Optimization of data directed acquisition in tandem mass spectrometry for proteomics

Carrillo, Brian

Department of Biomedical Engineering,

McGill University, Montreal

February, 2004

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master in Biomedical Engineering

Your file Votre référence
ISBN: 0-612-98517-2
Our file Notre référence
ISBN: 0-612-98517-2

Acknowledgements

I would like to express my gratitude towards Dr. Robert E. Kearney, my supervisor, for providing insight and guidance during the topic selection, and throughout the thesis preparation and completion.

I would also like to thank everyone at BMED and the RPMPN, especially Dr. Alex Bell and Dr. Daniel Boismenu for their answers to countless questions during all stages of my thesis.

Special thanks go to my lab buddies Corey and Kossi, who provided assistance in all areas of the research, and of course, provided hours of entertainment.

And finally I would like to thank all the extra editors who read, and in some cases re-read, my thesis to provide valuable insight to help make this thesis as clear as possible, thank you Gabriella, Nushi, Benoit, and of course, the lovely Mona.

Abstract

LC-QTOF tandem mass spectrometers behave according to user controlled switching parameters, duty-cycle and repetition rate, which guide the selection of peptides and the timing of their fragmentation. Using a novel algorithm which analyses all spectra simultaneously, it has been found that the majority of available peptides are not fragmented with the current switching scheme. Unfortunately, it is not practical to experiment with the mass spectrometer to determine optimal switching parameters. In this study, simulation coupled with intensity surface analysis was used as a method of evaluating mass spectrometer performance. Algorithms that mimic the mass spectrometer were created in order to simulate its response to various data sets. The simulations resulted in operating curves displaying the trade-off between quality and quantity of fragment spectra. The optimal operating curve demonstrated that the current switching scheme is sub-optimal, and that new switching parameters with fewer duty cycles and fewer repetitions should be selected.

Résumé

Le spectromètre à fragmentation de type LC-QTOF se comporte selon les paramètres de changement, le cycle d'opération et le taux de répétition définis par l'usager. Ces paramètres contrôlent la sélection des peptides et l'instant de leur fragmentation. En utilisant un algorithme qui analyse tous les spectres simultanément, il a été remarqué que la majorité des peptides disponibles ne sont fragmentés avec la méthode actuelle. Malheureusement, ce n'est pas évident de déterminer expérimentalement les paramètres optimaux par le spectromètre. Dans cette étude, une simulation couplée avec une analyse de surface en deux dimensions des intensités de peptides a été utilisée comme méthode d'évaluation des performances du spectromètre. Des algorithmes qui imitent la fonctionnalité du spectromètre ont mis au point dans le but de simuler la réponse de différentes données. Les résultats de simulation sont présentés comme des graphes d'opération illustrant le compromis entre la qualité et quantité des spectres de fragmentation. La courbe d'opération optimale démontre que le mode d'utilisation présente n'est pas l'idéal, et que de nouveaux paramètres de changement avec moins des cycles d'opération et moins de répétitions devront être sélectionnés.

Table of Contents

Table of Figures

1 Introduction

Proteins are macromolecules essential for cellular activity and survival. Estimates from the human genome indicate that 30,000 genes are encoding more than 100,000 different proteins.[1] It is also estimated that a typical mammalian cell may contain as many as 10,000 different proteins.[2] Proteins have a wide variety of functions, acting as enzymes, structural elements, hormones, receptors and transporters, as contractile elements, antibodies, toxins, and blood clotting agents. Since proteins perform most of the cell's biological function, characterizing their function and interactions, and localizing them within the cell is essential to understanding cell function.[3]

Proteomics focuses on the systematic simultaneous analysis of large numbers of proteins in biological samples.[4] In the past, protein analysis was carried out by isolating and characterizing one protein at a time. However with the arrival of technological advancements, protein analysis has become automated, making high-throughput proteomics possible. The benefits have been shorter analysis times, consistency in the analysis process, and the flexibility of multiple assays. One of the most important technological advancements was the introduction of mass spectrometry to proteomics.[5]

In a matter of seconds a tandem mass spectrometer can measure the mass of a peptide, fragment it, and measure the mass/charge (m/z) ratio of the fragment ions. Algorithms for peptide sequencing and identification use the fragment ion spectra to determine the sequence of amino acids that made up the original peptide.[6] Ideally the identification is straightforward, since the mass differences of the fragment ions correspond to the constituent amino acids. Unfortunately, in practice problems arise that make identification difficult. First, fragmentation is not an ideal process since expected ion fragments may not form while unexpected fragments may be created. Second, as with most signals, noise is a problem.

The tandem mass spectrometer can only acquire one spectrum at a time, either the initial survey spectrum (MS) or the fragmentation spectrum (MS/MS) of one peptide in the

survey spectrum; either acquisition takes finite time. MS spectra are necessary to identify the molecular weight of peptides, but MS/MS spectra are necessary to identify the sequence of the peptide. Time limits the total number of spectra that can be acquired and so there is a trade-off between the number of peptides analyzed and how comprehensively each peptide is analyzed.

Our objective was to investigate the effects of two mass spectrometer operating parameters, the duty cycle and the repetition rate, on the quality and quantity of detected peptides. The effects of the operating parameters were examined by simulating the behaviour of the mass spectrometer at all feasible parameter combinations, and evaluating performance with the aid of a new surface intensity analysis algorithm. The results define operating curves characterizing the tradeoff between quality and quantity.

2 Background

Proteins are polymers composed of amino acid monomers. There are 20 different amino acids (in humans) which share a basic structure: a carboxyl group and an amino group separated by a single carbon atom as illustrated in Figure 2-1; a side chain (R group) on the centre carbon differentiates the amino acids, and gives them their unique properties.

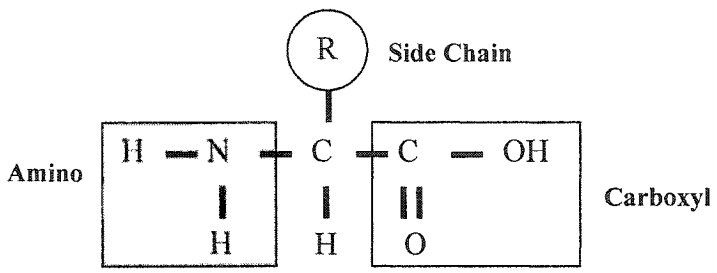

Figure 2-1: Amino acid structure

Shaded rectangle: amino group, clear rectangle: carboxyl group, circle: side chain

During protein synthesis, amino acids are joined together by peptide bonds creating polypeptide chains (see Figure 2-2).[2] When the process is completed the polypeptide chains are called proteins.

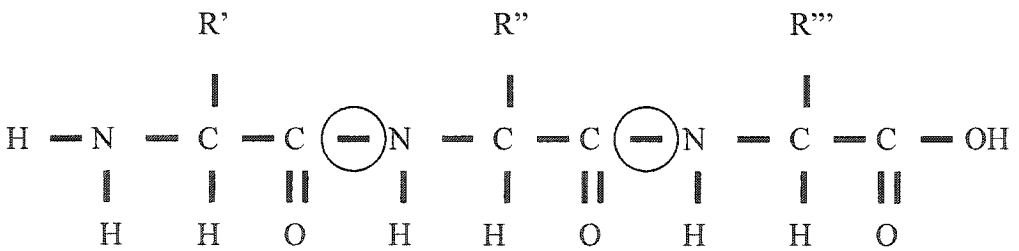

Figure 2-2: A three amino acid peptide, peptide bonds are encircled

2.1 Proteomics Pipeline

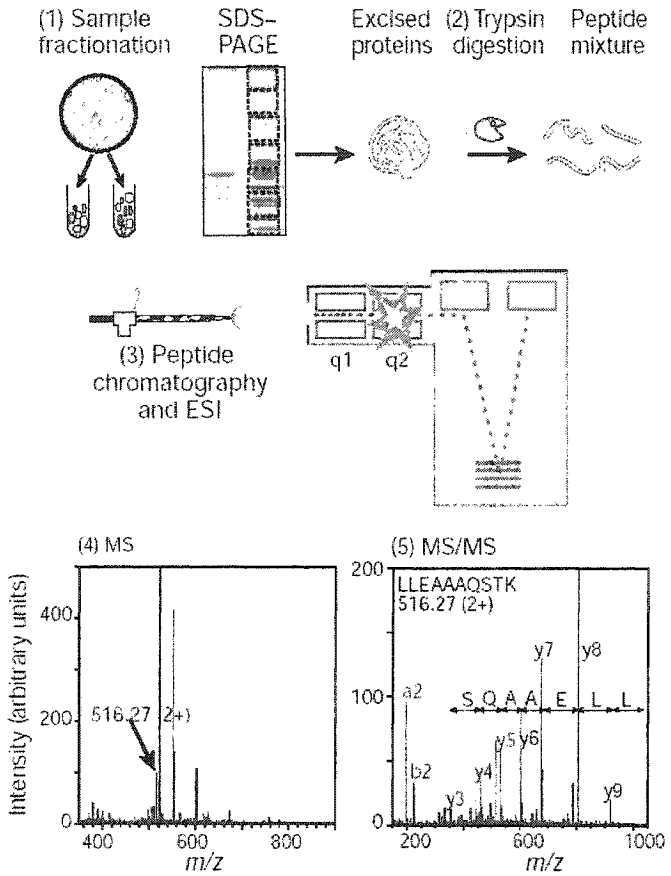

Figure 2-3: Generic mass spectrometry (MS)-based proteomics experiment. [7]

Figure 2-3 illustrates a typical proteomics pipeline showing the sequence of analysis used in high-throughput proteomics to identify proteins. In the first stage, proteins are extracted from the cell/tissue by biochemical fractionation and separated via gel electrophoresis to reduce sample complexity. In the second stage, a protease (usually trypsin) is used to digest the protein mixture into peptides of suitable size for mass spectrometer analysis. In the third stage, the peptides in the mixture are separated in time though the used of high performance liquid chromatography (HPLC), in preparation for entry to the mass spectrometer via electrospray ionization. In the fourth stage, the mass spectrum (MS) of the peptide mixture eluting at any instant is captured and peptides are detected. In the fifth stage, detected peptides are fragmented, and their tandem mass

spectra (MS/MS) are collected.[7] Note that during the time MS/MS spectra are acquired other peptides continue to elute from the HPLC and will not be sampled.

2.2 Protein separation and digestion

Proteins must be separated from each other to facilitate mass spectral identification. In high-throughput proteomics, proteins are commonly separated using 1D or 2D gel electrophoresis. Polyacrylamide gel electrophoresis is the technology of choice for separating complex protein mixtures.[8] Electrophoresis is based on the migration of charged particles in solution in response to an applied electric field. The rate of migration depends on a number of factors including the strength of the field, the protein size, and the viscosity of the gel. Proteins treated with sodium dodecyl sulphate (SDS) are denatured as SDS attaches to the polypeptide backbone; SDS also adds a negative charge to the protein in direct proportion to its length. Consequently, gel electrophoresis of proteins treaded with SDS separates them on the basis of their length (length is generally proportional to molecular weight) and not on the constituent amino acids of the protein.[9] The rightmost series of bands in Figure 2-4 illustrate a series of proteins isolated from human HeLa cells on a 1D gel. The leftmost series of bands are molecular markers with known masses.

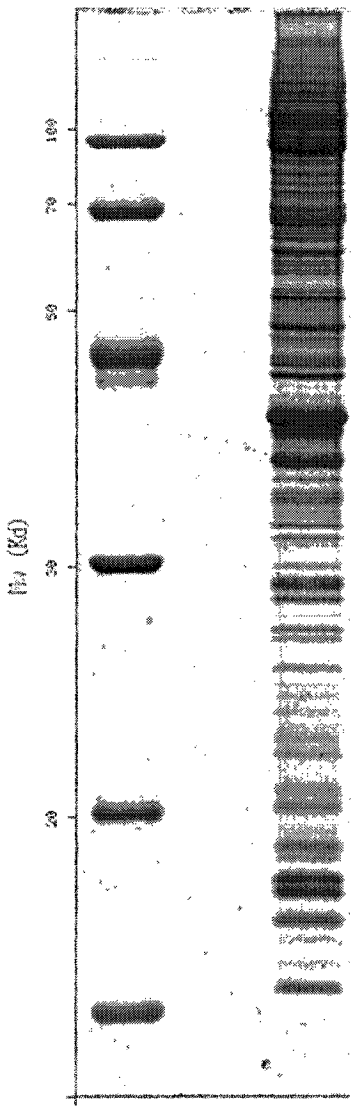

Figure 2-4: 1D gel of Coomassie Blue stained nucleolar proteins.

Human HeLa cells (right) and markers with known molecular weight (left)[10]

Even after gel separation, protein mixtures may be too complex for analysis by mass spectrometry. In addition, the mass of proteins vary widely which poses a problem for the spectrometer. In the Swiss-Prot database [11] (*KNOWLEDGEBASE RELEASE 42.9 STATISTICS*), an online database of proteins and associated information, peptide sequences range from 2 amino acids (261 Da) to 8797 amino acids (1,011,034 Da). Most mass spectrometers are unable to operate over such a large mass range; furthermore, the molecular mass of a protein cannot yet be measured with enough resolution to identify it

unambiguously. An additional confounding factor is that post-translational modifications may change the protein mass. Consequently, it is the general practice in high-throughput mass spectrometry to digest the proteins into smaller amino acid sequences (peptides) that are more consistent in length and thus more amenable to processing by the spectrometer. Once the sequence of a peptide(s) is known it can be used to search protein databases to find the parent protein.

Each gel slice (a gel is generally partitioned into discrete slices or blocks) will contain one or more proteins and so will generate many peptides. To resolve these peptides on the mass spectrometer, another separation step is necessary; liquid chromatography (LC) is used to separate peptides in time. LC separation of peptides begins by adsorbing all peptides onto the organic coating of beads packed in a column (tube). An acetonitrile solution is then run through the column and its concentration is varied with time along a predefined gradient. Peptides go into solution when their affinity for the packed column becomes less than their affinity for the acetonitrile solution.[12] Hydrophobicity, a measure of peptide affinity for organic molecules, varies from peptide to peptide and is determined by their amino acid sequence. Thus by choosing the gradient it is possible to control the number of peptides eluting at any time.

2.3 The Mass Spectrometer

The mass spectrometer measures the mass of molecular scale sized charged substances that can be transferred to a vacuum. Typically the mass spectrometer consists of three components: 1) the ion source, 2) the mass analyzer, and 3) the detector.

2.3.1 Ion Source

The mass spectrometer measures the mass of charged molecules; thus molecules must be ionized before they enter the spectrometer. A variety of different ion sources exist, including electron ionization, chemical ionization, field desorption, laser desorption, thermospray, and electrospray. The two most common ion sources in proteomics are matrix-assisted laser desorption (MALDI) and electrospray.

MALDI sources use a two step process to ionize molecules. First, the molecules of interest are mixed with an organic solvent solution, or matrix, that has a strong absorption band at a particular laser wavelength. The matrix solution is dried leaving matrix crystals containing the molecules to be analyzed. Secondly, an intense laser pulse heats a section of matrix causing it to sublimate. As the gas matrix expands, the molecules of interest are liberated. The ionization reactions however, occur through processes that are not yet fully understood.

