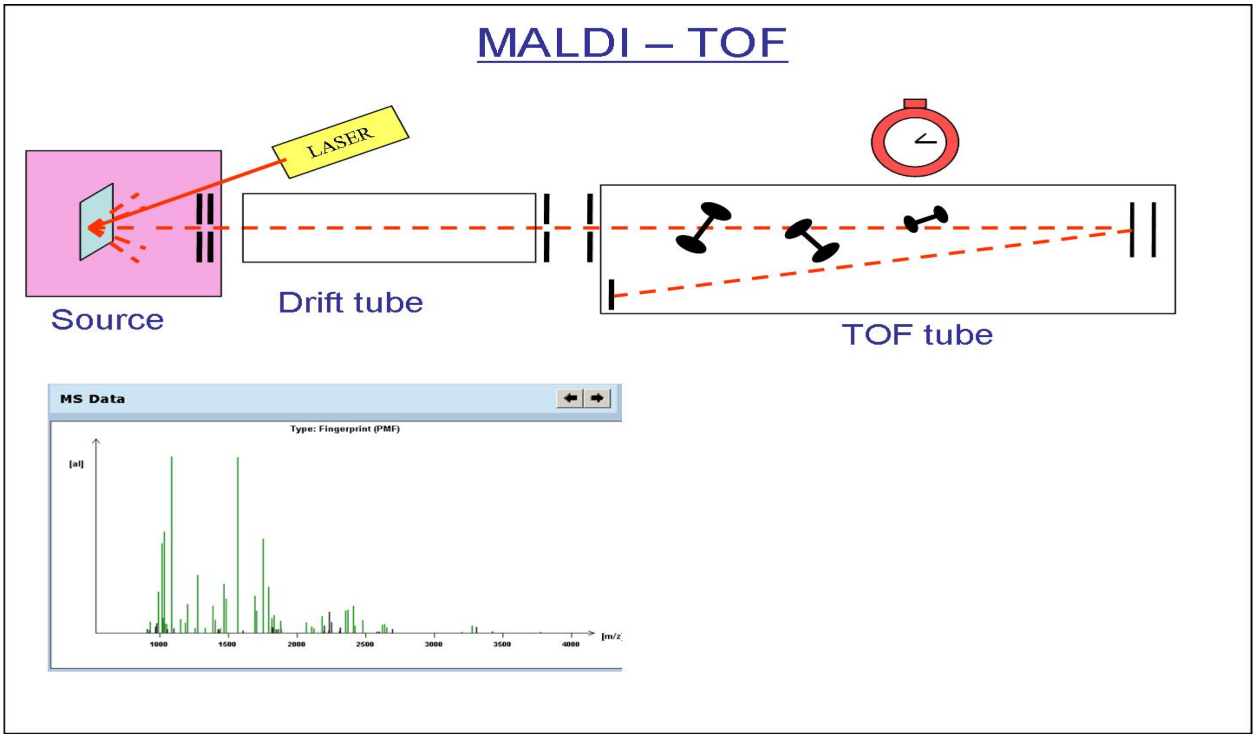

Figure 1: A representation of the MALDI-TOF Mass Spectrometry Method by Max Planck Institute For Plant Breeding Research (2003-2017) generated by Dr Thomas Colby.

2.4 Mass Spectrum

As earlier discussed, mass spectrometry as a vast analytical procedure yields charged particles or ions from the biomolecules or chemical substances to be analyzed. Various kinds of mass spectrometers use magnetic or electric fields to exert forces on charged particles and separates them based on their mass-to-charge ratios. The data and output from the mass spectrometry generates a vast sequence of value pairs. Each pair contains a measured intensity, depending on the amount of the biomolecule detected, and a mass-to-charge ratio, based on the molecular mass of detected biomolecules. A spectrum is the result of all these processes in a proteomic analysis; a plot of the mass-to-charge (m/z) values is on the horizontal index (x-axis) and relative abundances on the vertical index (y-axis), as shown in Figure 2. A base peak is the peak recorded with the highest intensity, usually set at 100% abundance in the spectrum, corresponding to the most abundant ion. The representation of the peptide in the current sample is by the peaks. A quadratic transformation is what aids the computation of the m/z values in the mass spectrometry [5, 11]. A small number of molecules (usually between 3 and 7) with known masses are used to generate a spectrum, which aids in establishing the coefficients for the quadratic transformation [5, 11]. Consequently, a count of peaks corresponding to the known masses in the spectrum is obtained. The method of least squares determines the coefficient, given the set of (time, mass) pairs [5, 11]. The final data spectrum is the line plot of pairs of intensities and m/z values resulting from calculations carried out in a preprocessing stage [5].

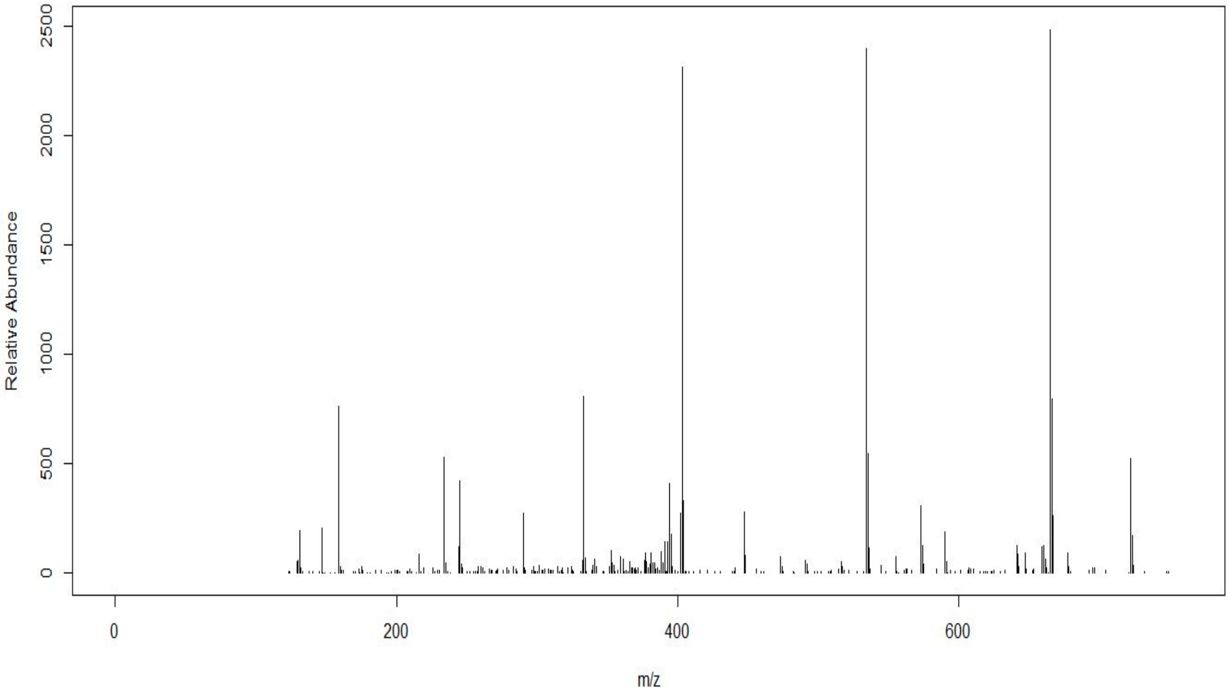

Figure 2: This figure shows a simplified representation of a mass spectrum as a line plot of pairs of intensities and m/z values.

2.5 Tandem Mass Spectrometry

Facilitating the identification of a considerable number of proteins and interspersed with fast speed, great ionization methods, high sensitivity and simple sample preparations, tandem mass spectrometry has emerged as a preferred method for high-throughput protein identification. Tandem mass spectrometry (MS/MS) is a two-stage mass spectrometry process with some form of fragmentation occurring in between the stages. The two stage of mass analysis allows the examination of individual ion fragmentation in a mixture of ions. Before analyzing with a mass spectrometer in the MS/MS experiment, a mixture of proteins is digested with an enzyme, thereby breaking the proteins into shorter peptides. The resulting peptide is then separated by liquid chromatography and converted into electrically charged particles. Various types of instruments and scan modes are employed in tandem mass spectrometry. The accomplishment of the multiple stage MS/MS experiment is by two key instrument types with individual mass spectrometer elements separated in space or employing a single mass spectrometer with the MS steps separated in time. The first category of instrument is an instrument in which two mass spectrometers are assembled in tandem with the separation elements physically separated and distinct. This is also referred to as tandem mass spectrometry in space or using a sequence of mass spectrometers in space, signifying the physical separation of the instrument components. The second type of instrument can be described as doing tandem mass spectrometry in time. The MS/MS instrument in this category comprises analyzers such as the quadrupole ion trap or Fourier transform ion cycloton resonance (FTICR) instrument that has the ability to store ions. Separation in this regard is attained with ions captured in the same place and multiple separation procedure taking place over time. The triple-quadrupole mass spectrometer, also referred to as QqQ, is the most generally used tandem mass spectrometer. The first and third quadrupoles in a triple-quadrupole mass spectrometry act as mass filters whiles the second quadrupole causes fragmentation of the analyte by allowing ions of any mass to pass through. Inferring from its name, the quadrupole mass analyzer, also known as the quadrupole mass filter, consists of four cylindrical rods that are fixed parallel to each other. In the context of mass spectrometry, the quadrupole plays the role of filtering sample ions, based on their m/z values. The product ion, precursor ion, and neutral loss scans are the three scan experiments commonly used in tandem mass spectrometry. As often is the case, in the product ions scan, ions of a particular m/z value are selected in the first mass spectrometer, transferred and analyzed in the second mass spectrometer after going through fragmentation. This results in the product ion spectrum. The process is more explained by the precursor ion first held in quadrupole 1 (MS1), undergoing a CID and fragmentation in Quadrupole 2 and scanning resulting in a spectrum of fragment, known as the product ion spectrum.

Collision-induced dissociation (CID), also known as collisionally activated dissociation (CAD), is used in tandem mass spectrometry to fragment a peptide and subsequently obtain a spectrum. It entails the collision of an ion with a neutral atom in the gas phase and successive separation of the ion. In a precursor ion scan, only ions with a given m/z value passes through the second mass spectrometry. Further elaborating the process, the m/z value of a particular product ion is held fixed at quadrupole 3 and quadrupole 1, scanned across the desired m/z range. A spectrum of precursor ions is then produced. The neutral loss scan employs a combination of scanning to identify precursor ions that fragment through a particular neutral loss amid the CID fragmentation. By this process, the first M/S scans all the masses and the second M/S scans at a particular offset different from the first mass spectrometry [7].

3 THE FUNDAMENTAL CONCEPTS OF FRAGMENTATION AND SPECTRUM ANALYSIS

The ability to meaningfully manage and interpret mass spectrometry data goes a long way in the maximization of true protein identification and the minimization of inaccurate identification. A fundamental goal in any protein identification method, in a quest to define a protein as already identified or novel, is to match an observed spectrum against a theoretical spectrum of the peptide in question [11].

An important technology in mass spectrometry-based proteomics is peptide sequencing. The sequencing of proteins began with the purification of a substantial amount and a technique known as Edman degradation was used, developed by Peer Edman (Epstein et al 1996) [13, 25]. This approach not only requires much expertise, but a considerable amount of sample is needed. The popularity generated by mass spectrometry, in terms of high sensitivity and efficiency, has supplanted Edman degradation as a technique in peptide sequencing. Protein analysis by mass spectrometry makes use of two main approaches: the top-down or bottom up approach [13]. The top-down method focuses directly on intact protein, and challenges such as difficulties in the separation of intact proteins, the complexity of the data and lack of automation limits the output of the top-down strategy [13]. The bottom-up approach, a fundamentally easier approach, focuses on peptides rather than intact proteins, generated from enzymatic digestion.

Proteins are complex molecules made up of strands of amino acids. The favorable consideration for peptides over proteins in the identification process is attributable to factors such as greater solubility of peptides, and the higher sensitivity of the mass spectrometer for smaller molecules such as peptides than for high mass proteins [13]. The difficulty of identifying intact protein leads to proteins being broken into short peptides and examined separately. The problem of identifying a protein is thus reduced to the problem of identifying peptides. The benefit primarily is higher sensitivity for smaller molecules by mass spectrometry and the better fragmentation behavior of peptides when compared with proteins. A very critical procedure in peptide sequencing is fragmentation in the mass spectrometer.

A peptide is formed when the amine and carboxylic acid functional groups in amino acids come together to form amine bonds in a chain of amino acid unit. A peptide is, therefore, a chain or sequence of amino acids, each of which is represented by one of 20 letters. That is, a peptide is a string over the 20 - letter alphabet of amino acids, with each amino acid assigned a non-negative molecular mass, measured in daltons (Da). A dalton is the standard unit used for denoting mass of an atomic or molecular scale. Table 1 shows a list of all 20 amino acids with their corresponding 3 letter and 1 letter codes. The single letter code is used throughout the thesis to simplify notation. Structurally, proteins and peptides are very similar, being made up of chains of amino acids that are held together by peptide bonds (also called amine bonds). The basic distinguishing factors are size and structure, with peptides being less well defined in structure than proteins, which can adopt complex conformations known as secondary, tertiary, and quaternary structures. Peptide fragmentation as pertaining to mass spectrometry, is triggered by collisions with residual gas where the bond breakage occurs through cleavage of the amine bonds [25]. The theoretical spectrum consists of m/z values and intensity of possibly occurring ions. Compounds, a molecule composed of atoms from different elements, can have the same molecular weight or m/z value but exhibit different chemical composition. However, breaking them into fragments allows peptides to be identified. Through an ionization process, the peptides are metamorphosed into electrically charged particles called ions. An ion is an electrically charged atom or group of atoms, formed by the loss or gain of one or more electrons. The peptide is broken into pairs of complementary fragment ions, with the most common ones being b and y ions. It is the intensities of the ions that are detected in the mass spectrometer and subsequently presented in a spectrum as a vertical line graph, representing ions having a specific mass-to-charge ratio (m/z) and relative abundance of the ion indicated by the length of the line. The theoretical spectrum of a peptide is a set of peaks with the position of each peak at the m/z value of each ion type. There are spikes at each peak location and zeros everywhere else [11]. Figure 3 shows the theoretical spectrum for the peptide *TGMSNVSK* using only the b and y ions. The b ions are represented by the solid lines and the dashed lines refers to the y ions.

Table 1: The 20 amino acids with their corresponding 3 letter and 1 letter codes.

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	Tryosine	Try	Y
Isoleucine	Ile	I	Valine	Val	V

Peptide fragment ions that emerge from tandem mass spectrometry are indicated by a specific notation [20]. The notation is assigned based on the fragment ion types that emerge by various bonds along the peptide backbone and side chain [27]. Peptide fragment ions are indicated by a, b or c if the charge is maintained on the N-terminus (also referred to as the amino-terminal fragment) and by x, y or z if the charge is retained on the C-terminus (also referred to as the carboxyl-terminal fragment). The process of generating the theoretical spectrum begins by splitting the true peptide sequence into all viable ion combinations. Albeit there are several ion types, the ability to correspond to cleavage of the amine bond, makes b and y ion the most useful sequence ion types, hence why they are used in practice. The b and y ions for a given peptide represents the two halves formed by splitting the original peptide between various amino acids. Figure 4 and Figure 5, portrays the two parts for the b and y ions respectively for the peptide $TGMSNVSK$ that is formed by splitting it between various amino acids.

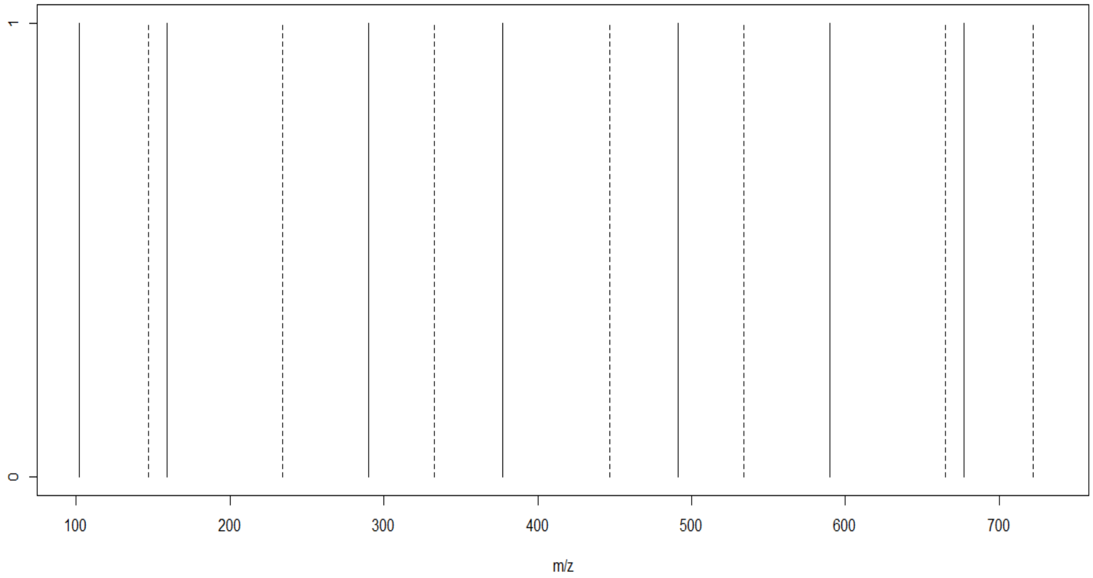

Figure 3: Theoretical spectrum for the peptide $TGMSNVSK$ using only b and y ions. The b ion is denoted by solid lines and the y ion is denoted by dashed lines. The presence of an ion is represented 1 and 0 represents the absence of an ion.

A b-ion is a fragment ion, that is the beginning of the peptide containing the N-terminus and is terminated by an amino acid with a free amine group $(-NH_2)$, resulting in the charge being retained by the amino-terminal fragment. Amines are organic compounds (hydrocarbons) whose functional group contains basic nitrogen atom with a single pair. Thus, an amine group is any group of organic compounds containing a nitrogen functionality. A y-ion extends from the C-terminus and is the complement of the b-ion. An ion is classified as y ion if the carboxyl-terminal fragment (C-terminus) retains the charge. The C-terminus also called carboxyl-terminus is the end of an amino acid, terminated by a free carboxyl group $(-COOH)$. A carboxyl group also referred to as the carboxy group and symbolized as $COOH$ is an organic functional group, having a carbonyl and hydroxyl group linked to a carbon atom.