Electrospray ionization (ESI) sources use a liquid (usually an acetonitrile solution eluting from the front end of an HPLC column) to carry peptides through a metallic capillary, and eventually into the vacuum of the mass spectrometer. A high voltage applied to the capillary results in a very high electric field at the capillary tip that causes charge to accumulate at the liquid surface. When sufficient charge accumulates surface tension is broken forming highly charged droplets which are passed though a heated inert gas to evaporate the acetonitrile solution. Peptides in the solution will retain the droplet's charge.

Figure 2-5 illustrates a schematic of an ESI assembly; ions travel from the high voltage needle towards and through the lower voltage sampling cone. The small orifice of the sampling cone provides an interface between the ambient atmospheric pressure and the vacuum of the mass spectrometer.

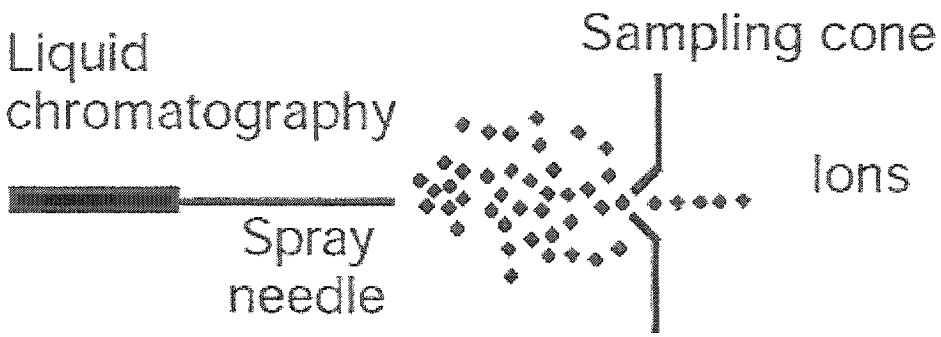

Figure 2-5: An electrospray ionization (ESI) assembly[7]

2.3.2 Mass Analyzer

The second component of the mass spectrometer, the mass analyzer, separates ions based on their *m/z* ratio. There are many types of mass analyzers, but they can be divided into two main categories: scanning analyzers (quadrupoles and ion traps) which process ions sequentially based on their mass-to-charge ratio, and simultaneous mass analyzers (time-of-flight and ion cyclotron analyzers) which process all ions together.[13] Quadrupoles and the time-of-flight (TOF) analyzers are used most often in high-throughput proteomics.

The quadrupole is a set of four parallel metal rods (either cylindrical or parabolic) that are energized by modulating frequencies and voltages. Ions entering the quadrupole oscillate as they travel the length of the rods. By suitably choosing the voltages and frequencies applied to the rods, ions of specific *m/z* will exhibit stable oscillations and traverse the length of the rods uninhibited; all other ions will have unstable oscillations, strike the rods, discharge, and therefore not be detected. By varying the voltages and frequencies with time, the quadrupole mass analyzer can scan a mass range to create a mass spectrum.

The time-of-flight tube separates molecules based on the time required to traverse a fixed distance. At one end of the tube, all molecules are initially imparted with the same kinetic energy by a high voltage source. Consequently, molecules with small mass will be accelerated to a higher velocity than those with larger mass and so will traverse the length of the tube more quickly. A detector placed at the opposite end of the tube records the time that molecules arrive. The *m/z* of a molecule is directly proportional to the square of the time required to traverse the length of the flight tube[13]:

$$m/z = \left(\frac{2V_s e}{d^2}\right)t^2 \qquad (1)$$

where, V_s is the accelerating potential, d is the length of the flight tube, and *e* is the charge of a single electron.

Figure 2-6 shows a schematic of a TOF spectrometer. Ions are propelled down the tube by the high voltage source, and their arrival time is recorded by the detector.

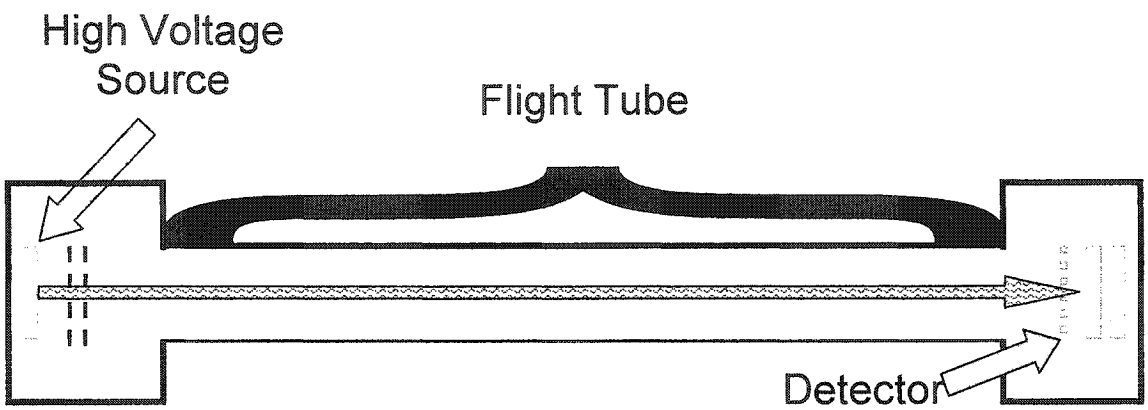

Figure 2-6: Time of flight mass spectrometer schematic

2.3.3 Detectors

The final component of the mass spectrometer is the detector which converts the beam of ions from the mass analyzer into a useable signal. Two types of detectors exist: one type allows the direct measurement of the ions (e.g. photographic plates and Faraday cages); the other amplifies the signal before recording (e.g. micro channel plates (MCP) and photon multiplier detectors).[13]

The MCP is the detector frequently used in TOF instruments; it consists of a plate containing parallel cylindrical holes coated with a semi-conducting material that releases secondary electrons when struck by ions, thus multiplying the initial ion. An accelerating voltage of ~1kV is applied across the plate to ensure electrons traverse through the plate. This cascade of electrons can cause gains in the order of 10^5. Several plates can be combined for further amplification.[13]

2.4 Tandem mass spectrometry

Mass spectrometer analysis of separated peptides allows for peptide mass to be accurately determined, but does not provide any structural or chemical information. Gay et al. [16] demonstrated that, in the SWISS-PROT protein database, there are thousands of peptides having the same mass to a resolution of 10^{-5} Da, a resolution much higher than most

spectrometers can provide. The Micromass Q-TOF mass spectrometer, for example, provides a resolution in the order of 10^{-2} Da.

Tandem mass spectrometry has the potential to resolve this redundancy. Tandem mass spectrometry uses two mass analyzers. The first mass analyzer selects a single peptide mass from the initial mass spectrum (MS) by filtering out all other masses. The single peptide is then fragmented in a collision cell and the second mass analyzer acquires the resulting fragmentation spectra (MS/MS). Since peptides fragment at known locations, the fragment spectrum can be used to determine the amino acid sequence of the peptide.[13]

The switching behaviour of the mass spectrometer is controlled via two "switching parameters". The first parameter, the **duty cycle**, indicates the maximum number of peptides that can be selected for fragmentation in a single MS spectrum. The second parameter, the **repetition rate**, controls the number of times each peptide is fragmented. The switching parameters are denoted A-B, where A is the duty cycle and B is the repetition rate.

2.5 LC-QTOF mass spectrometer

The LC-Q-TOF incarnation of a mass spectrometer is commonly used in high-throughput proteomics for the analysis of complex samples. Liquid chromatography (LC) helps to reduce the complexity of samples before injection into the mass spectrometer.

Figure 2-7 illustrates the path traveled by ions as they are processed and detected by the Q-TOF mass spectrometer. The ions first pass through the quadrupole analyzer (Q_1) where, if the spectrometer is operating in MS mode all ions will pass through unabated, or if the spectrometer is operating in MS/MS mode then only the ions with the selected *m/z* ratio will be able to pass. When operating in MS/MS mode, the selected peptides will collide with uncharged gas molecules (usually nitrogen) in the collision cell (q_2). The kinetic energy transferred in the collision causes the peptides to fragment in a process known as collision-induced dissociation (CID). At the entrance to the flight tube (TOF)

kinetic energy is imparted to the ions by the pusher, changing their trajectory. The reflector (or reflectron) inverts the direction of ions, through the use of an electric field, in order to attenuate slight inconsistencies in the velocity of ions with identical mass. Ions with slightly higher velocities will penetrate deeper into the reflector, and thus take longer trajectories and more time. Finally, when ions reach the detector they are counted and given a timestamp.

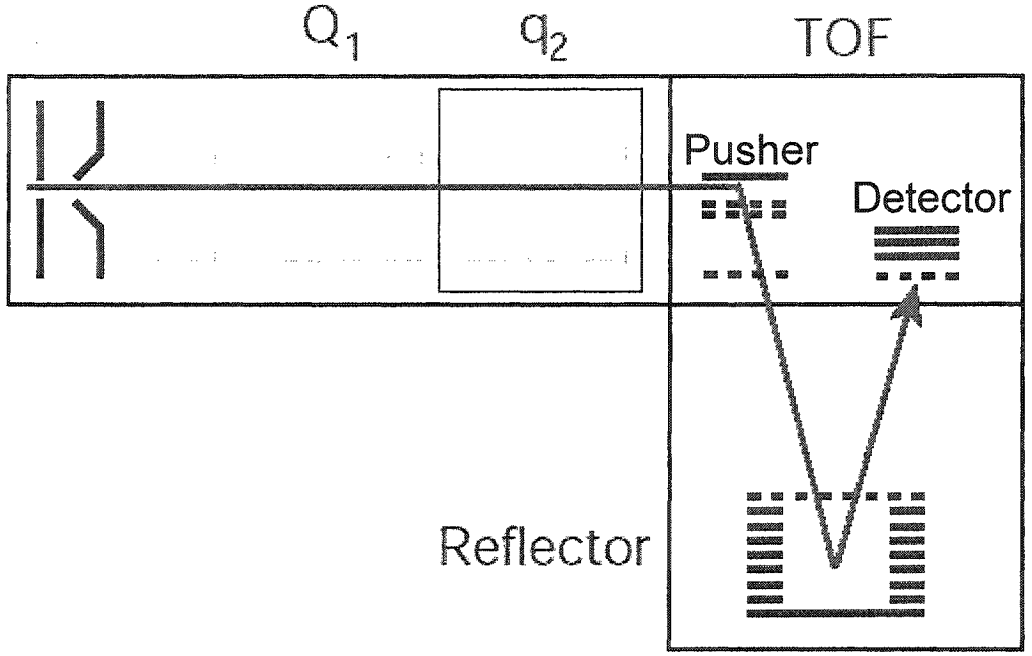

Figure 2-7: Schematic of a Q-TOF mass spectrometer[7]

2.6 Isotopic Distribution

The resolution of the many mass spectrometers allows a single peptide to be resolved into many isotopic peaks whose spacing is dependant on the peptide charge. The first peak, the peak with the smallest *m/z*, called the mono-isotopic peak, is composed of atoms with only the base isotope (C_{12}, N_{14}, O_{16}, S_{32}). The remaining peaks are composed of atoms containing one or more of the larger naturally occurring isotopes (C_{13}, N_{15}, O_{18}, S_{34}). The relative intensity of these remaining peaks is determined by the atomic composition of the peptide, and the ratio of isotopes that these atoms contain. Natural carbon, for example,

consists of 98.89% C_{12} and only 1.11% C_{13}. Although the percentage of large isotopes is small, they have a dramatic impact on the peak distribution of larger molecules.

In a hypothetical molecule containing 100 Carbons, probability theory predicts that in roughly two thirds of the cases, at least one carbon atom would be a heavier isotope. In a large population of these 100 carbon molecules the relative abundances of the different isotopes would behave according to Table 1.[14]

Isotope Number	m/z	Percent Total Intensity
0	1200.00000	32.85
1	1201.00335	36.77
2	1202.00671	20.38
3	1203.01006	7.45
4	1204.01342	2.02
5	1205.01677	0.43

Table 1: The isotopic distribution for a theoretical molecule containing only 100 Carbons

As the mass of a peptide increases, so does the relative intensity of the non-mono-isotopic peaks. These peaks eventually dominate the spectrum (as does the second isotope in Table 1), and eventually the mono-isotopic peak will have negligible intensity.

If the atomic composition of a peptide is known, a simple binomial expansion of the number of atoms and the proportions of large isotopes can predict the relative abundance of the various isotopes. Molecules are present only in integer amounts so the counting statistics can be modeled in terms of Poisson distributions. Therefore, given a mono-isotopic peak and its chemical formula, the relative intensity of other isotopic peaks can be predicted.

However, in high-throughput proteomics, the chemical composition of the peak under consideration is not known so the isotopic peak distributions cannot be determined. One

way to estimate composition is to define an average amino acid. Breen et al. [15] computed the average amino acid from large protein databases and found it to have the chemical formula of $C_{10}H_{16}N_3O_3$. By concatenating this average amino acid with itself, various mass peptides can be constructed to map most mass ranges. Gay et al. [16] attempted a brute force method where the isotopic distribution was computed for every peptide in a large protein database and polynomials were fitted through the data to map the mass range. Figure 2-8 shows the weighted mean heights of isotopic peaks for peptides between 800 and 3000Da extracted from the SWISS-PROT database. The second isotopic peak ($^{+1}$M) rises over the mono-isotopic peak ($^{+0}$M) near 1800 Da, well within the detection range of the LC-QTOF mass spectrometer, especially if the molecules are multiply charged. The inset shows the intensity of the mono-isotopic peak relative to the other isotopic peaks.