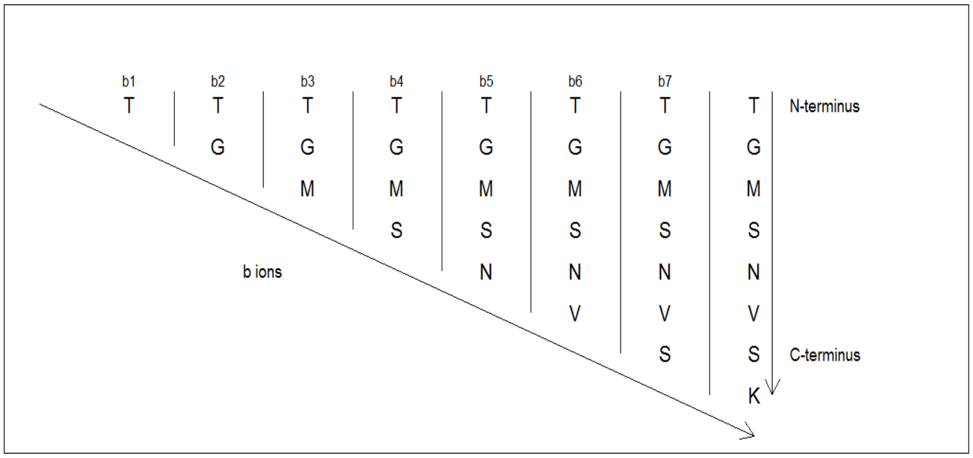

Figure 4: Illustration of the fragmentation for *b* ions.

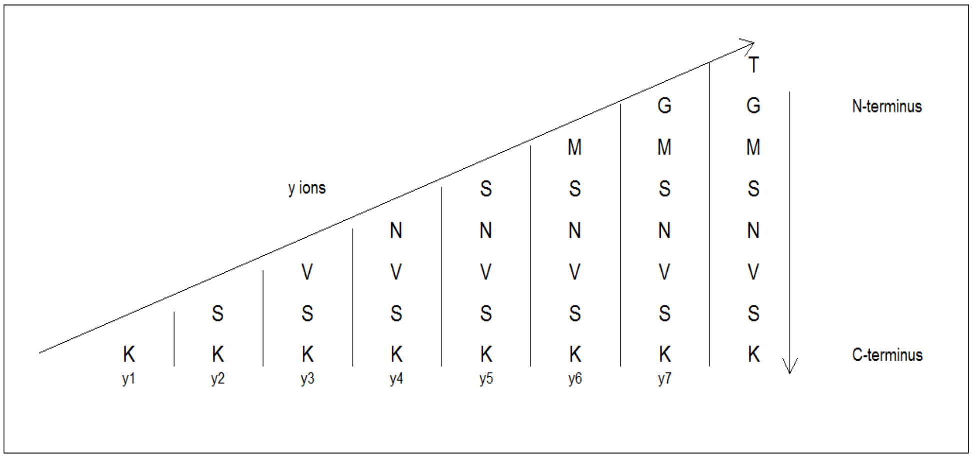

Figure 5: Illustration of the fragmentation for *y* ions.

On the other hand, a carbonyl group is a carbon double-bonded to an oxygen and a hydroxyl as an OH group, composed of a hydrogen atom covalently bonded to an oxygen atom. In effect, the carboxyl group has both a carboxyl and a hydroxyl group attached to the same carbon atom.

After the peptides are broken down into fragmented ions, the mass of each ion is determined. The mass for any given ion is found by adding the parent mass of the peptide to an offset value. Thus, for any given ion, the mass is found by:

$$\sum_{i=1}^{k} m(p_i) + \delta_\ell$$

where k is the number of amino acids in the ion sequence, p_i is the amino acid in the i th position, $m(p_i)$ is the mass of the amino acid in the i th position, ℓ denotes the type of ion such that $\ell \in (b, y)$, and δ_ℓ is the offset for ion type ℓ. Table 2 shows a list of all twenty amino acids together with corresponding mass in daltons (Da).

Table 2: The 20 amino acids with their corresponding masses in daltons.

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

Peptide fragmentation in tandem mass spectrometry is determined by offsets that correspond to the peaks and represents the ion types produced by a given mass

Table 3: Information about ion types. Here 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 -17.05)
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+ 2.3)/2
y-$H_2O - NH_3$	C	-16.10	(M-16.10)
y-$H_2O - H_2O$	C	-17.15	(M-17.15)

spectrometer. That is, the offsets match up to be the peaks in a given spectrum, and thus denote the different ion types created in the given mass spectrometer (Dančik et al.,1999). The offsets, determined by Dančik et al (1999), are the result of either N-or C-terminal cleavage. Table 3 lists the different ion types with their terminus position, the offset value, and calculation of the mass of the ion.

Dančik et al. (1999) developed an offset frequency function to describe ion type tendencies for specific mass spectrometers. To enable software to accurately analyze spectra obtained from any type of mass spectrometer, the offset frequency function was introduced. The usage of the offset frequency function to ascertain the ion-types specific to a mass spectrometer is useful in determining the ordering of amino acids in a fragment sequence.

Consider the peptide *TGMSNVSK*, as an example for illustrating the theoretical spectrum. There are seven b ions and seven y ions that we generate by splitting the peptide. The ions we obtain from the b ions are as follows: *T*, *TG*, *TGM*, *TGMS*, *TGMSN*, *TGMSNV*, and *TGMSNVS*. The first b ion, *T*, has a mass of $101.048 + 0.85 = 101.898$ Da. The second b ion, *TG*, has a mass of $101.048 + 57.0215 + 0.85 = 158.9195$ Da. Continuing with the process, we obtain the following additional b ions: *TGM*, *TGMS*, *TGMSN*, *TGMSNV*, and *TGMSNVS* with masses 289.9595, 376.9915, 491.0345, 590.1029, and 677.1349 daltons, respectively. Refer back to Figure 4 for the illustration of the splitting of the b ions and the seven ions generated.

Similarly, for the y ions, we obtain the following ions: *K*, *SK*, *VSK*, *NVSK*, *SNVSK*, *MSNVSK*, and *GMSNVSK*. The first y ion, *K*, has a mass of $128.095 + 18.85 = 146.9450$ Da. The second y ion, *SK* has a mass of $87.032 + 128.095 + 18.85 = 233.9770$ Da. Continuing with the splitting, we obtain the following additional ions : *VSK*, *NVSK*, *SNVSK*, *MSNVSK*, and *GMSNVSK* with masses 333.0454, 447.0884, 534.1204, 665.1604, and 722.1819 daltons, respectively. Refer back to

Figure 8 for the illustration of the splitting of the y ions and the seven ions generated. In summary, the theoretical spectrum for the peptide *TGMSNVSK* is the set of masses 101.8980, 158.9195, 289.9595, 376.9915, 491.0345, 590.1029, 677.134, 722.1819, 665.1604, 534.1204, 447.0884, 333.0454, 233.9770, 146.9450 daltons.

4 A BAYESIAN MODEL

We focus on refining a Bayesian model in an earlier work by Lewis (2013), a model that seeks to identify the true peptide based on the observed spectrum [11]. The Bayesian model in identifying this true peptide uses a Markov chain Monte Carlo (MCMC) algorithm to simulate candidate peptide sequences from the posterior distribution [11].

4.1 Pre-Processing of Data

Data produced by mass spectrometry is extremely large, depicting the abundance (intensity) of biomolecules showing certain mass-to-charge ratio (m/z) values [14]. Generally, the broad processes involved in mass spectrometry-based proteomics experiments consist of a data generation phase, data preprocessing and a phase for analyzing data [14]. Preceding spectra data analysis is the data preprocessing phase, a process that removes or curtails problems with data by way of spectrum noise and contaminant clean up. Coming from a process of sample preparation, sample ionization and activities with the instrument itself, it is not surprising that mass spectrometry data are quite noisy [14]. As a result, the observed spectrum first needs to be thresholded [14]. In the process, peaks with intensity values below a threshold will be removed, and emphasis given to m/z values having intensities above the threshold [11]. The data employed in our model consist of the retained intensity values and their corresponding m/z values [14]. Each integer m/z value is assigned a distinct threshold value, which is computed and denoted by $T = (T_1, T_2, ..., T_{q^*})$. q^* represents the total number of m/z values [11]. In thresholding, both a constant and a moving threshold are calculated. [11]. This has become necessary because the mass spectrometer does not always capture all peaks at the beginning and the end of the spectrum [11]. Focusing on only a constant threshold has the ability to eliminate peaks that are truly signal peaks and not noise peaks. Thresholding is therefore done using a combination of a constant and moving thresholds as a weighted average of the thresholds [11].

4.2 Likelihood for the Bayesian Model

For our Bayesian model, we work at improving a likelihood function in earlier work by Lewis (2013) that incorporates additional ions to the model. The likelihood function we specify, and employ in our model, gives a measure of how well the observed spectrum and theoretical spectrum agree [11]. As a cardinal part of any Bayesian inference, we establish our parameters and models. There is an overall goodness of fit measure, which penalizes any candidate peptide whose theoretical spectrum does not align well with the observed spectrum [11]. However, a candidate peptide is rewarded by this overall goodness of fit measure if the theoretical spectrum aligns nicely with the observed spectrum [11]. As discussed earlier, data generated by mass spectrometers are plagued by noise, owing to various factors therefore, we do not discount the possibility of noise peaks in the data set even after thresholding. This has necessitated the incorporation of an overall goodness of fit measure tasked to penalize a candidate peptide displaying too many noise peaks near the theoretical spectrum [11]. Signal peaks in a mass spectrometer appear as a local maxima in the spectrum. However, not every signal peak is captured by the mass spectrometer.

The challenge is amplified when a peptide with low abundance is concealed by noise, resulting in increasing false rate of peak detection. To ameliorate this trend, we include an indicator function in our likelihood function that represents the presence or absence of a peak [11].

4.3 Explaining the Likelihood Function

The original likelihood function from the work of Lewis (2013) is comprised of the b and y ions and is defined as:

$$L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\eta}, \kappa_1, \kappa_2) \propto \kappa_1^s \exp(-\kappa_1 S_1) \kappa_2^{t-s} \exp(-\kappa_2 S_2). \tag{1}$$

The size of the likelihood function is driven by two overall goodness of fit measures S_1 and S_2.

$$S_1 = \sum_{i=1}^{p} \lambda_i^b \; \min_j d(x_j - \tau_i^b) + \lambda_i^y \; \min_j d(x_j - \tau_i^y) \tag{2}$$

$$S_2 = \sum_{j=1}^{t} \min_{i,k} |x_j - \tau_i^k| \tag{3}$$

with $d(x_j - \tau_i^k) = \min\{|x_j - \tau_i^k|, \delta\}$. The interpretation of our parameters will be explained in our refined one.

The proportion of detected ions was calculated for each of the ions in Table 3 for all $1,206$ peptides from our data. We assume the ion is present if the observed m/z value is within 0.5 Da of the theoretical m/z for that given peptide. As expected, the b and y ions had the highest proportion ions. We found that $b - H_2O$ was the next ion with the highest percentage for this data. Similar results were found in Dančik et al (1999). Figure 6 provides the histogram, the mean, and five number summary for the proportion of detected ions for the b, y, and $b - H_2O$ ions. Based on this work, we chose to add the $b - H_2O$ ion into the likelihood.

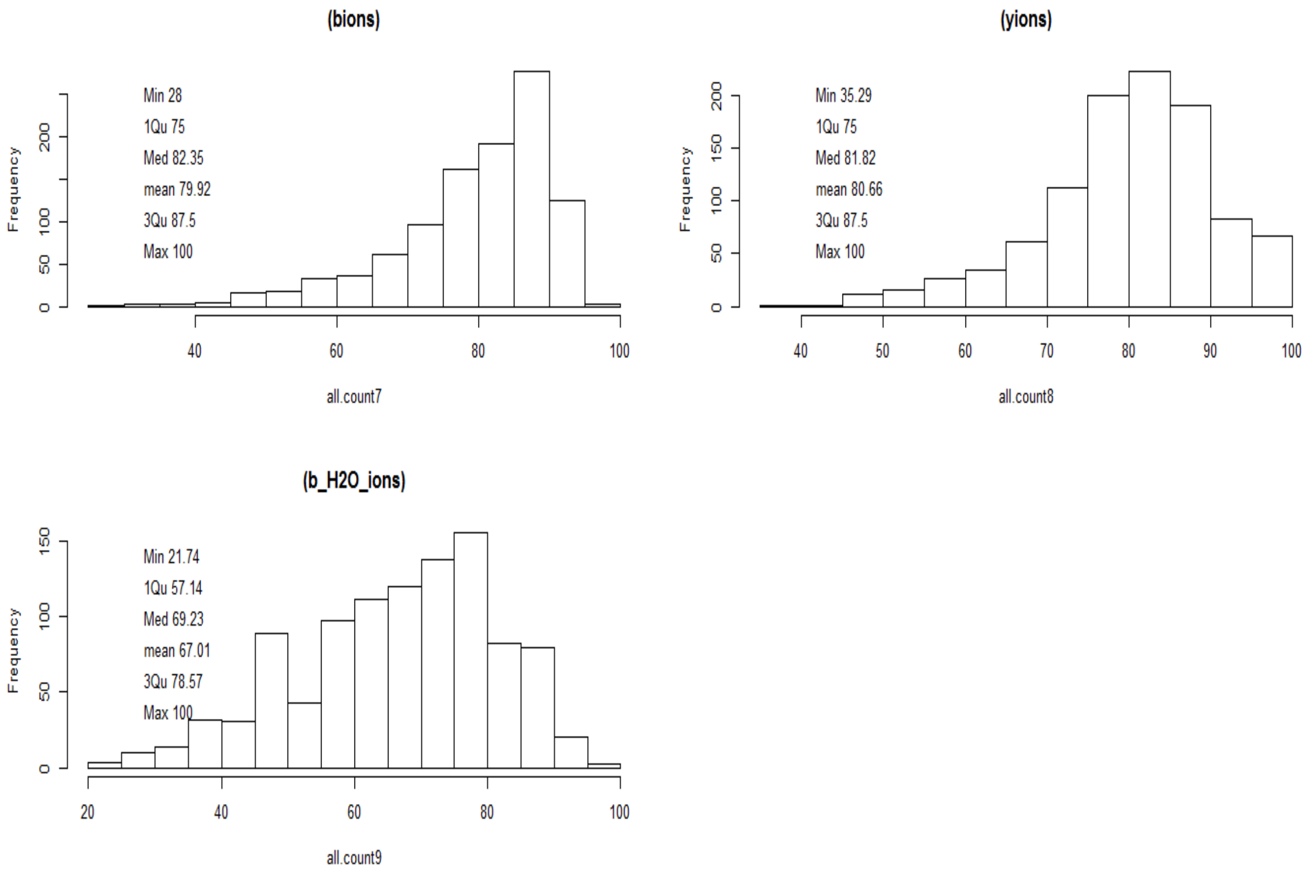

Figure 6: Numerical summary and plot for b, y and $b - H_20$ ions of the proportion of the detected ions.

The revised likelihood function with the additional $b - H_2O$ ion has the form

$$L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\eta}, \kappa_1, \kappa_2) \propto \kappa_1^s \exp(-\kappa_1 S_1) \kappa_2^{t-s} \exp(-\kappa_2 S_2) \tag{4}$$

where our parameter vector $\boldsymbol{\theta} = (\tau_1^b, \ldots, \tau_p^b, \lambda_1^b, \ldots, \lambda_p^b, \tau_1^y, \ldots, \tau_p^y, \lambda_1^y, \ldots, \lambda_p^y,$ $\tau_1^{b-H}, \ldots, \tau_p^{b-H}, \lambda_1^{b-H}, \ldots, \lambda_p^{b-H})$, $\boldsymbol{\eta}$ signifies string of amino acids for the candidate peptide and $\boldsymbol{X}$ is the observed pairs of m/z values and intensities for a particular spectrum [11]. Here t represents the total number of m/z values after thresholding and s is the total number of b, y and $b - H_2O$ ions combined. The notations b, y and $b - H$ as used in this context denotes b ions, y ions and $b - H_2O$ ions respectively. Let x_j be the observed m/z values with peaks that are above some specific threshold $\boldsymbol{T}$, $j = 1, ..., n$. We write $\mathcal{N}$ to denote the set of observed noise peaks, where for label $j \in \mathcal{N}$ we have $|x_j - \tau_i^k| \geq \delta$ for all i, k, or there exists a $j' \neq j$ such that $|x_j' - \tau_i^k| < |x_j - \tau_i^k| < \delta$ for some i, k, where i= 1,..., p and $k \in \{b, y, b - H\}$ [11]. We take $\delta = 3$ Da. The goodness of fit measures of the candidate spectrum to the observed spectrum that drives the size of the likelihood function are defined as:

$$S_1 = \sum_{i=1}^{p} \lambda_i^b \ \min_j d(x_j - \tau_i^b) + \lambda_i^y \ \min_j d(x_j - \tau_i^y) + \lambda_i^{b-H} \ \min_j d(x_j - \tau_i^{b-H}) (5)$$

$$S_2 = \sum_{j=1}^{t} \min_{i,k} |x_j - \tau_i^k| \tag{6}$$

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

As previously discussed, not every peak in the theoretical spectrum is captured by the mass spectrometer, due to different factors, mentioned in section 4.1. This phenomenon has the potential of increasing the value of S_1 if the peak is missing, as the m/z value for the next closest peak may be far from the candidate peak and thus increasing the likelihood value, which could cause the model to incorrectly estimate the true peptide [11]. In order to forestall this from throwing off our model, a penalty term, $\delta = 3$ Da is incorporated to mitigate the effect of the missing peak [11].