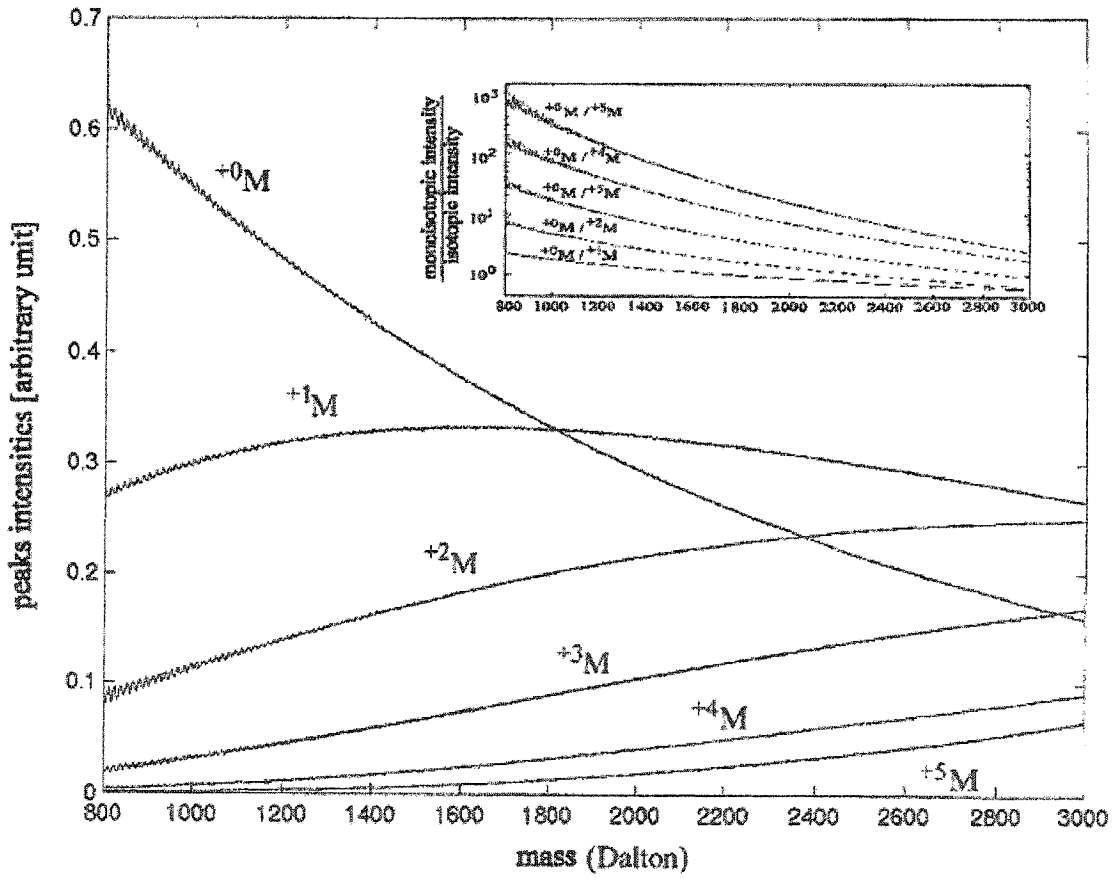

Figure 2-8: Isotopic peak intensities for 800-3000Da peptides from SWISS-PROT [16]

2.7 Spectral Processing

Peptides are not always easily identified within the raw spectra provided by the mass spectrometer. There are a number of reasons for this:

1) each peptide is represented by several isotopic peaks with heights varying according to its composition

2) each isotopic peak is in itself a distribution, which is Gaussian-like, that spreads the peak across the *m/z* axis and is dependant on the instruments resolving power

3) several peptides may appear close to each other, causing both their isotopic distributions and their peak distributions to overlap

4) finally noise signals can corrupt the spectra.

Processing the spectra using signal processing techniques can help to alleviate these problems. Figure 2-9 shows a section of a typical raw spectrum which illustrates three of these points. The peak labeled 'a' is the mono-isotopic peak of the peptide cluster 'abcd'. Peak 'a', is not localized to a single *m/z* but has a finite width. The arrow indicates a section of considerable noise, although noise can be seen throughout the spectrum, including within peaks.

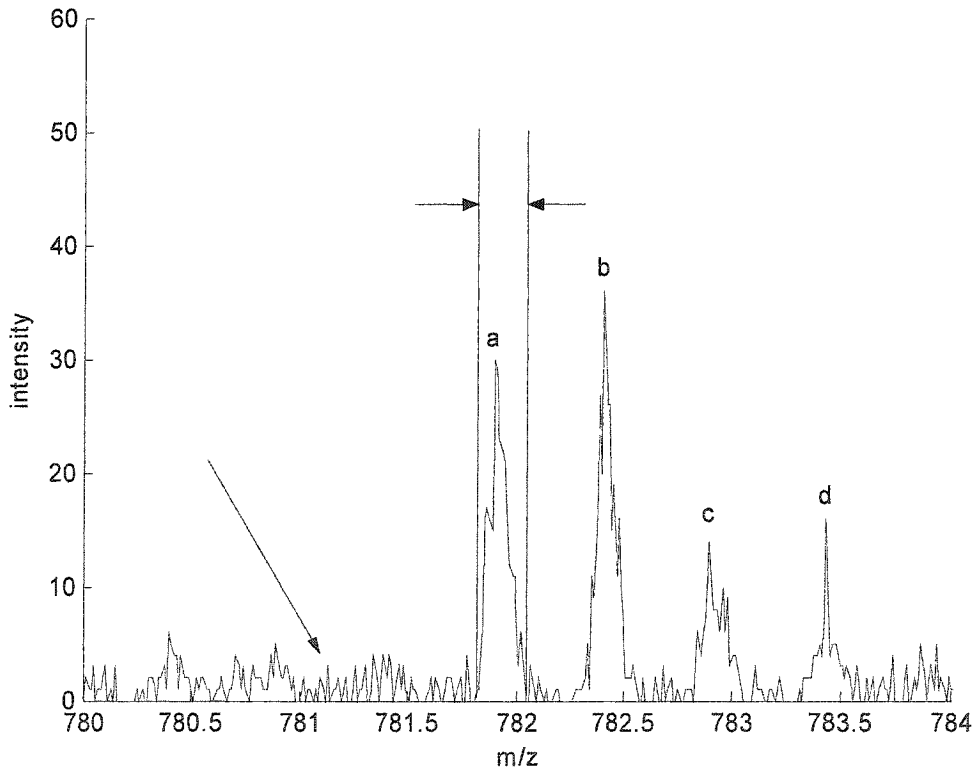

Figure 2-9: A section of a typical raw spectrum obtained from an LC-QTOF.

2.7.1 Filtering

The spectra recorded by the mass spectrometer are, like most experimental signals, corrupted by noise. In mass spectrometry, the main sources of noise are the chemicals used in sample preparation and electronics. While it may be difficult to separate and identify the various signals, it is possible to model the properties of the ideal signal and use these parameters to aid in filtering. Lekpor et al. [17] processed calibration spectra and determined that the ideal peptide signal contains information only in a low-frequency range whose bandwidth varies with *m/z*. Figure 2-10 shows the frequency content of peaks, computed from the flight time (~10s of μs) of ions in a TOF tube, at various *m/z* locations. Peaks at low *m/z* have higher frequency content then those at a high *m/z*. A piecewise linear filtering of the spectrum using cut off frequencies tuned to each peak can reduce high-frequency noise without modifying the information of the peptide ion peaks.

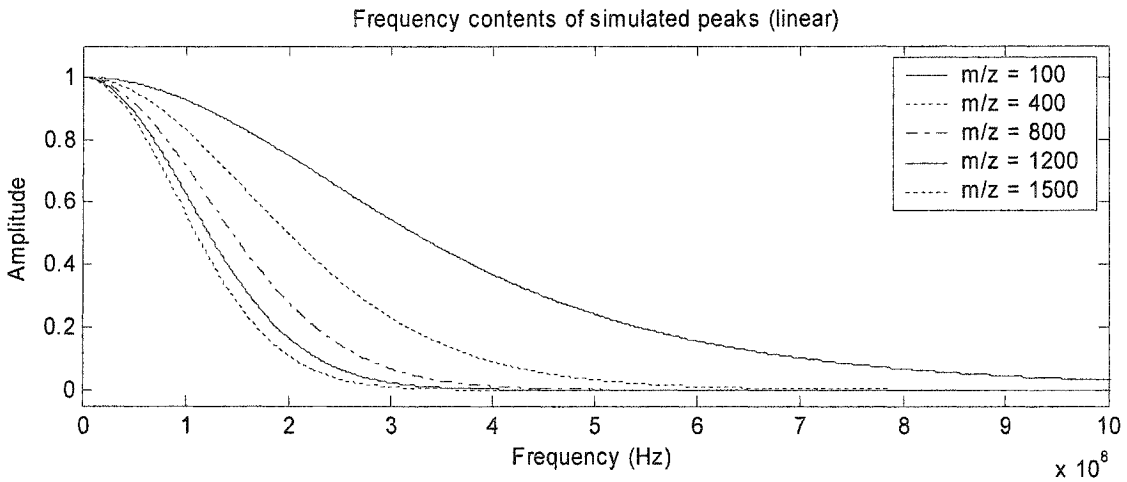

Figure 2-10: Frequency content of peaks at various m/z [17]

2.7.2 Peak Picking

Peak picking is the process that tries to resolve the problem of each peak existing as an *m/z* distribution. Peak picking converts the peak distribution to a single, "correct" *m/z* value. The algorithms used for peak picking in mass spectra vary considerably between laboratories. Each laboratory selects its own method of identifying and localizing peaks. Hastings et al. [18] used a local maxima method in conjunction with a peak in chromatogram to identify "true" peaks. Breen et al.[15] used a watershed technique to

isolate peaks and then computed the centroid of isolated peaks to localize them. Finally, Zhang et al. [19] used simple global maxima to isolate the most intense peak, processed it, then iteratively searched the rest of the spectra. Each method has its own strengths and weaknesses, and each peak picking method will likely need to be tweaked to suit a particular instrument and protocol.

2.7.3 Deisotoping

Deisotoping is the process whereby all the peaks in an isotopic distribution are folded back to the mono-isotopic peak and all peak intensities summed. This process also identifies the charge on the peptide so that its uncharged mass can be determined. This is a data reduction step that characterizes peptides by two parameters, mass and charge, instead of the location and intensity of a set of peaks. Deisotoping methods fall into two general categories, probabilistic methods, and heuristic methods. Probabilistic methods attempt to find peptide models whose behaviour best fit the available data. Maximum entropy, a probabilistic method, iteratively tries to optimize the match between a calculated spectrum and the observed spectrum. Maximum entropy is computationally complex and thus has not been used extensively.[20] Heuristic methods, on the other hand, attempt to exploit known patterns or parameters within the data to simplify deisotoping. Heuristic methods use stepwise decision making and are, in general, computationally simple. However, they are based on "rule-of-thumb" observations and may not generalize well.

Deisotoping is also occasionally referred to as deconvolution. Deconvolution is the unraveling of overlapping signals into their constituent signals. Overlapping spectra are usually automatically resolved by deisotoping algorithms.

2.8 Problem Formulation

During tandem mass spectrometry experiments, there is only a finite time when a peptide elutes from an HPLC to determine its overall mass (to select for fragmentation) and its fragmentation pattern (to determine its amino acid sequence). This time limitation limits

the acquisition of enough information required to unambiguously identify the peptides in a sample. Exacerbating the situation, several peptides may elute simultaneously, or have overlapping elution profiles. Multi-peptide elution causes the mass spectrometer to sacrifice fragmentation spectra quality to capture information on more of the available peptides.

Figure 2-11 depicts a set of several spectra from a typical experiment run with a 4-5 duty cycle. Each pixel row represents one spectrum, while pixel greyscale intensity is proportional to the intensity of the originating spectra. The figure illustrates the elution of several peptides over a two minute period, in just a 50 m/z mass range. In the figure, dark spots indicate peaks, sets of dark spots indicate peak clusters (peptides), and the two sets of white circles show the peptides selected for fragmentation. It is obvious that most of the peptides are ignored.

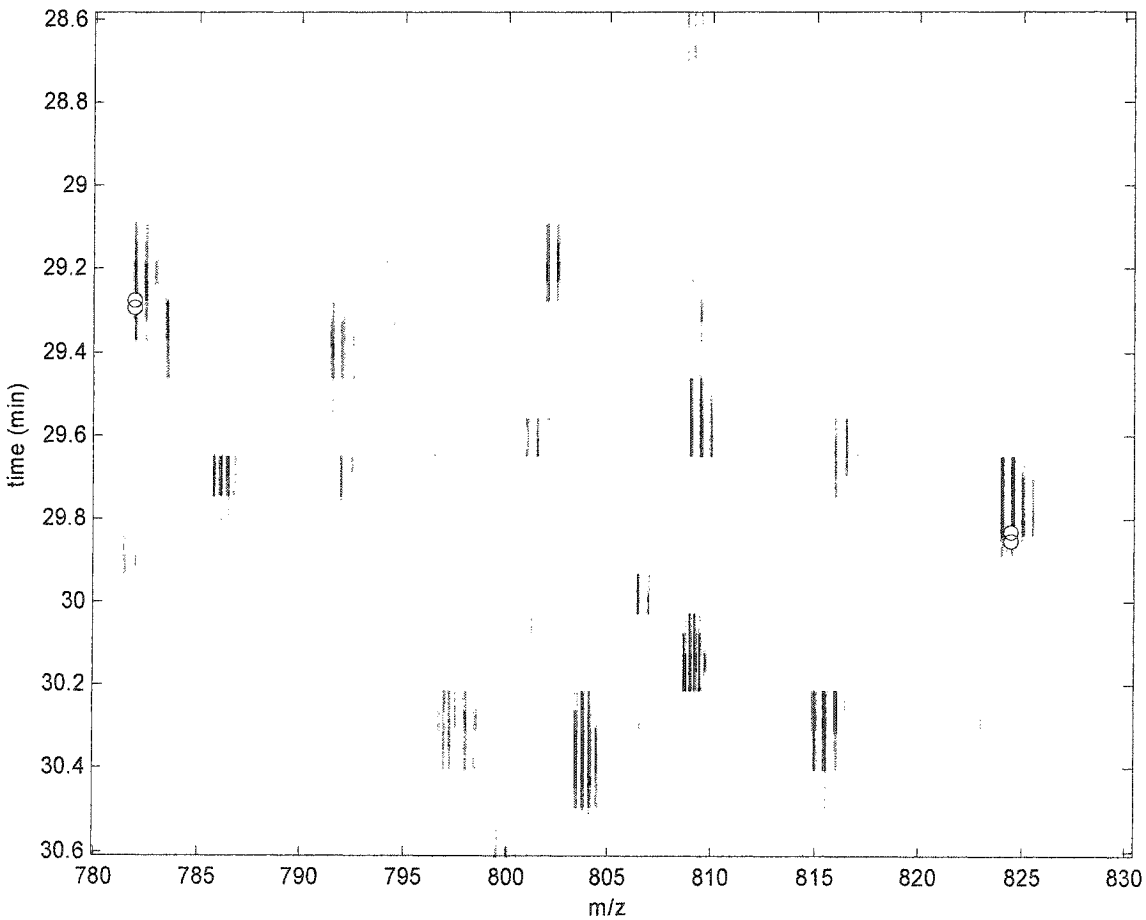

Figure 2-11: Section of a 30 min gradient, with locations of MS/MS (from an LC-QTOF)

Too many peptides and not enough time leads to a classic tradeoff problem: what is the optimal balance of quality and quantity to maximize the amount of information collected? The mass spectrometer's switching parameters control this tradeoff. Finding these optimal settings will allow high-throughput proteomics to take full advantage of mass spectrometry resources.

3 Experimental Overview

This thesis investigates the effects that switching parameters have on the quality and quantity of peptides identified in the LC-Q-TOF mass spectrometer. The experiments used data from a sample containing ~1000 proteins. The sample was run in MS only mode at four different gradient lengths and repeated four times at each gradient. Data were collected in this manner to avoid the time gaps created by acquiring MS/MS spectra, thus ensuring that all of the peptide peak profiles were observed.

Surface intensity analysis was used to analyze each dataset to determine where and when (time & *m/z*) each peptide eluted. The surface intensity algorithm is more robust than traditional peptide detection algorithms as it processes multiple spectra simultaneously. Surface intensity analysis leads to better peptide localization and fewer errors.

The same data was then processed via traditional single-spectrum algorithms to simulate the operation of the mass spectrometer. The simulations were run with an array of switching protocols, duty cycles from 1-20 and repetitions from 1-10, to determine the number of peptides that would be fragmented, and to estimate the spectrum quality based on the intensity of the peptide in the MS spectrum.