Moreover, τ_i^b, τ_i^y, and τ_i^{b-H} are the m/z values for the b, y and $b - H_2O$ ions respectively, of the candidate peptide and κ_1 and κ_2 signify weights [11]. Here κ_1 follows a Gamma prior distribution with a shape parameter a_1 and a scale parameter b_1, while κ_2 follows a Gamma prior distribution with a shape parameter a_2 and a scale parameter b_2. We also have λ_i^b, λ_i^y and $\lambda_i^{b-H} \in \{0, 1\}$ as indicator functions that signify whether the ith b, y and $b - H_2O$ has a corresponding observed peak, where $i = 1, ..., p$. Hence, $\lambda_i^b = 1$ denotes the presence and $\lambda_i^b = 0$ denotes the absence of a b-ion at position i, $\lambda_i^y = 1$ denotes the presence and $\lambda_i^y = 0$ denotes the absence of a y-ion at position i [11]. Similarly, $\lambda_i^{b-H} = 1$ denotes the presence and $\lambda_i^{b-H} = 0$ denotes the absence of a $b - H_2O$ ion at position i. S_1 separately fixes each b-ion, y-ion and $b - H_2O$ ion of the candidate peptide and measures the closeness of the nearest peak to it. That is, S_1 measures the sum of the minimum absolute distances between the closest observed m/z above a threshold and each m/z peak value of the candidate peak [11]. S_2 fixes each observed peak and measures the closeness of the nearest "candidate peak" to it. More formally, S_2 measures the closeness of the nearest candidate peak to each observed peak. S_1 and S_2 are simply two reasonable measures of closeness, and our likelihood combines them (weighting them by κ_1 and κ_2) [11]. It stands to reason that, the closeness of the candidate peaks to the observed peaks makes S_1 small. Similarly, the presence of fewer noise peaks or closeness of the noise peaks to the candidate peak makes S_2 small [11].

4.4 Priors

As earlier discussed, the collision-induced dissociation (CID), also referred to as collisional activated dissociation (CAD), entails the collision of an ion with a neutral atom in the gas phase and successive separation of the ion. Thus, the fragmentation process is very much reinforced by CID during the gas phase. The b and y ions have the ability to correspond to cleavage of the amine bond among several ions, making them the most useful sequence ion types. The inspiration about cleavages in the amino acid pair and cleavage pair abundance, comes from Huang et al [9], who estimated the average bond cleavage abundance for each amino acid pair for both the b and y ions for gas-phase dissociation spectra [9, 11]. It is during the CID that a peptide bond fragments, resulting in cleavage into distinct fragments with a cleavage pair such as the b and y ion pair present in the peptide. Consider the peptide $TGMSNVSK$, for example. Recall from Chapter 3, that one of the seven b ions of the peptide $TGMSNVSK$ is TG, and thus the complement $MSNVSK$ as the y ion. The cleavage between the amino acids G and M results in these complementary ions. To develop prior information to identify the true peptide, we are inspired with information from Huang et. al (2004) concerning cleavage pair abundance and when to expect cleavages in the amino acid pairs.

4.5 Cleavage Prior

From the prior information, we obtain prior probabilities of "seeing" the b-ion peak and y-ion corresponding to each complementary pair of ions in a peptide. The cleavage prior used in our posterior is from the work by Lewis (2013).

Let $\boldsymbol{\lambda} = (\lambda_1^b, \ldots, \lambda_p^b, \lambda_1^y, \ldots, \lambda_p^y)$ be a vector of binary indicators of whether a peak occurs in the spectrum at the m/z values corresponding to each b- and y-ion. We define a cleavage pair prior as (here p is the number of cleavage pairs):

$$\pi(\boldsymbol{\lambda}) = \prod_{i=1}^{p} P(\lambda_i^b, \lambda_i^y)$$

with

$$\begin{aligned} P(\lambda_i^b = \lambda_i^y = 1) &= \rho_i^{by} \times \gamma_i \times \beta_i \\ P(\lambda_i^b = 1, \lambda_i^y = 0) &= \rho_i^{by} \times (1 - \gamma_i) \times \beta_i \\ P(\lambda_i^b = 0, \lambda_i^y = 1) &= \rho_i^{by} \times \gamma_i \times (1 - \beta_i) \\ P(\lambda_i^b = \lambda_i^y = 0) &= 1 - \rho_i^{by} + [\rho_i^{by} \times (1 - \gamma_i) \times (1 - \beta_i)] \end{aligned}$$

where ρ_i^{by} is the geometric mean of the average relative abundance of bond cleavages of b and y ions for a particular amino acid pair for $i = 1, ..., p$ derived from Huang et al.(2004), γ_i is the probability of the presence of a y ion, and β_i is the probability of the presence of a b ion [11]. Further details about the cleavage prior can be found in the work by Lewis (2013).

4.6 Sequence Prior

There is also a sequence prior from the work by Lewis (2013). This is a prior distribution for a particular sequence (string) of amino acids in a peptide [11]. The string or sequence prior, $\pi(\boldsymbol{\eta})$ which represents the probability of any particular amino acid sequence, computes the probability of a sequence of amino acids appearing consecutively in a peptide sequence [11].

The sequence prior is defined as

$$\pi(\boldsymbol{\eta}) = \sqrt{\pi(\boldsymbol{\eta}_F) \times \pi(\boldsymbol{\eta}_R)} \tag{7}$$

Here $\boldsymbol{\eta}$ is the ordered sequence of the amino acids in the current peptide under consideration, $\pi(\boldsymbol{\eta})$ is a probability for this particular sequence, $\pi(\boldsymbol{\eta}_F)$ is the joint probability of any particular amino acid sequence calculated from left to right, $\pi(\boldsymbol{\eta}_R)$ is the joint probability of any particular amino acid sequence calculated in the reverse direction [11]. In-depth information about the sequence prior can be found in the work by Lewis (2013).

4.7 Prior for κ_1, κ_2

Our previous discussion of the likelihood function has κ_1 having a Gamma prior distribution with a shape parameter a_1 and a scale parameter b_1 and κ_2 follows a Gamma prior distribution with a shape parameter a_2 and a scale parameter b_2 [11]. Thus, our concentration parameters, κ_1 and κ_2, are estimated to have independent Gamma (a_1, b_1) and Gamma ((a_2, b_2) prior distributions respectively [11]. These are independent of the other parameters.

4.8 Posterior Distribution

From Bayes' Theorem, the posterior density can be written as

$$\begin{aligned}\pi(\boldsymbol{\eta},\, \boldsymbol{\omega},\, \boldsymbol{\lambda},\, \kappa_1, \kappa_2 | \boldsymbol{X}) &\propto L(\boldsymbol{X}|\, \boldsymbol{\omega},\, \boldsymbol{\tau},\, \boldsymbol{\eta},\, \kappa_1, \kappa_2) \times \pi(\boldsymbol{\lambda}) \times \pi(\boldsymbol{\eta}, \boldsymbol{\tau}) \times \pi(\kappa_1, \kappa_2) \\ &= L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\eta}, \kappa_1, \kappa_2) \times \pi(\boldsymbol{\lambda}) \times \pi(\boldsymbol{\eta}) \times \pi(\kappa_1, \kappa_2).\end{aligned}$$

where $\boldsymbol{\lambda}$, $\boldsymbol{\eta}$, and κ_1, κ_2 are assumed independent. The set of m/z locations given by $\boldsymbol{\tau} = (\tau_1^b, \ldots, \tau_p^b, \tau_1^y, \ldots, \tau_p^y, \tau_1^{b-H}, \ldots, \tau_p^{b-H})^T$ are determined by the sequence $\boldsymbol{\eta}$, which signifies the string of amino acids, and as such $P(\boldsymbol{\tau}|\boldsymbol{\eta}) = 1$. Here $\boldsymbol{\omega}$ is defined as $(\lambda_1^b, \ldots, \lambda_p^b, \lambda_1^y, \ldots, \lambda_p^y, \lambda_1^{b-H}, \ldots, \lambda_p^{b-H})$. It is worth stressing that the posterior density is only known up to a constant and the actual form of the posterior density is quite complicated [11].

4.9 A Markov Chain Monte Carlo Algorithm (MCMC)

The complexity of our posterior density and the fact that it does not represent a known distribution leads us to use a Markov chain Monte Carlo algorithm (MCMC) method to sample the parameters [1, 23, 24, 26].

A starting peptide for the MCMC is found by taking a random sample from the results of PepNovo. Since the mass spectrometer yields the overall weight for the corresponding observed spectrum, we only consider candidates with the correct mass (within a tolerance) [11].

Once we have a reasonable "initial candidate" peptide, we evaluate how closely it matches the observed spectrum by calculating the posterior probability of that candidate [11]. Then we propose a new candidate peptide and use the Metropolis-Hastings algorithm to decide whether to "accept the move" to the new candidate peptide, or whether to "reject the move" and retain the current candidate peptide. This produces a Markov chain in which each state is a candidate peptide (a sequence of letters). The initial peptide is called the current peptide and denoted $\boldsymbol{\eta}_{curr}$. The $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ vectors are pre-determined and remain constant throughout the algorithm. The vector $\boldsymbol{\lambda}_{curr}$ is generated using the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ vectors.

In summary, the algorithm starts by generating a new candidate peptide by randomly replacing one, two, or three amino acids of the current peptide with one, two, or three amino acids while still ensuring the total weight is within a tolerance of 0.5 of the true weight [11]. The posterior probability for both the new and current peptide is calculated and denoted as ζ_1 and ζ_2, respectively [11]. As part of the MCMC procedure, proposal densities are calculated for our priors. The proposal densities are $q(\lambda_{curr}|\lambda_{new})$, $q(\lambda_{new}|\lambda_{curr})$, $q(\boldsymbol{\eta}_{curr}|\boldsymbol{\eta}_{new})$, and $q(\boldsymbol{\eta}_{new}|\boldsymbol{\eta}_{curr})$ [11]. We then generate a value from a standard uniform distribution, $U \sim U(0,1)$. If $U < \left(\frac{\zeta_1}{\zeta_2} \times \frac{q(\lambda_{curr}|\lambda_{new})}{q(\lambda_{new}|\lambda_{curr})} \times \frac{q(\boldsymbol{\eta}_{curr}|\boldsymbol{\eta}_{new})}{q(\boldsymbol{\eta}_{new}|\boldsymbol{\eta}_{curr})}\right)$, then the new peptide becomes the current peptide, and $\boldsymbol{\lambda}_{new}$ becomes $\boldsymbol{\lambda}_{curr}$. Otherwise, both the current peptide and $\boldsymbol{\lambda}_{curr}$ remain unchanged [11] and the algorithm starts again [11].

After a large number of iterations, the algorithm stops. Whichever peptide (among those searched) that has the largest posterior probability is our estimate for the true peptide [11]. To ensure that our chain is irreducible, every 1000 steps we propose a completely random peptide (regardless of the current peptide) [11]. The state space is assumed finite, because for any given spectrum, the peptide cannot be arbitrarily long. Since our algorithm uses only the m/z values that have intensities above a threshold, we will instead generate a spectrum with signal and noise peaks that are already assumed to be above a threshold. For a given peptide, we will know the locations of the true peaks. In-depth discussion of this algorithm can be found in the work by Lewis (2013).

4.10 Simulated Annealing

To aid our search of the true peptide and optimization of our posterior density from a large parameter space, simulated annealing, a probabilistic technique, is explored. The idea fundamentally is designed similar to the physical process of heating a substance and lessening the defect by reducing the temperature. The procedure consolidates our model with a temperature parameter, allowing for more exploration at high temperature values and restraining of the exploration at lower temperature. Our likelihood function now becomes :

$$L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\eta}, \kappa_1, \kappa_2)^{1/T} \propto \kappa_1^s \exp(-\kappa_1 S_1)\kappa_2^{t-s} \exp(-\kappa_2 S_2)$$

from our old likelihood function of :

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

where T is the temperature parameter. We set a temperature value for our large temperature at 500 catering for the first 95% of iterations. The last 5% of iterations caters for our small temperature parameter and this is set to 1, ensuring we are back to our likelihood function.

5 RESULTS

Our motivating data set is from the Pacific Northwest National Laboratory (PNNL) which can be accessed publicly online [2]. For these data, the true peptide is known for each spectrum [11]. Consisting of $1,206$ peptides, the dataset has peptides with lengths ranging from 7 to 31 amino acids with an average length of 15.16 [11]. The total mass for each peptide is given along with the set of intensities and m/z values. Proceeding our spectra data analysis is the data preprocessing phase [11]. We remove the doubly charged parent ion from the dataset and use a threshold of 75% to remove "noise peaks" [11]. We set our tolerance level to be 0.5 Da. As indicated earlier, κ_1 and κ_2 follows a Gamma prior distribution and signifies the weights parameters with s being the number of b, y , $b-H_2O$ as used for our model [11]. The constants for this distribution is set to be $a_1 = 5.5$, $b_1 = 0.1$ and $a_2 = 3$, $b_2 = 100$. We set the initial components of $pb_1 = 0.05$ and $py_1 = 0.10$. These probabilities are fixed low because the mass spectrometer seldomly captures the first b and y ion [11]. We set all other pb_i and py_i to be equal to 0.80 for $i = 2, ..., p$. Thus, presence or absence probability vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are set to be $(0.05, 0.80, \ldots, 0.80)$ and $(0.10, 0.80, \ldots, 0.80)$. Going by these percentages, there is a 5% chance that we will see the first b ion and 80% chance of seeing the other b ions. The parameters set are the same used by Lewis (2013). After extensive numerical simulation and search, the best parameters for our new ion, $b - H_2O$ were determined. We set the initial components of $pb - H_1 = 0.05$

and $pb - H_2 = pb - H_3 = 0.10$. All other $pb - H_i$ are set to be equal to 0.60 for $i = 4, ...p$. As with the b and y ions, the first $b - H_2O$ ion is set to be low because the mass spectrometer rarely captures the first ion. Because the $b - H_2O$ ions tend to have low intensity values near the beginning of the data, they are more likely to be considered noise and will be removed in the pre-processing step. Therefore, it will appear that the peak is missing and the likelihood will be penalized for the missing peak. Hence, we set probability for the next two $b - H_2O$ ions to be low as well. We observe from the beginning that because these intensities are low, it is more likely to be considered noise.

Using the same parameter values as Lewis (2013), we set the large temperature to be 500 catering for the first 95% of iterations [11]. The small temperature parameter is set to 1 for the last 5% of iterations [11].

5.1 Example 1

Figure 7 is a plot of the observed versus the theoretical spectrum including only the b and y ions for the peptide *TGMSNVSK* and the Figure 8 is a plot of the observed versus the theoretical spectrum including only the b, y, and $b - H_2O$ for the peptide TGMSNVSK. In both plots, one can see there is noise in the center of the spectrum but over all the theoretical and observed spectrum align fairly well. One can see from Figure 8, by incorporating the additional ion, $b - H_2O$, the observed spectrum is better aligned with the theoretical spectrum because what was once identified as noise is now correctly being identified as a $b - H_2O$.

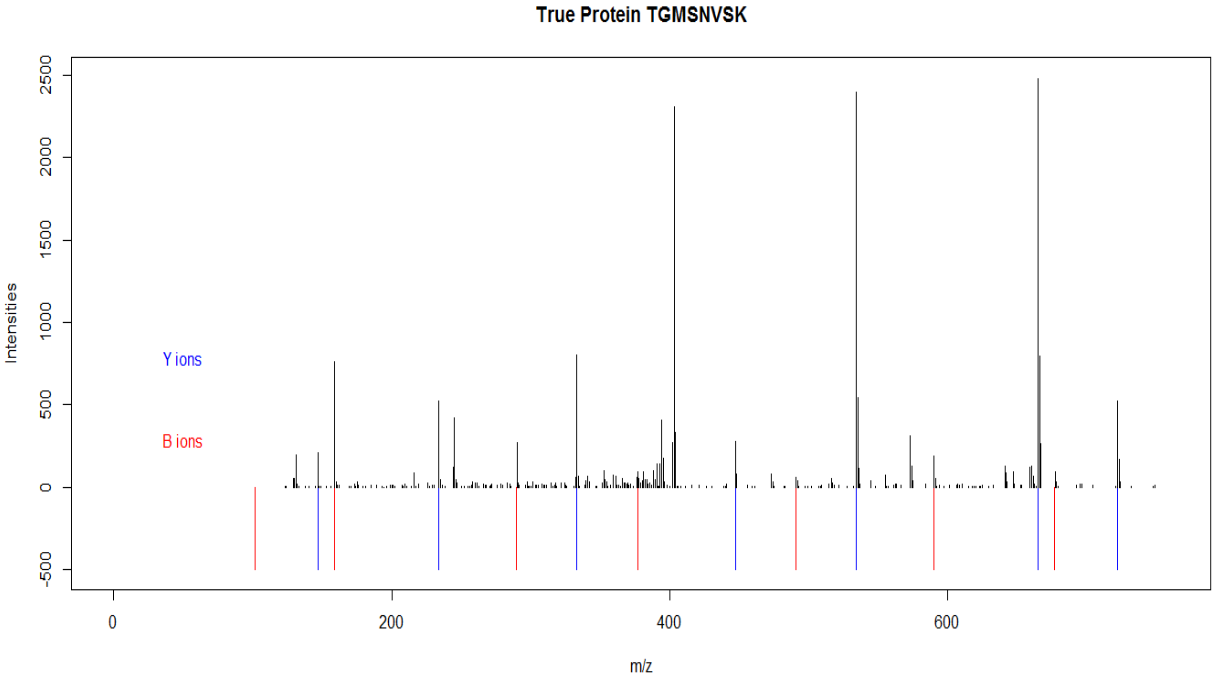

Figure 7: The observed spectrum plotted against the theoretical spectrum for the peptide *TGMSNVSK*, showing only two ions used in our model: b and y ions. Plotted below the zero axis is the theoretical spectrum and the observed spectrum is plotted above the zero axis.