Finally, the simulations were compared to the results of surface intensity analysis to determine how many of the available peptides the simulations fragmented, and how close those fragmentations were to the maximum intensity of the peptide profile. This comparison provides a measure of performance for the mass spectrometer, and will determine the optimal switching parameters.

Figure 3-1 depicts an overview of these processing steps as information flows from raw data through the various processing stages.

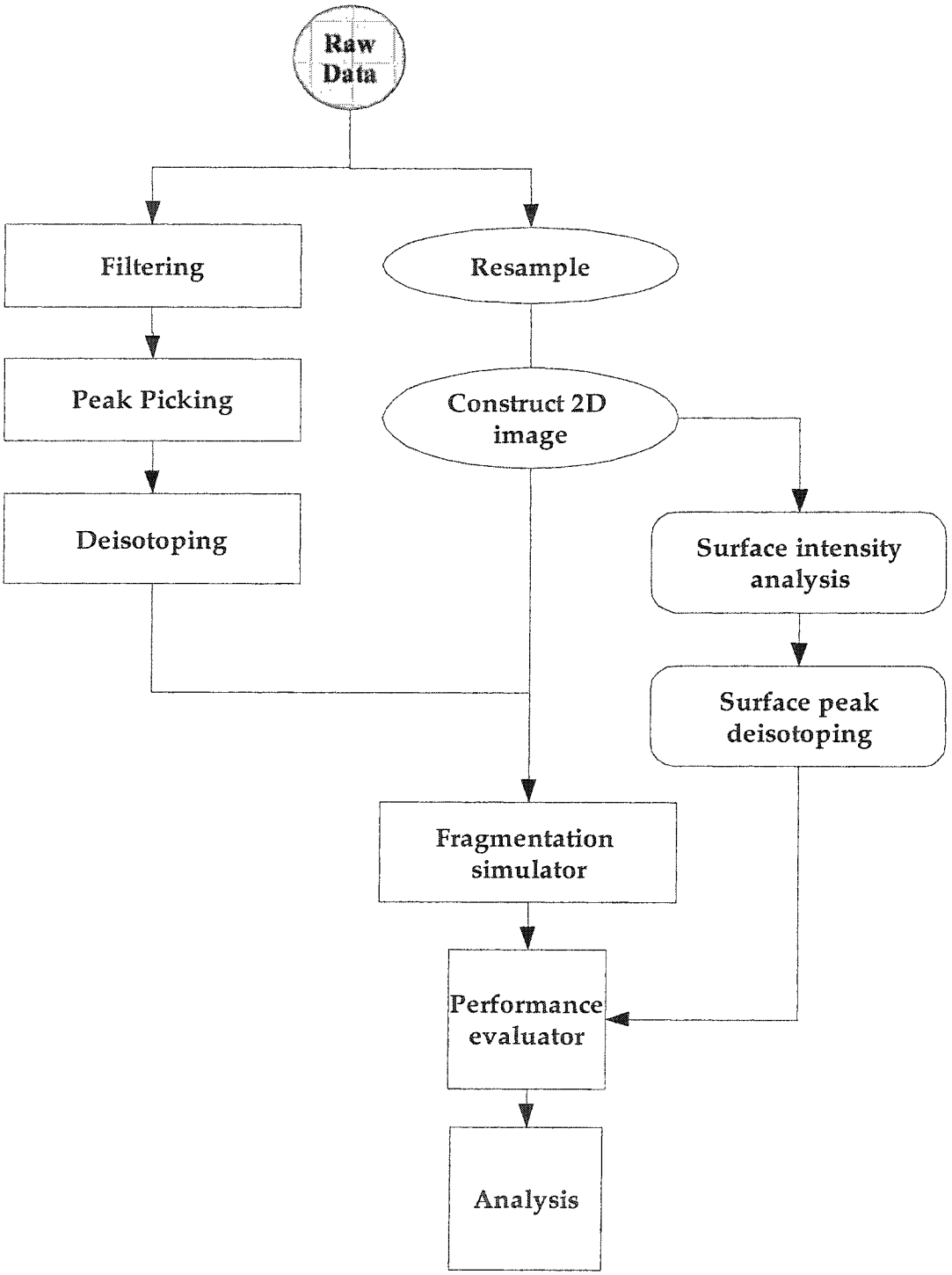

Figure 3-1: Information processing flowchart

Rectangle: simulation of mass spectrometer (Chapter 4)

Ellipse: data transformation to allow surface analysis (Chapter 5)

Rounded Rectangle: surface analysis (Chapter 5)

Square: comparison and performance measurement (Chapter 6 & 7)

3.1 Data

The mass spectrometry data used in this thesis was acquired using the method described by Lavoie et al. [21]:

> Total microsomes were obtained by differential centrifugation of rat liver homogenates (Paiement and Bergeron, 1983)[1]. They were resuspended in sucrose to give a final concentration of 1.38 M, placed under a step-gradient of 1.0, 0.86, and 0.25 M sucrose, and centrifuged using a Beckman SW 60 rotor at 300,000 g_{av} for 60 min. A subfraction containing smooth microsomes and low density rough microsomes (1.17 g/cm^3) was obtained from the upper half of the 1.38 M sucrose step above the residual pellet after centrifugation. This fraction, characterized as LDMs, was washed once by centrifugation and resuspension in 0.25 M sucrose at 100,000g_{av} (Lavoie et al., 1996)[2]. High density rough microsomes were prepared as previously described (Paiement and Bergeron, 1983).

The extracted fractions were then separated using the method described by Wasiak et al. [22] with the exception that the SDS-PAGE was a 7-15% gradient with 4 molar urea:

> ... proteins were separated by SDS-PAGE and stained with Coomassie blue. The gel lane was then cut horizontally into 62 even sized gel slices. The slices were dehydrated in acetonitrile and washed by two cycles of 10 min in 100 mM $(NH_4)_2CO_3$ before the addition of an equal volume of acetonitrile. The completely destained gel slices were then treated for 30 min with 10 mM dithiothreitol to reduce cystinyl residues and for 20 min with 55 mM iodoacetamide to effect alkylation. After an additional round of $(NH_4)_2CO_3$ and acetonitrile washes, the slices were extracted with acetonitrile at 37°C. They were then incubated with trypsin (6 ng/µl in 50 mM $[NH_4]_2CO_3$) for 5 h at 37°C and the peptides were first extracted in 1% formic acid/2% acetonitrile followed by two further extractions with additions of acetonitrile. All treatments were performed robotically using a MassPrep Workstation (MicroMass).

Figure 3-2 illustrates the gel containing the peptides analyzed.

---

[1] Paiement, J., and J.J.M. Bergeron. 1983. Localization of GTP-stimulated core glycosylation to fused microsomes. *J. Cell Biol.* 96:1791–1796.

[2] Lavoie, C., J. Lanoix, F.W.K. Kan, and J. Paiement. 1996. Cell-free assembly of rough and smooth endoplasmic reticulum. *J. Cell Sci.* 109:1415–1425.

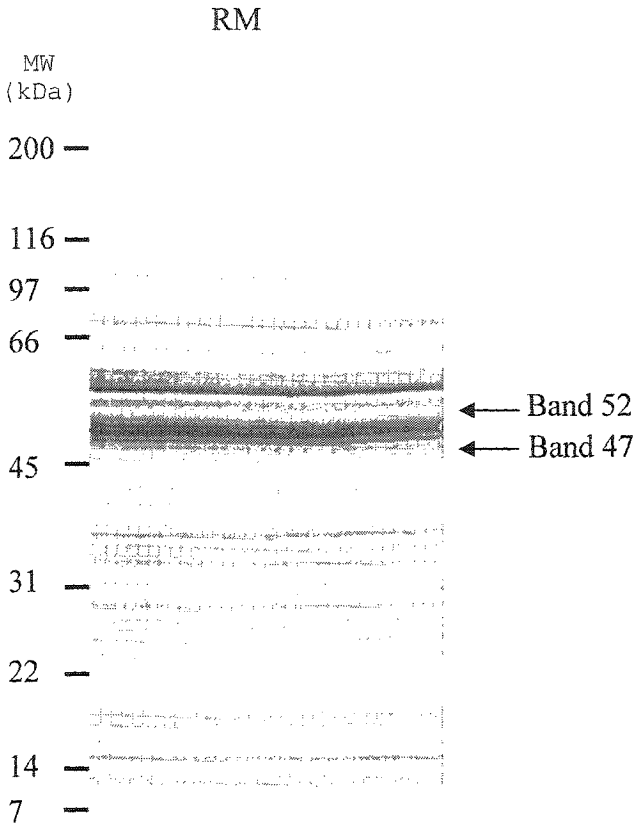

Figure 3-2: 1D gel of stained rough membrane proteins with extracted bands

Numbers on the left show the molecular weight of standards.

Peptides from bands 47 through 52 were pooled and analyzed by mass spectrometry using the protocol described in Wasiak et al. [22]:

> Extracted peptides were applied to a reverse phase guard column and then eluted in-line to a 10 cm by 75μm PicoFrit column filled with BioBasic C18. The column was eluted at 200 nl/min with a linear gradient of 5–70% acetonitrile/0.1% formic acid. Four gradients, 30min, 60min, 120min, and 240min in quadruplicate were used. A 2,000-V charge was applied to the PicoFrit column such that the eluted peptides are electrosprayed into a Cap liquid chromatography quadrupole time-of-flight MS (MicroMass). The mass spectrometer collected MS scans only.

4 Simulation

The behaviour of the mass spectrometer was simulated to avoid the time consuming and resource taxing alternative of running thousands of nearly identical experiments. Simulation allows for the modification of all parameters and the calculation of all possible outcomes in the order of minutes, whereas experimentation would require weeks or months. Simulating the exact behaviour of the mass spectrometer is not possible since the algorithms contained within the software are proprietary and not available for use outside the mass spectrometer. However, the rules that govern the software are known and are:

- Only doubly and triply charged peptides are considered for fragmentation.
- If multiple peptides are detected within a single scan, the most intense peptides are considered for fragmentation first.
- A peptide will not be considered for fragmentation if a peptide with the same mass and charge has been fragmented within a user specified time frame.

Algorithms that adhere to these rules were created to mimic the ion selection behaviour of the mass spectrometer. The data for these algorithms were 16 sets of real MS only spectra from a sample believed to contain ~10,000 peptides.

Figure 4-1 displays a small section of a typical MS spectrum containing two peptides with different charge states. This sample spectrum will be used to illustrate the effects of the various processing stages. In simulating the mass spectrometer, our approach was to analyze each spectrum using optimal methods so that any differences can be attributed to switching parameter issues. Actual implementations, however, are likely to be suboptimal due to time constraints.

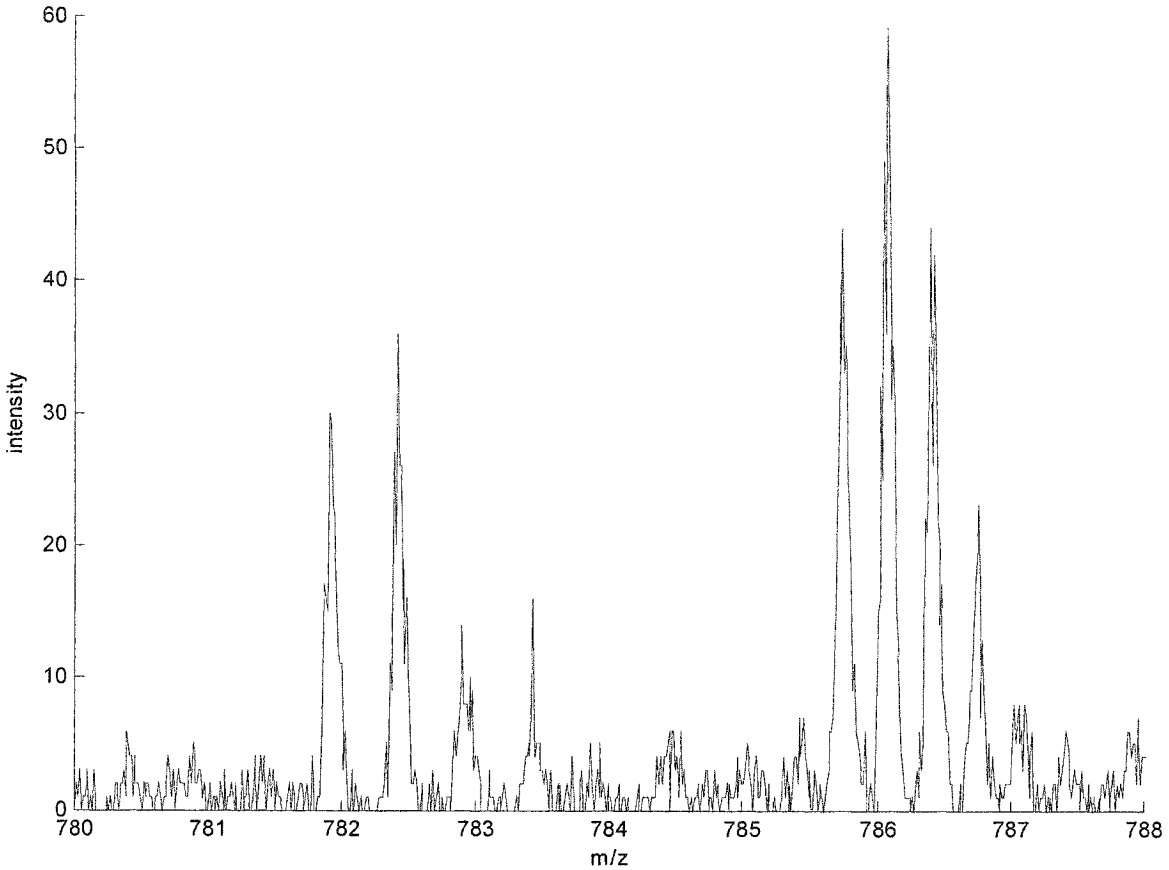

Figure 4-1: Section of a typical unprocessed spectrum

A doubly charged peptide at ~781.9 *m/z*, and a triply charged peptide at ~785.8 *m/z* are visible.

4.1 Spectrum filtering

Filtering was based on the Lekpor [17] method of piecewise linear filtering. The algorithm slices the spectrum into pieces 50 *m/z* wide. Each piece is then filtered using a Butterworth low pass filter with an order and cutoff frequency dependant on the *m/z* of the piece. These values were pre-computed by Lekpor based on the modeled peak shape of calibration data. A rectifying function is then used to remove any negative intensities (meaningless in mass spectrometry) introduced by the impulse response of the filter as they may cause problems in later processing stages. Figure 4-2 shows the result of filtering on the sample spectrum shown in Figure 4-1. The peaks are evidently, more consistent in shape and are smoother.