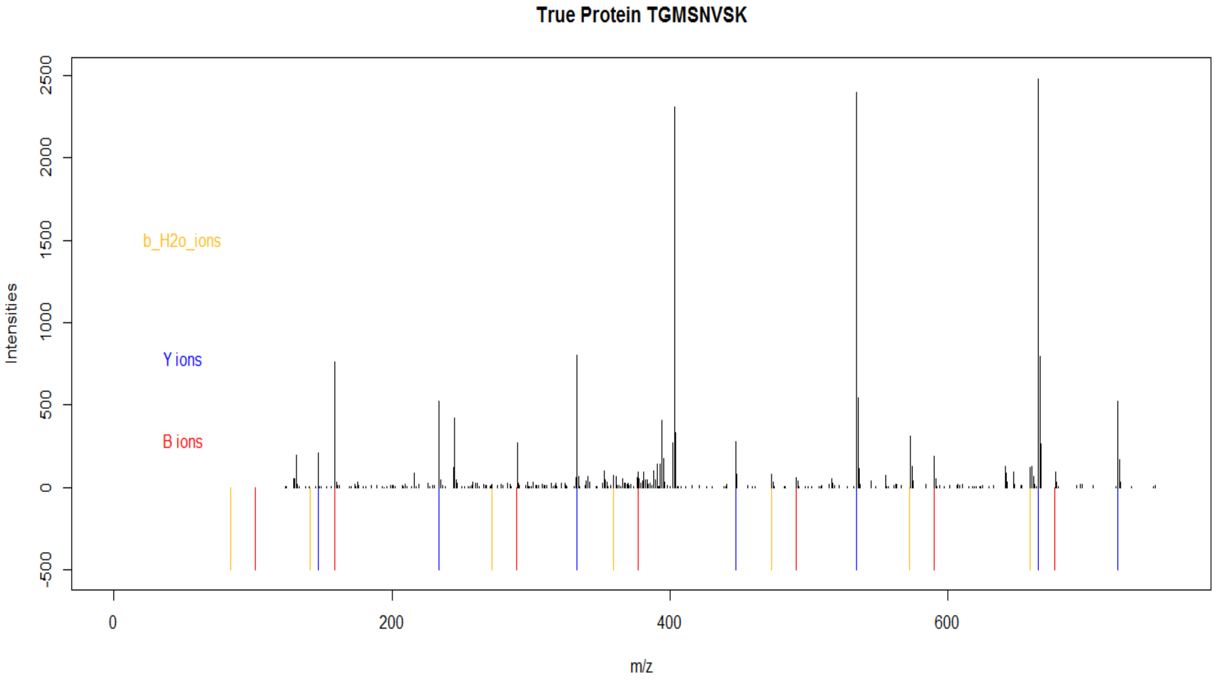

Figure 8: The observed spectrum plotted against the theoretical spectrum for the peptide *TGMSNVSK*, showing the three ions used in our model: b ions, y ions and $b - H_2O$ used in our model. Plotted below the zero axis is the theoretical spectrum and the observed spectrum is plotted above the zero axis.

Table 5.1: The top estimated peptides from the *MCMC* algorithm along with their corresponding log posterior densities and log likelihood for the peptide *TGMSNVSK*. The true peptide is in bold.

Peptide	**Log Posterior**	**Log Likelihood**
TGMSNVSK	35.21	63.30
TGMDSVSK	14.17	46.46
ASMSGGGEQ	7.26	37.22
ASEFGVSK	2.40	31.35
KCSRGSK	-0.07	32.62
GTMSGGVTN	-1.73	27.01
ASMSGKDK	-3.77	26.19
SAAFNWK	-3.99	27.04

A starting peptide, *CQSNDAK* was obtained from the results of PepNovo, which has a total mass of within 0.5 Da of the weight of the true peptide. Our best estimate for the true peptide is *TGMSNVSK* after $200,000$ iterations with a log posterior density of 35.21 (up to a constant). The top estimated peptides for this example along with their corresponding log likelihood value is shown in Table 5.1. We see the true peptide is estimated as having the largest log posterior density.

5.2 Example 2

Figure 9 shows the plot of the observed spectrum for the peptide *DLVESAPAALK*. Figure 10 is a plot of the observed versus the theoretical spectrum including the b, y, and $b - H_2O$ ions for the peptide *DLVESAPAALK*. In both plots, the theoretical and observed spectrum align fairly well. However, there is noise in the center of the spectrum. One can see from Figure 10, by incorporating the additional ion, $b - H_2O$, the observed spectrum is better aligned with the theoretical spectrum because what was once identified as noise is now correctly being identified as a $b - H_2O$.

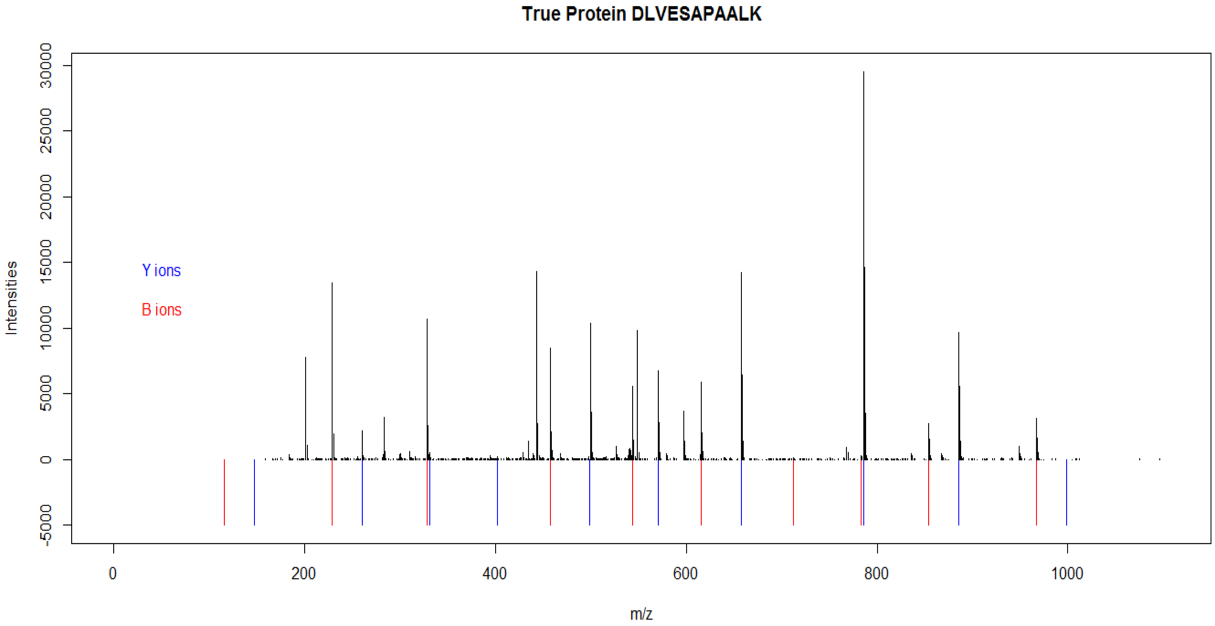

Figure 9: The observed spectrum plotted against the theoretical spectrum for the peptide *DLVESAPAALK*, showing only two ions used in our model: b and y ions. Plotted below the zero axis is the theoretical spectrum and the observed spectrum is plotted above the zero axis.

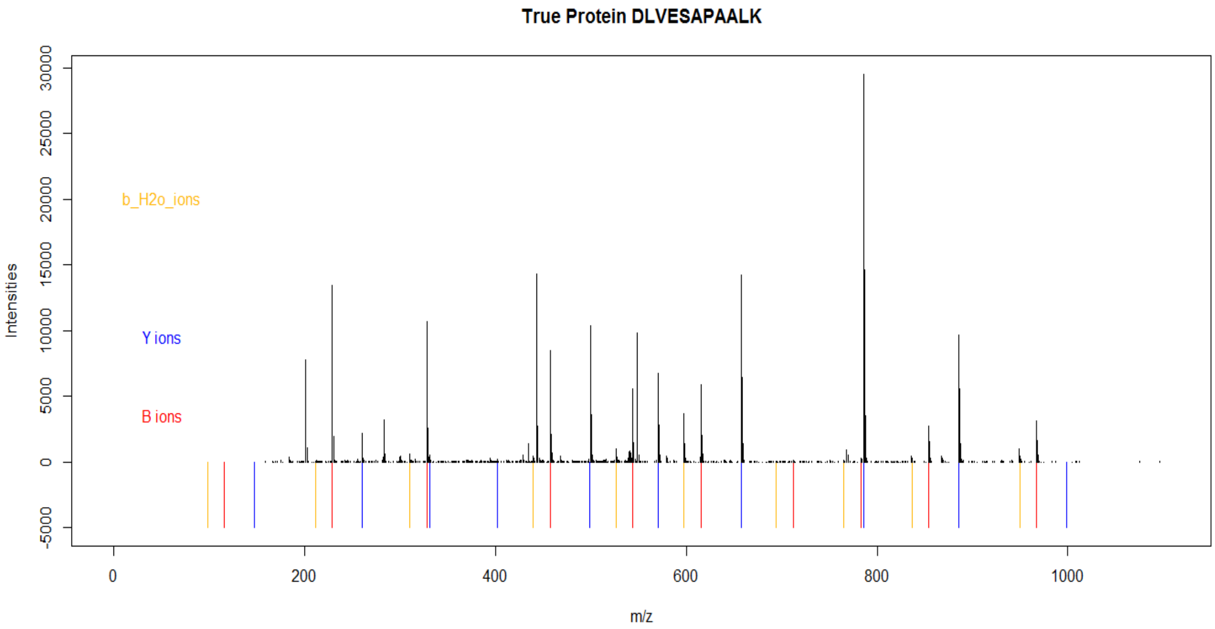

Figure 10: The observed spectrum plotted against the theoretical spectrum for the peptide *DLVESAPAALK*, showing the three ions used in our model: b ions, y ions and $b-H_2O$ used in our model. Plotted below the zero axis is the theoretical spectrum and the observed spectrum is plotted above the zero axis.