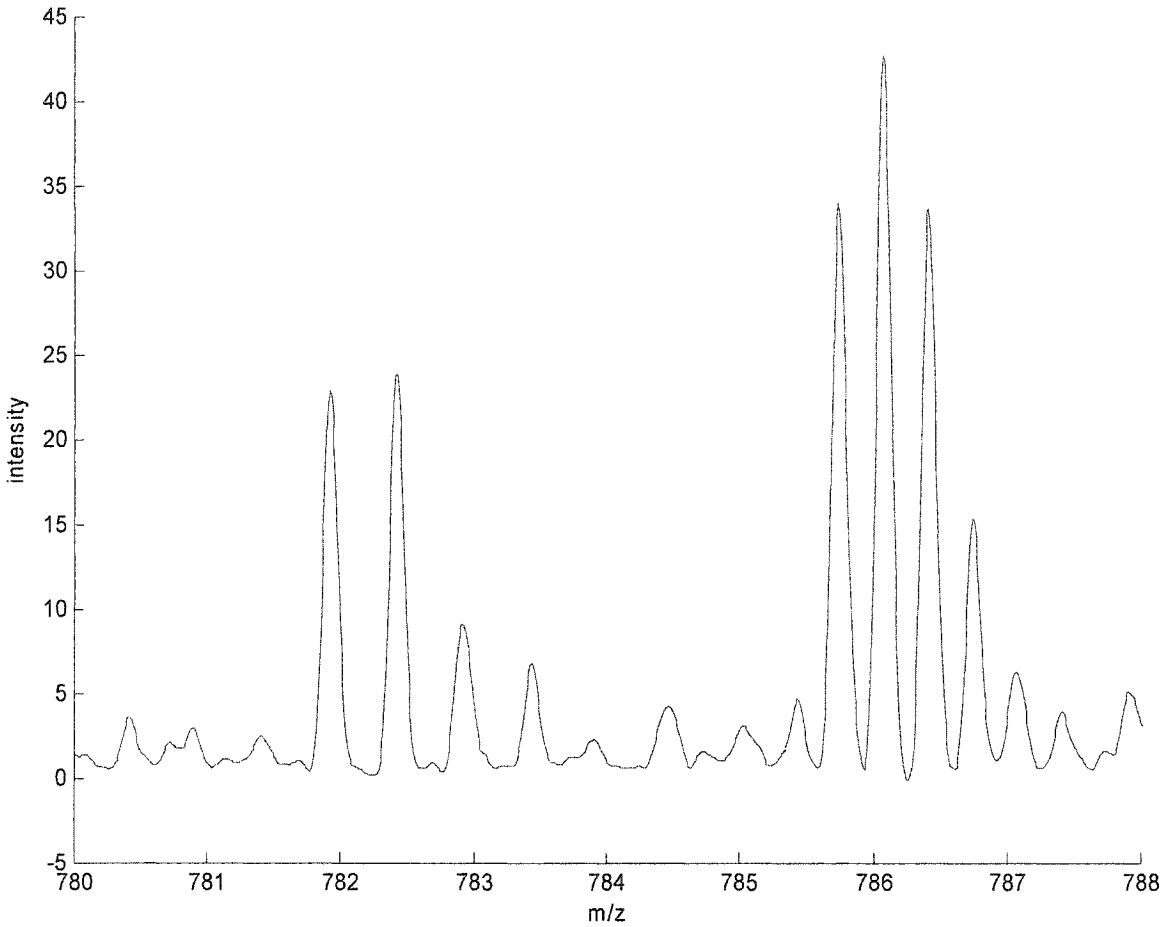

Figure 4-2: Spectrum in Figure 4-1 after filtering

4.2 Peak Picking

Peak picking was simulated using a custom heuristic algorithm that used experimentally determined peak properties as general rules for peak finding. They include:

1) the width of a peptide peak varies linearly with *m/z,*

2) small peaks within close proximity to a larger peak are hidden and cannot be resolved,

3) the most intense data point within the peak defines the *m/z* location.

The peak picking algorithm combines these "rules", and uses a divide and conquer approach to improve speed.

The algorithm (Appendix 9.1) begins by finding the most intense data point in the spectrum, and computes its expected peak width based on its *m/z*. The peak intensity and location is recorded; the peak and all data points within the calculated peak width are removed from the spectrum. The two pieces of the remaining spectra are processed in the same manner as if they were complete spectra. The algorithm concludes when all data points have been processed. All of the recorded peaks are combined to form a complete peak list. Figure 4-3 shows the result of applying the peak picking algorithm to the filtered spectra of Figure 4-2.

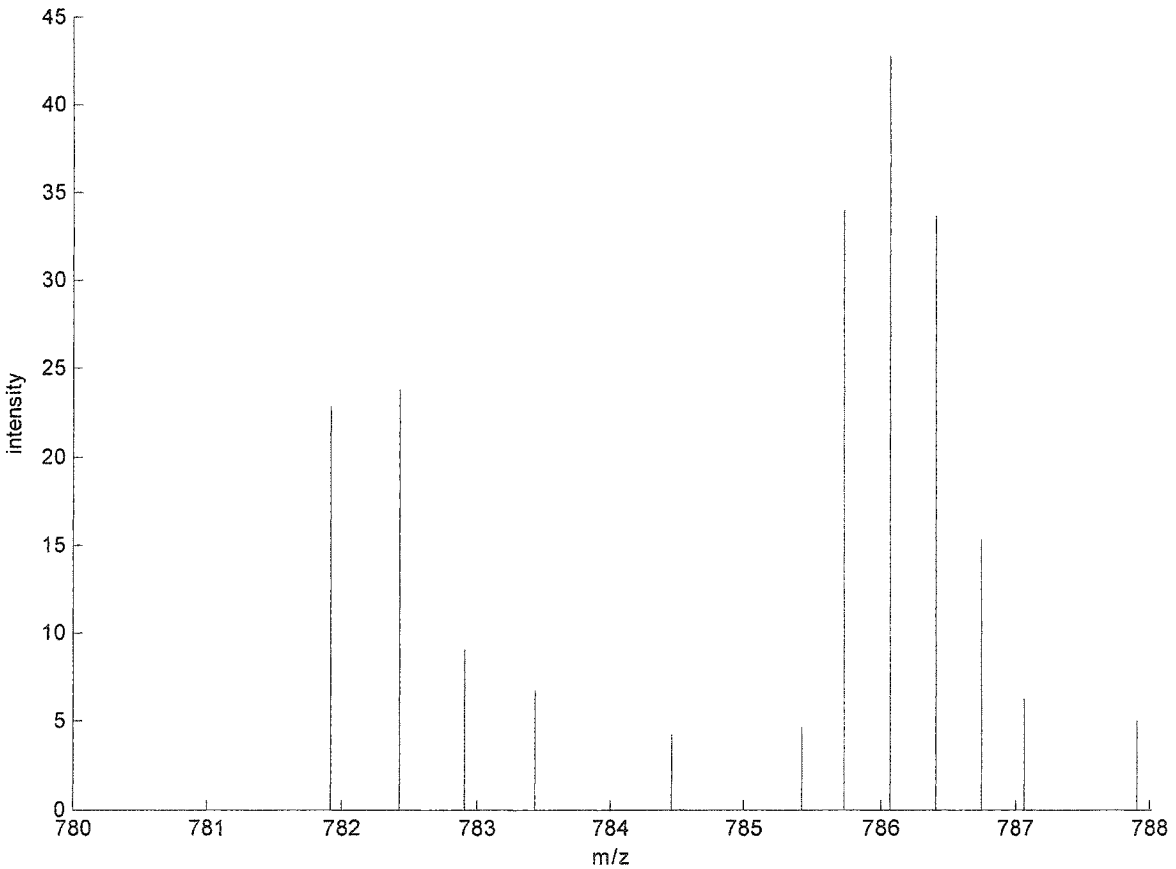

Figure 4-3: Spectrum in Figure 4-2 after applying the peaking algorithm

4.3 Deisotoping

Deisotoping was achieved using an extension of the Wehofsky [23] algorithm that was designed to deal with singly charged peptides from a MALDI source. Our algorithm (Appendix 9.2) isolates and merges isotopic peaks spread by multiple charges from an

ESI source. The algorithm makes the assumption that the data has been perfectly peak picked, and thus no peaks are erroneous. The application of an intensity threshold, or any other means of peptide filtering, can be used to clean the data before or after deisotoping. In this case, the spectral filtering algorithm was used to remove erroneous peaks. The algorithm scans the *m/z* axis sequentially, taking the first peak it encounters, and determines if there are any peaks in the vicinity which might indicate the presence of peptides. The presence of peptides would be attributed to a set of correctly spaced peaks. The inter-set spacing would define the charge state of the peptide.

For a given charge state, a theoretical average isotopic distribution exists which dictates the relative intensity and spacing of the peaks within a peak set. The average isotopic distribution is generated using the Breen method and data.[15] For a given mass, the theoretical distribution of atoms defines the relative intensities of the spectral peaks. Small deviations (~20%) of peak height from theoretical distributions can be attributed to molecules containing atomic compositions that vary from the average.[16] Larger deviations from this theoretical distribution indicate that the identified mass/charge state does not adequately describe the peptide, or that more than one peptide is present in the considered set. If the peak set adheres to the theoretical average then a peptide is detected and recorded, the peak set is removed from the spectrum, and the processing of the spectrum continues.

Figure 4-4 displays the result of deisotoping the sample spectrum. Two large peptides are detected, one doubly charged and one triply charged. Several smaller singly charged peptides are associated with peaks that could not be attributed to isotopic clusters. A small triply charged peptide is also located directly ahead of the large peptide, which the algorithm predicts is an overlapping peptide since its spacing is consistent with a triply charged peptide. It will be shown in using surface intensity analysis that this is the not the case, but that the peak is associated with a singly charged cluster that has just begun to elute.

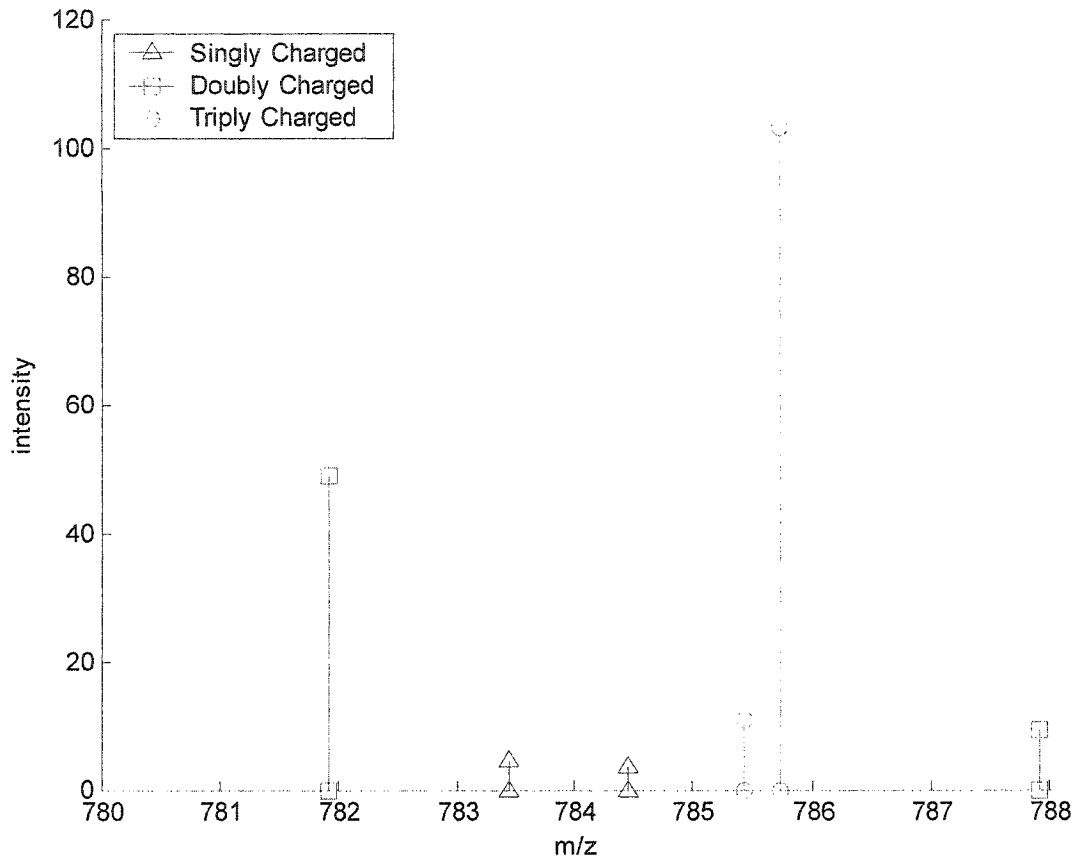

Figure 4-4: The spectrum in Figure 4-3 after applying the deisotoping algorithm.

The deisotoping process generates a list of possible peptides containing their location in *m/z* and their charge state. However, the mass spectrometer only considers doubly or triply charged peptides for fragmentation because:

1) these peptides generally have longer amino acid sequences which aids in protein identification,

2) doubly charged peptides require less energy to fragment in the mass spectrometer, and

3) all of the fragments of multiply charged peptides retain a charge, and thus contribute more information to the fragmentation spectrum.

Consequently, the list of possible peptides was pruned to contain peaks associated with doubly or triply charged clusters.

4.4 Validation

To gauge the quality of the algorithms presented, they were tested to determine their response to extreme noise conditions. To verify the accuracy of the algorithms, an 'ideal' spectrum was created containing a peak cluster defined by the virtual peptide 'ACDEFGHIKLMNPQR'. The peaks of the ideal spectrum were then convolved with a Gaussian curve to simulate the spreading of peaks within the mass spectrometer. By design, the sum of the intensity of all peaks in the cluster was 100. When the algorithm processed the noise-free spectrum, 89% of the peptide intensity was accounted for. The missing 11% can be attributed to the several peaks falling below threshold.

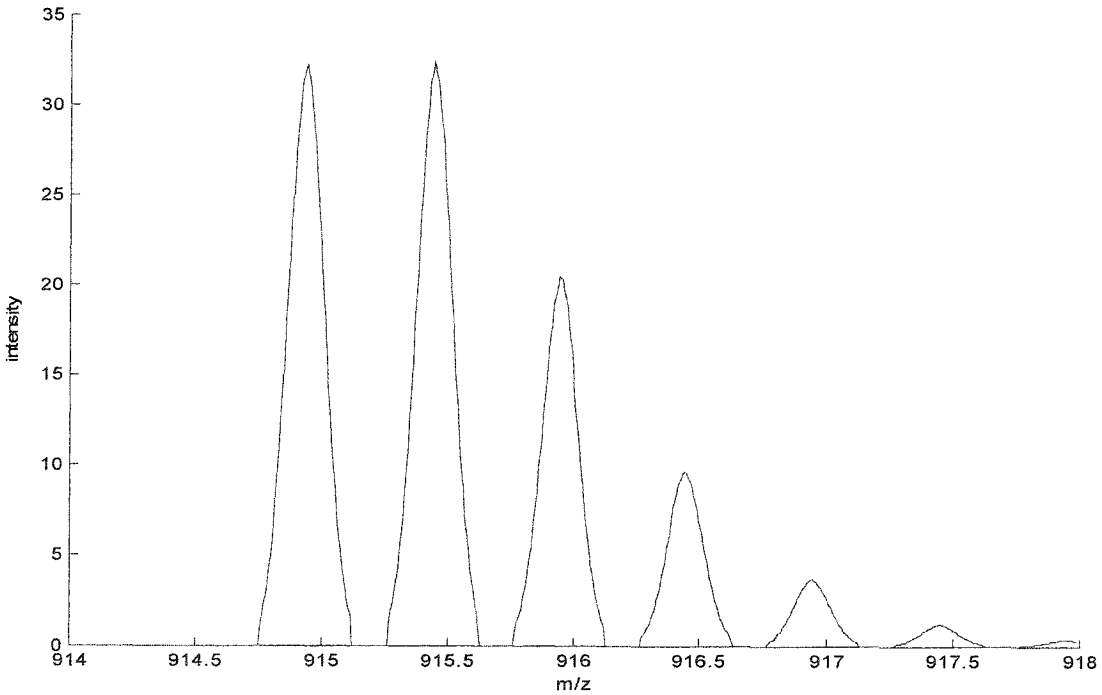

Figure 4-5: The ideal spectrum used to validate peptide detection algorithms

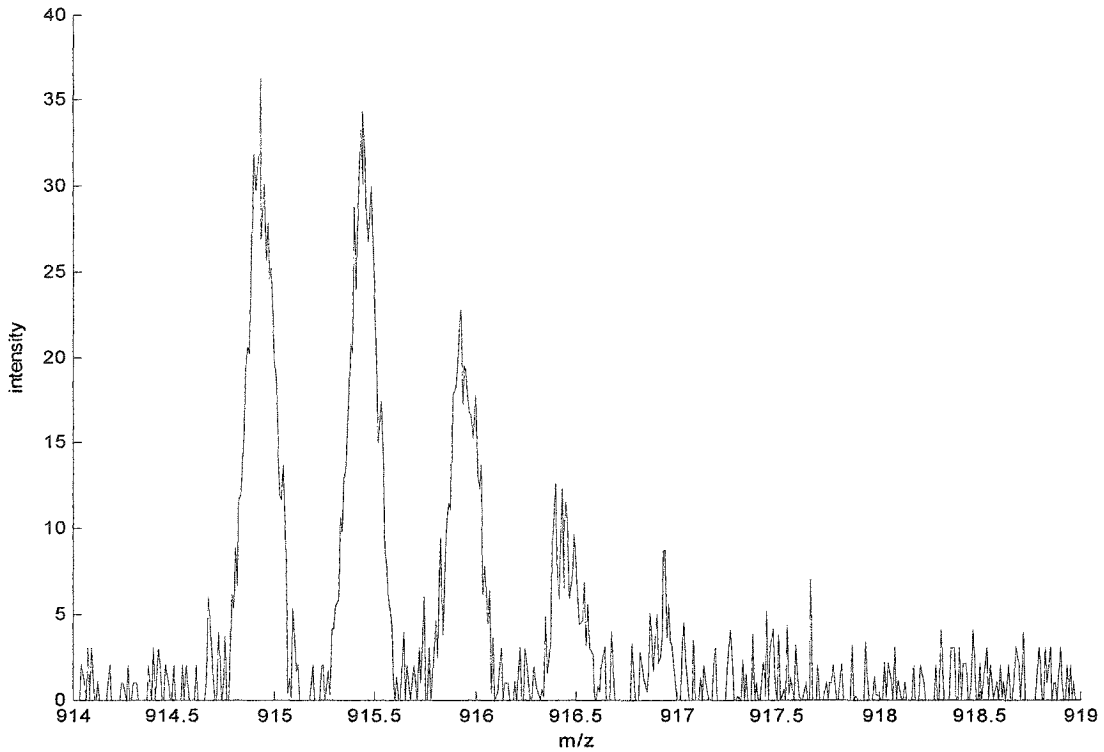

Figure 4-6: The spectrum of Figure 4-5 with a signal-to-noise ratio of ~4.6

The addition of noise to the spectrum was simulated 2000 times with the signal-to-noise ratio varying randomly from 1.4 to 11. The signal-to-noise ratio in mass spectrometry is defined as the height of the largest peak in the cluster relative to the maximum intensity of the noise. The minimum signal-to-noise ratio in mass spectrometry is 2, but minimum limits of 5 to 10 are more generally applicable. It is not uncommon to see ratios as high as several hundreds, and even thousands. The algorithms processed the noisy spectra in an attempt to find the correct peak location, charge and intensity. The histogram of Figure 4-7 shows the percentage of intensity accounted for in the 2000 simulations. The majority of the simulations show intensity detections close to 90% similar to the noise free spectrum.

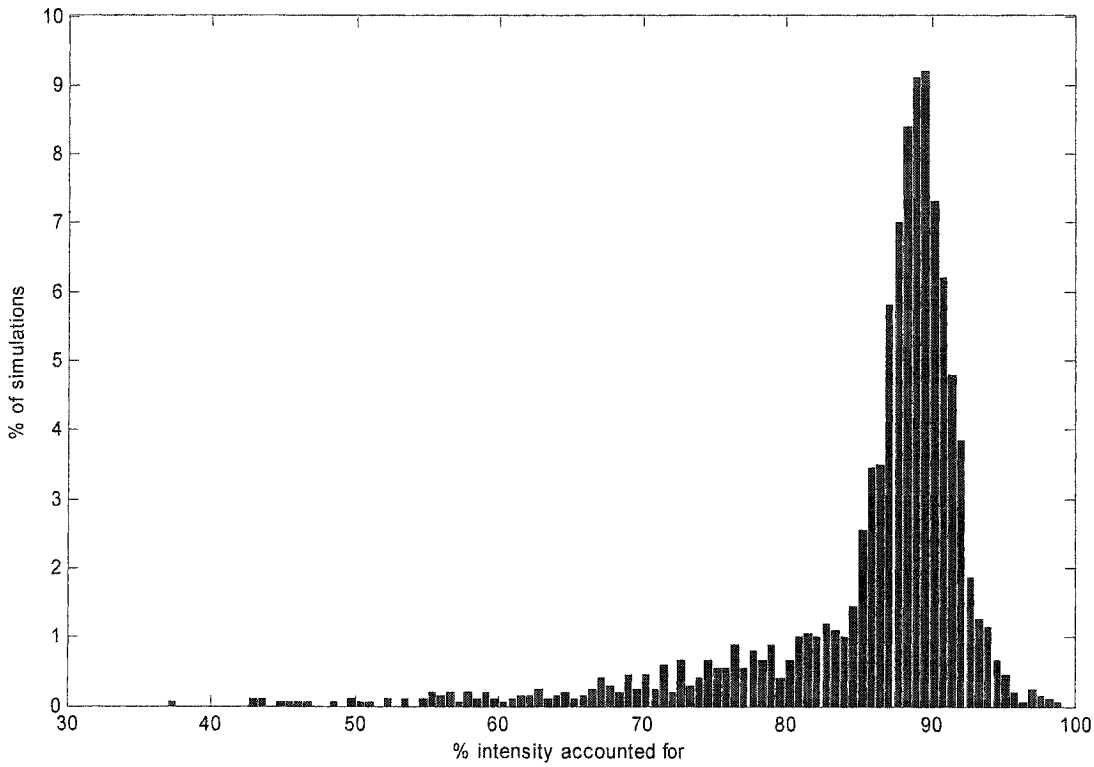

Figure 4-7: Histogram of intensity accounted for in 2000 noise simulations

While a large percentage of the simulations were able to account for most of the signal intensity, a small minority of outliers appear to be problematic. However, when searching protein databases for sequence matches, the mass and charge of a peptide are more important criteria than its intensity [24]. The same simulations show (Figure 4-8) that the peptide was always located within 0.04 m/z of its target, an error comparable to the calibration error of the mass spectrometer. In all cases the peptide's charge state was determined correctly.

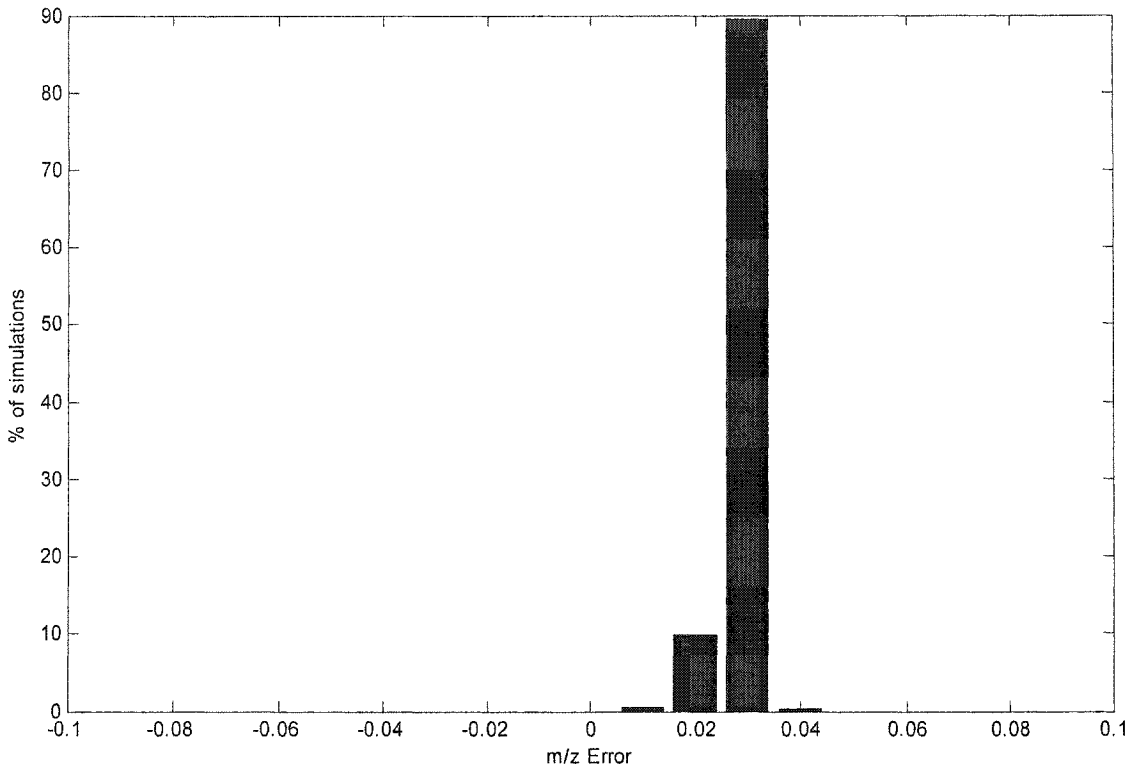

Figure 4-8: The m/z error of 2000 simulations

4.5 Fragmentation Simulation

Fragmentation simulation mimics the mass spectrometer processing of deisotoped data for a set of switching-parameters. The switching-parameters are varied such that each feasible parameter combination is generated. Each parameter combination is then simulated.

The peak list is searched, one MS spectrum at a time, for peaks whose deisotoped intensity is above the mass spectrometer minimum threshold. If no peptides are found, the process continues to the next scan. If peptides are present, they are sorted by intensity and individually added to the list of fragmented peptides if: 1) the same peptide does not appear on the list within the user specified time, and 2) the number of peptides on the list is less than that allowed by the duty cycle.

The repetition rate dictates the replication of newly added peptides on the list. Finally, the number of peptides added to the list in a single scan regulates how many scans need to be skipped, since skipping represents the time necessary to acquire the fragment scans. The list of fragmented peptides contains the scan number and the *m/z* of the peptides.

Figure 4-9 illustrates the selection and fragmentation of peptides following a 4-5 duty cycle. First, a MS scan is taken (circle), from which four peptides are selected for fragmentation (MS/MS 1-4). The fragmentation spectrum for each peptide is then taken once. The signal strength allows for the peptides to be repeatedly fragmented (27.9 to 28.15 min). If the fragmentation spectrum signal becomes too weak, the mass spectrometer stops acquiring spectra for the weak peptides (MS/MS 1, 3 &4). When the limit of 5 repetitions is reached (MS/MS 2), the cycle is over, and the mass spectrometer captures another MS spectrum to find new peptides. *Note: The intensities in the graph are meaningless.*

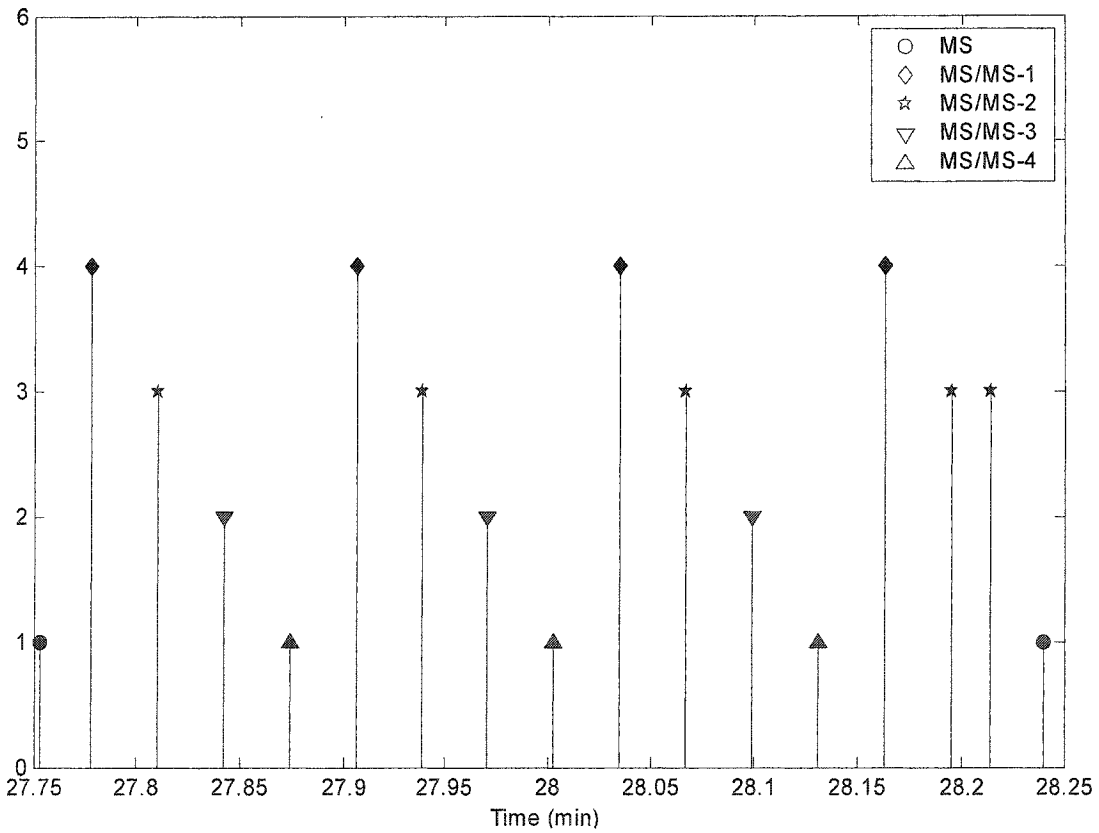

Figure 4-9: Implementation of a 4-5 switching scheme on a real sample

5 Surface Intensity Analysis

The mass spectrometer evaluates spectra one at a time, in less than one second, to estimate the presence and location of peptides. However, during offline processing, the complete data set can be used to yield improved results by using all available spectra.

Current peptide detection algorithms generally use two dimensional (*m/z* and intensity) techniques for identifying peptides either in a single scan, or in the average of several scans. The liquid chromatographic stage of the LC-Q-TOF introduces a third dimension, retention time. The set of acquired MS scans can therefore be regarded as a three dimensional dataset where intensity is a function of *m/z* and time. This dataset can be processed using image processing methods similar to those used to identify "spots" in images of 2D gels.