Table 5.2: The top estimated peptides from the *MCMC* algorithm along with their corresponding log posterior densities and log likelihood for the peptide *DLVESAPAALK*. The true peptide is in bold.

Peptide	**Log Posterior**	**Log Likelihood**
DLVESAPAALK	102.99	141.08
VEVETPQALQ	73.63	109.28
NNVESAPAALK	72.03	109.47
LDVEYYALQ	72.02	111.07
EVPAFAPAALK	69.73	107.81
NNVESFYLQ	68.96	107.47
EVVNGANVALK	68.52	107.74
NNVETPKALK	67.99	105.21

Setting an initial peptide of $QVVESLFK$ from the results of PepNovo, our best estimate for the true peptide is $DLVESAPAALK$ after $50,000$ iterations. It is worth noting why a smaller number of iterations was used compared to Example 1. First, note that the length of the peptide in Example 1 is shorter, 8 compared to 11, respectively. We noticed the refinement worked better for peptides whose sequences were longer because they have more data. The top estimated peptides for this example along with their corresponding log likelihood value is shown in Table 5.2, where the true peptide is estimated as having the largest log posterior density.

6 DISCUSSION

To strengthen research in the proteomics field, we refined a Bayesian model, which relies on prior information of the peptides in generating the best estimate of the true peptide. The complexities of our posterior density influenced us to use a Markov chain Monte Carlo algorithm coupled with simulated annealing to obtain the posterior probabilities. By incorporating an additional ion into the original Bayesian model proposed by Lewis (2013), we saw the true peptide was estimated having the largest posterior density in both examples. We understand to have more competitive results, numerous peptides of varying length must be used to determine if we see similar results.

As part of future development, the Bayesian model can be refined by exploring more ions such as $b{-}NH_3$, $y{-}H_2O$, $y{-}NH_3$ etcetera. The goodness of fit measures of the candidate spectrum to the observed spectrum that drives the size of the likelihood function could be extended as follows:

$$S_1 = \sum_{i=1}^{p} \lambda_i^b \ \min_j d(x_j - \tau_i^b) + \lambda_i^y \ \min_j d(x_j - \tau_i^y) + \lambda_i^{b-H} \ \min_j d(x_j - \tau_i^{b-H}) + \lambda_i^{b-N} \ \min_j d(x_j - \tau_i^{b-N}) + ... \tag{8}$$

$$S_2 = \sum_{j=1}^{t} \min_{i,k} |x_j - \tau_i^k| \tag{9}$$

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

Upon adding $b - H_2O$ to the likelihood function, we observe that these additional ions are not as prevalent as the b and y ions. Due to the smaller prevalence of this ion type, the algorithm had to be run for longer iterations and thus, taking longer to estimate the peptide. As future work, more avenues need to be explored to improve upon ways of making our simulation quicker.

The cleavage prior is set to cater for only two ions, the b-ion and y-ion. As earlier discussed in Section 4.5, this computes the probabilities of "seeing" the b and y ion peaks corresponding to each complementary pair of ions in a peptide. This could be extended to include other potential ions already discussed.

We hope our refined Bayesian approach to peptide identification, will be an invaluable contribution to the challenging scope of peptide identification, interspersed with complex and computationally challenging data set.

.

BIBLIOGRAPHY

[1] Andrieu C, de Freitas, N, Doucet, A., and Jordan , M. An introduction to MCMC For Machine Learning. 2003.

[2] Ansong,C, Tolic, N, Purvine, S, Porwollik, S, Jones, M, Yoon, H.P.S, Martin, J.,Burnet, M, Monroe, M, Venepally, P, Smith, R, Peterson, S, Heffron, F, McClelland, M and Adkins, J. . Experimental annotation of post-translational features and translated coding regions in the pathogen salmonella tryphimurium, bmc genomics. 2011.

[3] applied proteomics inc. The Conversation Of The Body. `https://appliedproteomics.com/pipeline/`, 2015.

[4] Clalre M. Fraser, Jonathan A. Elsen Steven L. Saizberg. Microbial Genome Sequencing. 406, 17 August 2000.

[5] Baggerly K.A Coombes, K.R and J.S. Morris. Pre-processing mass spectrometry data, fundamentals of data mining in genomics and proteomics. 2007.

[6] Corie Lok. Mining the microbial dark matter. 522, 16 June 2015.

[7] E de Hoffman. Tandem mass spectrometry: A primer journal of mass spectrometry. 1996.

[8] Edited by Rossen Donev. *Advances in Protein Chemistry and Structural Biology, Protein Structure and Diseases*, volume 83. Academic Press, May 11, 2011.

[9] Huang, Y., Triscari, J.M, Pasa-Tolic, L., Anderson, A.G., Lipton, M.S, Smith, R. D, Wyosocki, V.H. Dissociation behavior of doubly-charged tryptic peptides : Correlation of gas-phase cleavage abundance with ramachandran plots. 2004.

[10] International Service For The Acquisition Of Agri-biotech Applications. Pocket K No. 15: 'Omics' Sciences: Genomics, Proteomics, and Metabolomics, November 2006.

[11] Lewis, C.N. (2013). *Protein Identification Using Bayesian Stochastic Search, (Doctoral dissertation). Retrieved from http://scholarcommons.sc.edu/etd/2674.*

[12] Chibo Liu. The application of seldi-tof-ms in clinical diagnosis of cancers, journal of biomedicine and biotechnology. Vol. 2011:6 pages, .doi:10.1155/2011/245821, 2011.

[13] Yan Luo. Application of proteomics mass spectrometry to the keap1/nrf2 chemoprevention pathway. 2008.

[14] M. Cannataro, P.H Guzzi, T. Mazza, and P. Veltr. Preprocessing, Management, and Analysis of Mass Spectrometry Proteomics Data. 2005.

[15] Mark P. Molloy and Frank A. Witzmann. Proteomics: Technologies and Applications. 4th September,2001.

[16] Leo McHugh and Jonathan W. Arthur. Computational methods for protein identification for mass spectrometry data. February 29, 2008.

[17] Morris, J. S, Baggerly, K. A, Gutstein, H. B, Coombes, K. R. Statistical contributions to proteomic research, methods in molecular biology. 2010.

[18] National Cancer Institute, Office Of Cancer Clinical Proteomics Research,`https://proteomics.cancer.gov/whatisproteomics`. What is Cancer Proteomics?

[19] Norman N. Hoffman M.D,Inc.,Gary H. Hoffman M.D, Elman Flroozmand M.D,Liza M. Caplendo M.D,Stephen Yoo M.D . Proteomics-The Incisionless Cure May be Closer Than You Think.

[20] Roepstorff P and Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptide. 1984.

[21] Paul R. Graves and Timothy A. J. Haystend. Proteomics: Technologies and Applications. Vol.66, 1 March, 2002.

[22] Research Unit for Molecular Medicine, Faculty of Health Sciences and Aurhus University Hospital,Skejby,Arhus, Denmark. Protein Misfolding And Degradation In Genetic Diseases. 1999.

[23] Robert, C. P and Casella , G. Monte Carlo Statistical Methods. 1994.

[24] Sorensa, D and Gianola, D. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. 2002.

[25] Hanno Steen and Marthias Mann. The abc's(and xyz's) of peptide sequencing. Vol. 5, September 2004.

[26] Tierney, L. Markov Chains For Exploring Posterior Distributions. 1994.

[27] Qingfen Zhang Vicki H. Wysocki, Katheryn A. Resing and Guilong Cheng. Mass spectrometry of peptides and proteins. Methods 35(2005).

[28] Wikipedia, the free encyclopedia, `https://en.wikipedia.org/wiki/Omics`. 2017.

[29] John R. Yates. Mass spectrometry and the age of the proteome. Vol. 33, 1998.

[30] Zoltan Szabo, Tamas Janasky. Challenges And Developments In Protein Identification Using Mass Spectrometry. 2015.

VITA

THEOPHILUS B.K. ACQUAH

Education

MS Mathematical Sciences,
East Tennessee State University, 2017.

B.Ed. Mathematics,
University of Cape Coast, Ghana 2011.

Professional Experience

Graduate Teaching Associate (ETSU),
Teaching Math 1530 - Probability & Statistics
(Fall 2016),
Teaching Math 1530 - Probability & Statistics
(Spring 2017)

Graduate Teaching Assistant,
Provided tutoring services at the
Center For Academic Achievement,
August 2015 - July 2016

Mathematics Tutor,
Mfantsiman Girls Senior High School,
Saltpond,Ghana,
June 2013 - Aug 2015

Head of Mathematics Department,
Obama College, Ghana,
June 2012 - April 2013

Professional Development (software)

Statistical & Mathematical
SAS, R, SPSS, Minitab, Matlab

Scripting Languages
PHP, HTML, Python, Latex

Affiliations

American Mathematics Society, 2015-2017,
Abstract Algebra Club, ETSU,
Math & Stats Club, ETSU