5.1 Uniform Resampling

For the MS data to be interpreted as an image, it must be uniformly resampled to give equally spaced samples. The time data is uniform to start with since scans are taken every second. However, the *m/z* domain is not uniform since TOF mass spectrometers acquire data in the time domain and apply a non-linear transformation to determine *m/z*. Moreover, points that do not contain information are omitted from the spectrum (to save space) resulting in very non-uniform samples. To have uniformly sampled data, the *m/z* domain can be transformed back to the native time domain, or resampled into a uniform *m/z* domain. The latter option was chosen since most of the processing will be done in the *m/z* domain. Consequently, a linear interpolation algorithm was used to resample the *m/z* domain of the raw data at constant 0.01 *m/z* increments. Linear interpolation was chosen for both its speed and its predictability. Higher order interpolation is not useful since the derivatives that are preserved with these schemes are not needed. Furthermore, the derivatives cannot be computed accurately because the data is too noisy.

5.2 Spectral Stacking

The uniformly resampled scans were stacked, to form a 2D matrix with *m/z* and time as its axes. This two dimensional matrix was then manipulated as an image where the spectral intensity is represented by greyscale intensity. Figure 5-1 illustrates the results of stacking 80 consecutive scans. Four unique peptides (isotopic peak clusters) are clearly visible, while a fifth (at ~782, 1500) is visible close to background intensity.

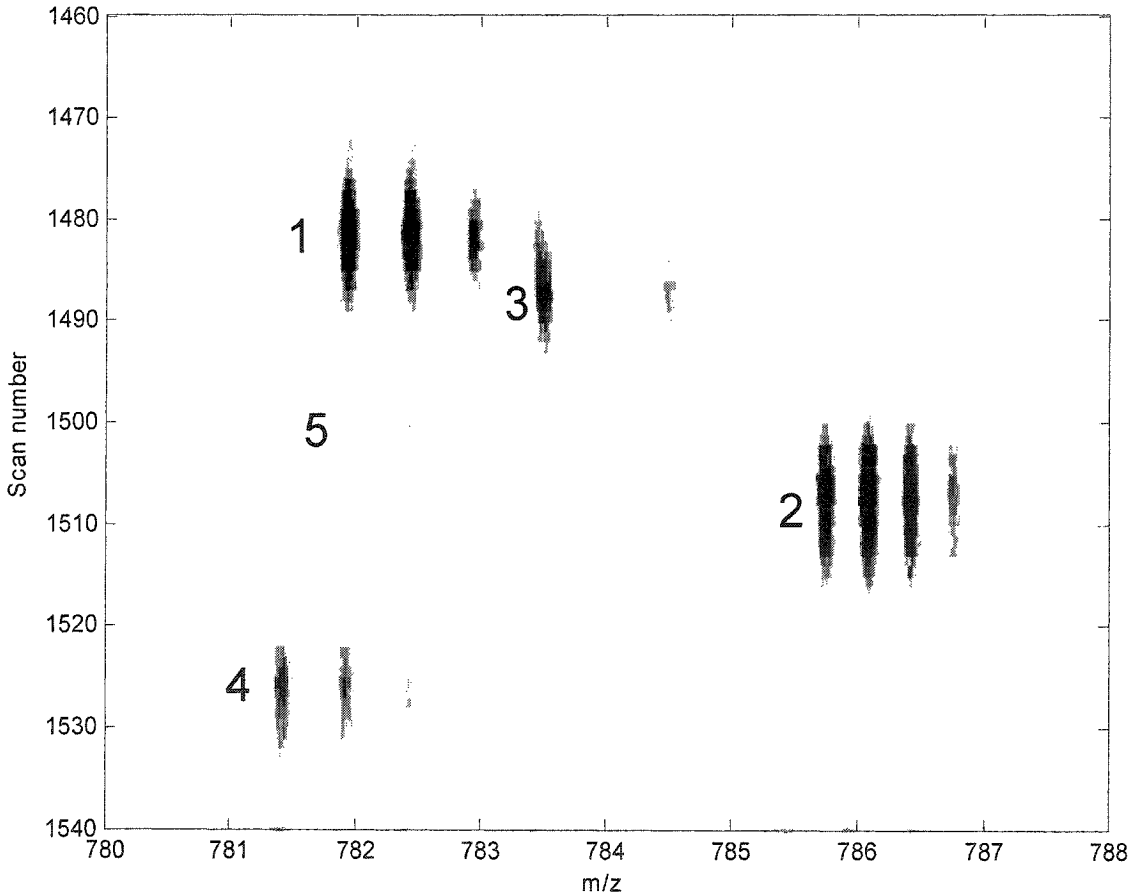

Figure 5-1: Subset of stacked spectra, scan number 1500 defines Figure 4-1

5.3 Surface smoothing

A surface smoothing filter was used to improve the signal-to-noise ratio of the matrix prior to peak picking. High frequency noise poses the major problem for peak picking since it may introduce new local maxima. Consequently, the image was filtered with a 4x4 square averaging filter, which is a low-pass filter with a cutoff frequency of 0.116

cycles/S (cycles per sample). The frequency content of peptide peaks vary with *m/z*. Gaussian models of these peaks with a minimum width of 12 samples (from peaks at 175 *m/z*) contain frequency content up to 0.0581 cycles/S; wider peaks have lower frequency content. In the time domain, peaks generally last for a minimum of 12s, thus leading to frequency content up to 0.0581 cycles/S. The filter cut-off frequency is sufficiently higher than the highest frequency in the signal, ensuring that only noise is attenuated while the signal remains unchanged. To increase the order of the filter, and hence improve attenuation of noise signals, the filter was applied three times. This procedure also has the effect of decreasing the cut-off frequency to 0.0686 cycles/S, which is still high enough to avoid signal attenuation. Figure 5-2 demonstrates the results of smoothing the sample spectra. The background intensity is decreased, while the peptide 'spots' are better resolved and therefore are identified with less ambiguity.

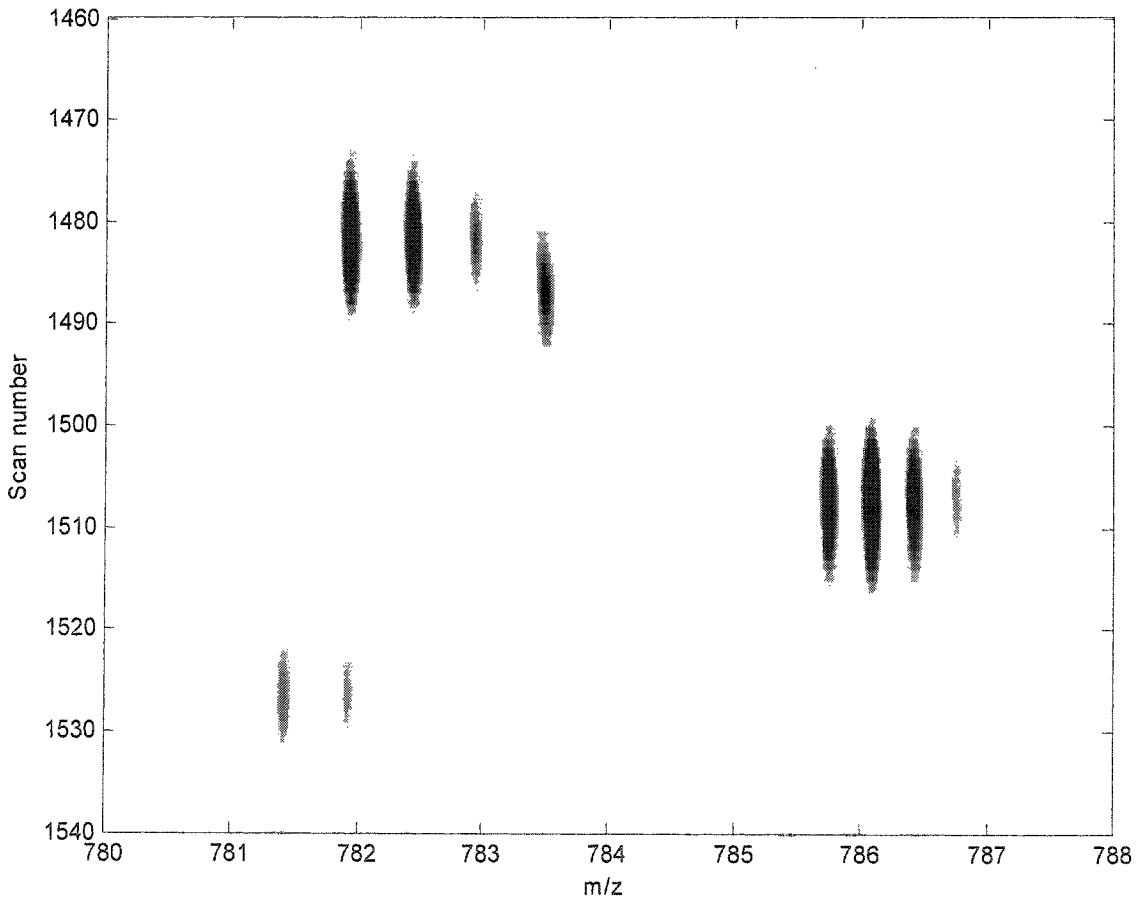

Figure 5-2: Smoothed spectra of Figure 5-1

5.4 Surface maxima

A peak picking algorithm is required to localize peptide peaks within the two dimensional matrix. Since surface smoothing removes most noise, the peptide peak is expected to be the highest point on the peptide profile, which can be easily identified as any local maximum. Our algorithm used an 8-connected local maxima algorithm to ensure that the point chosen is larger than all adjacent matrix entries, including the diagonals. This two dimensional peak picking returns a list of peaks with *m/z*, time and intensity co-ordinates. Figure 5-3 illustrates the peaks selected from the sample spectra. The white circles indicate the location of the peptide 'spot' maximum intensity.

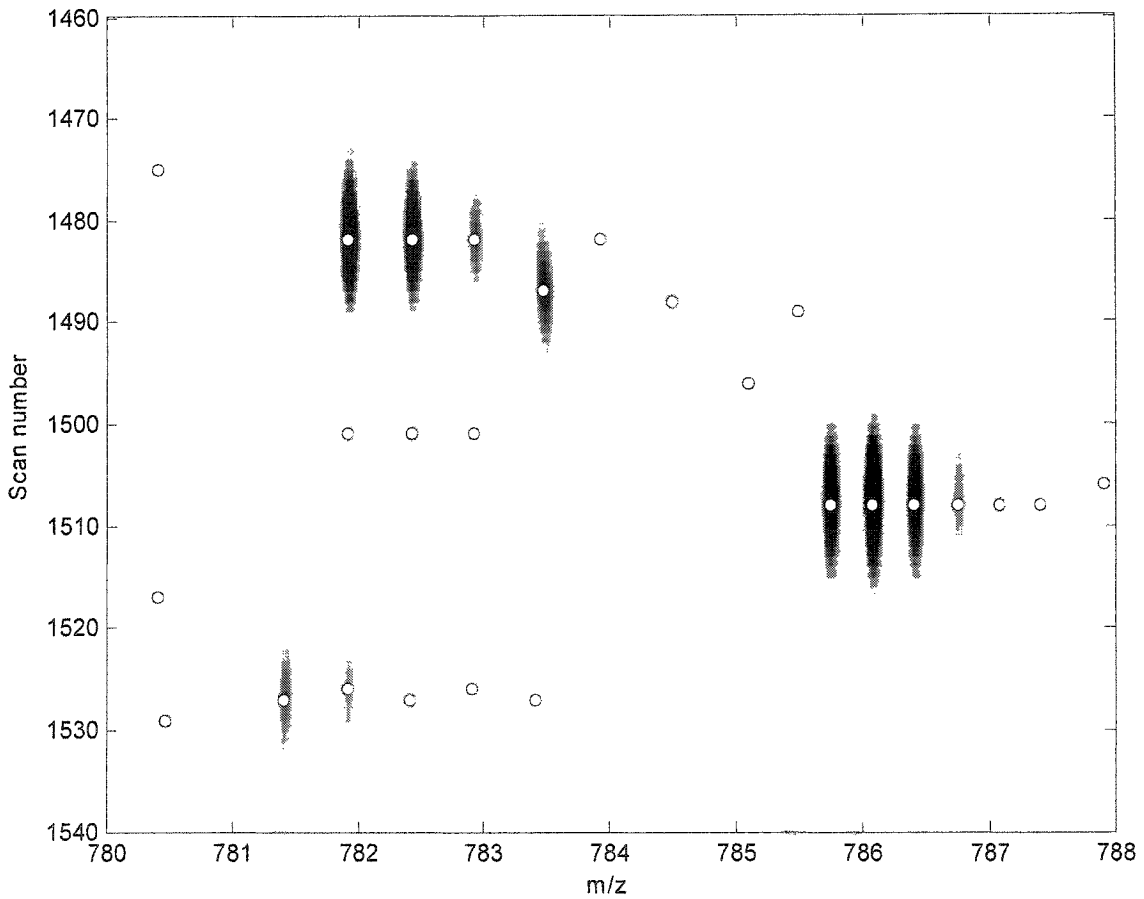

Figure 5-3: Peak picked spectra of Figure 5-2

5.5 Deisotoping

Deisotoping a surface is much simpler than traditional single spectrum deisotoping as isotopic distributions need not be modeled, and overlapping peptides are resolved by their time profiles. Peaks are grouped into sets, composed of peaks which appear within a one second interval (i.e. collinear on the x-axis), with an *m/z* spacing consistent with a specific charge state. Each set defines a peptide and its isotopic distribution. The sets can be collapsed to the mono-isotopic peak's parameters: charge state, *m/z,* and elution time.

Figure 5-4 displays the result of deisotoping the identified peaks of Figure 5-3. The algorithm correctly identified all five peptides that were identified manually, as well as a sixth that is only partially displayed.

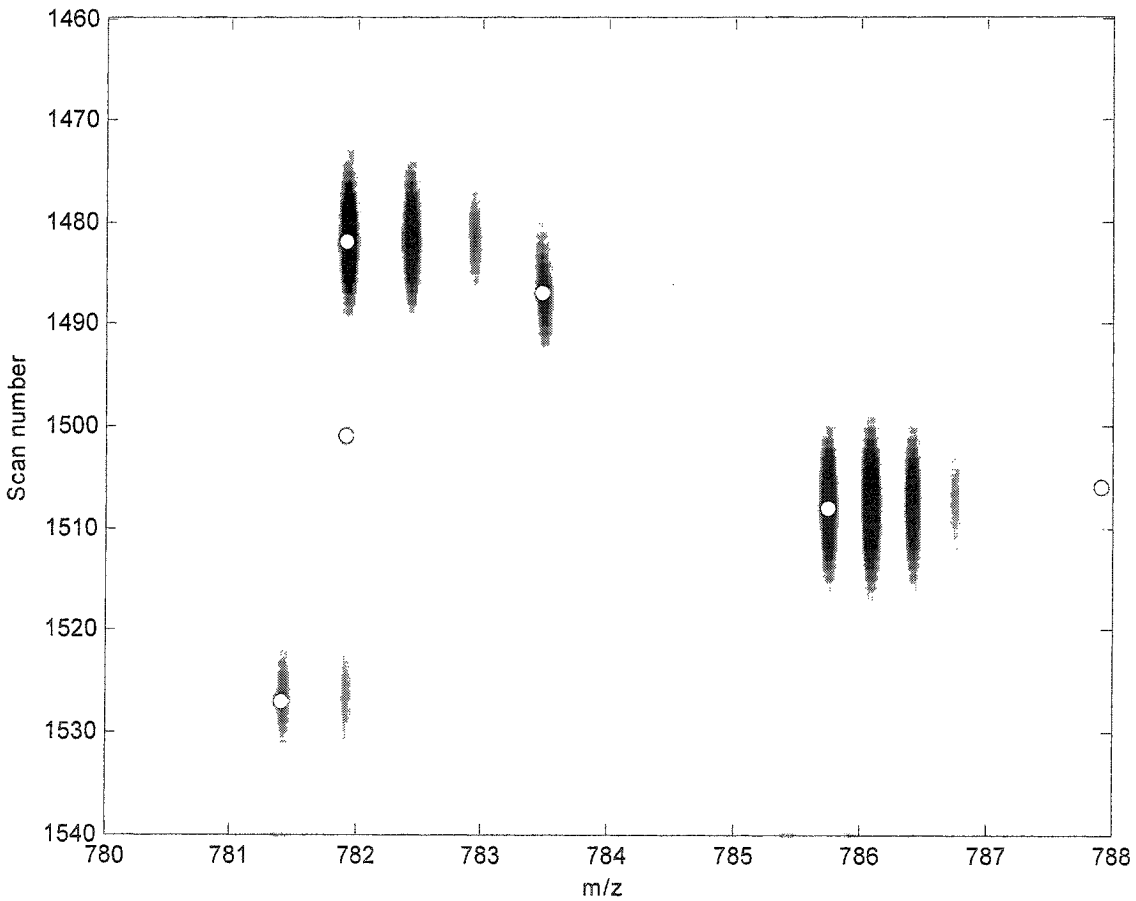

Figure 5-4: Deisotoped peaks of Figure 5-3

The surface intensity analysis algorithms easily localize peptides and their mono-isotopic peaks. The results are less prone to error as filtering attenuates noise in two dimensions instead of just one. The spectrum of Figure 4-1 is the 1500^{th} scan of Figure 5-4. Of the six peptides identified by traditional algorithms, two were in error. These peptides were low in their elution profile, causing their peak intensities to be small. Using surface intensity analysis, the identification of peptides is made at the top of the elution profile, thus making best use of the available information. The result of the analysis is the set of "all" identified peptides whose peak locations and profiles were used as a reference to grade traditional algorithms.

6 Performance Evaluation

Ideally, the mass spectrometer would fragment every available peptide at the apex of its peak profile. Thus, the evaluation of the switching parameter simulations was based on two criteria: 1) The number of peptides fragmented, and 2) the magnitude of the fragmentation location relative to the apex.

The mass spectrometer simulations produce lists of peptide fragments containing coordinates in terms of scan number (time) and *m/z* that are used to find the intensity of the peptide on the intensity surface array. Fragmentations are considered valid if the surface intensity is above the spectrometer's fragmentation threshold; otherwise it is considered an error for the simulation algorithms. If valid, the peptide intensity is used as the intensity score for that fragmentation.

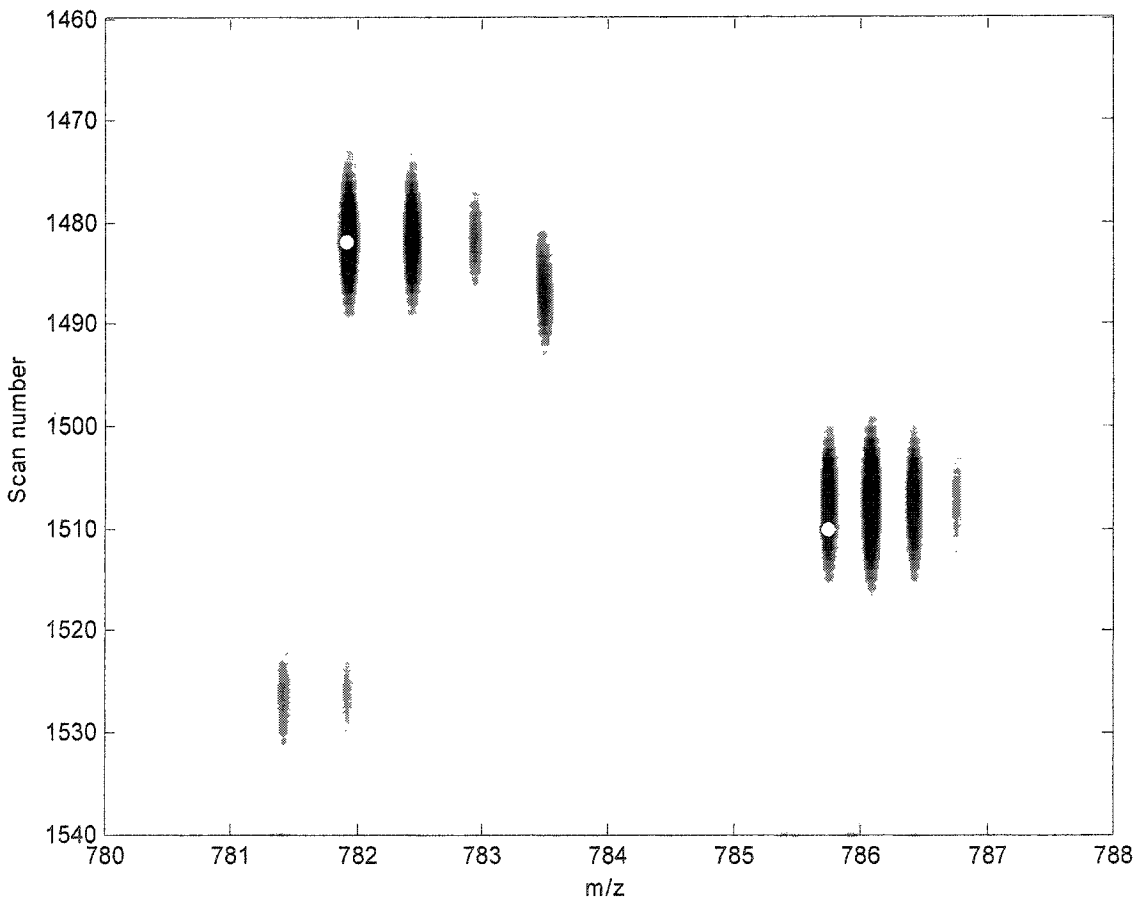

Figure 6-1: A 1-1 duty cycle simulation. (White dots show fragmentation location)

The white dots in Figure 6-1 indicate where fragmentation would have occurred had the mass spectrometer followed a 1-1 duty cycle. Fragmentations are close to the center of the peak profile in both axes indicating a quality fragmentation location. Unfortunately, only two of the six peptides detected using intensity surface analysis were fragmented (see Figure 5-4), as there are likely peptides with higher intensities outside the area depicted in the graph.

Nearly all of the simulation experiments allow a peptide to be multiply fragmented. In tandem mass spectrometry, it is customary to average these repeated spectra to improve the signal-to-noise ratio before further processing. To mimic this signal improvement in our simulations, intensity scores were summed and divided by the square root of the number of repetitions. This incorporates the improvement of signal quality caused by spectral averaging without creating a bias towards large numbers of repeated fragment scans. Figure 6-2 illustrates the simulation of a switching parameter set that allows up to three repetitions (1 second each), contrasting the one repetition of Figure 6-1. Here, only the most intense peptide was fragmented. The fragmentation location is further from the apex of the peak, closer to the end of the peak profile. This will lead to poorer quality fragmentation spectra.

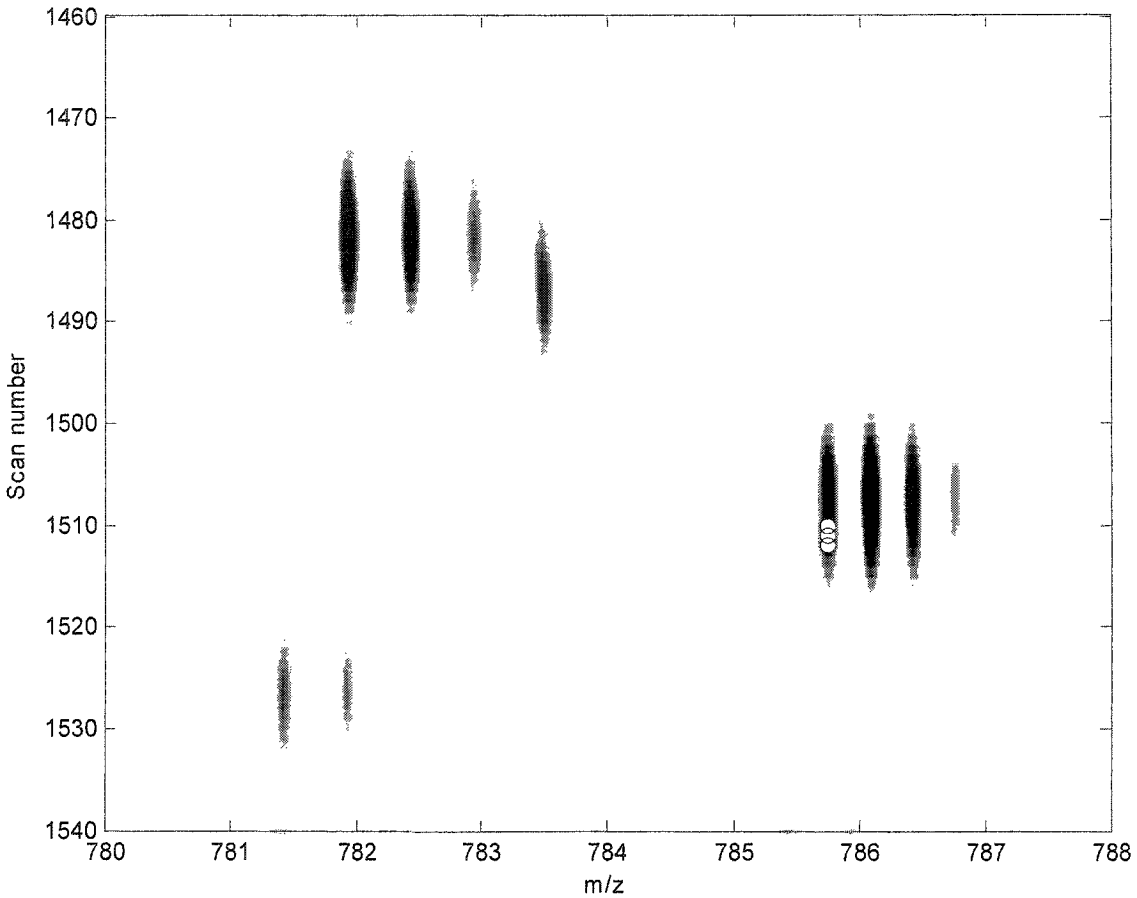

Figure 6-2: A 1-3 duty cycle simulation. (White dots show fragmentation location)

7 Results

The peptide intensity scores were analyzed in two ways. First they were compared to the maximum possible peptide intensity, as determined by intensity surface analysis, to gauge absolute quality. And second, the intensity scores were averaged for each switching parameter and all sets compared against one another to determine which performed best.

7.1 Quality versus maximum

The intensities where fragmentation occurred were first compared to the maximum intensity of the peptide to examine the strengths and weaknesses of various parameter sets. Ideally, all peptides would be fragmented near their peak intensity. Approximately 3300 peptides were detected via surface intensity analysis, however, no matter which switching parameter sets were used, the simulations only fragmented a small fraction of the available peptides.

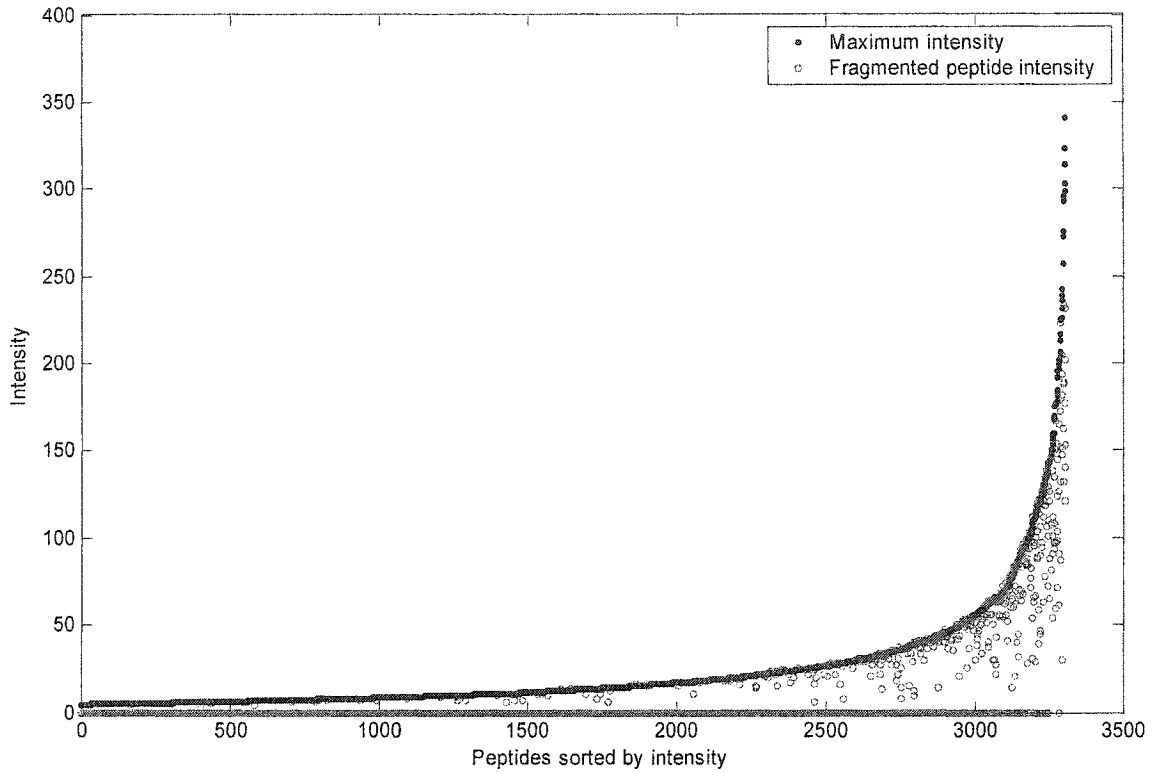

Figure 7-1: Peptide fragment and maxima intensity for a 1-1 switching parameter.

(465 peptides fragmented)

As an example, Figure 7-1 and Figure 7-2 display the maximal peptide intensity and the intensity where fragmentation occurred for the 1-1 and 5-4 switching parameters. In the 1-1 parameter set, more than twice as many peptides were fragmented than for the 5-4 parameters. This increase comes at a cost of intensity since the repetition rate trades off a number of peptides for higher intensity scores. For the 5-4 parameter, seven peptides had intensities of fragmented peptides higher than 250 whereas the 1-1 duty cycle had none. The repetition during the 5-4 parameter simulations allows for intensities higher than the maximum, which is not possible for a 1-1 parameter. Along the x-axis are the peptides that were not fragmented (~2800-3000).

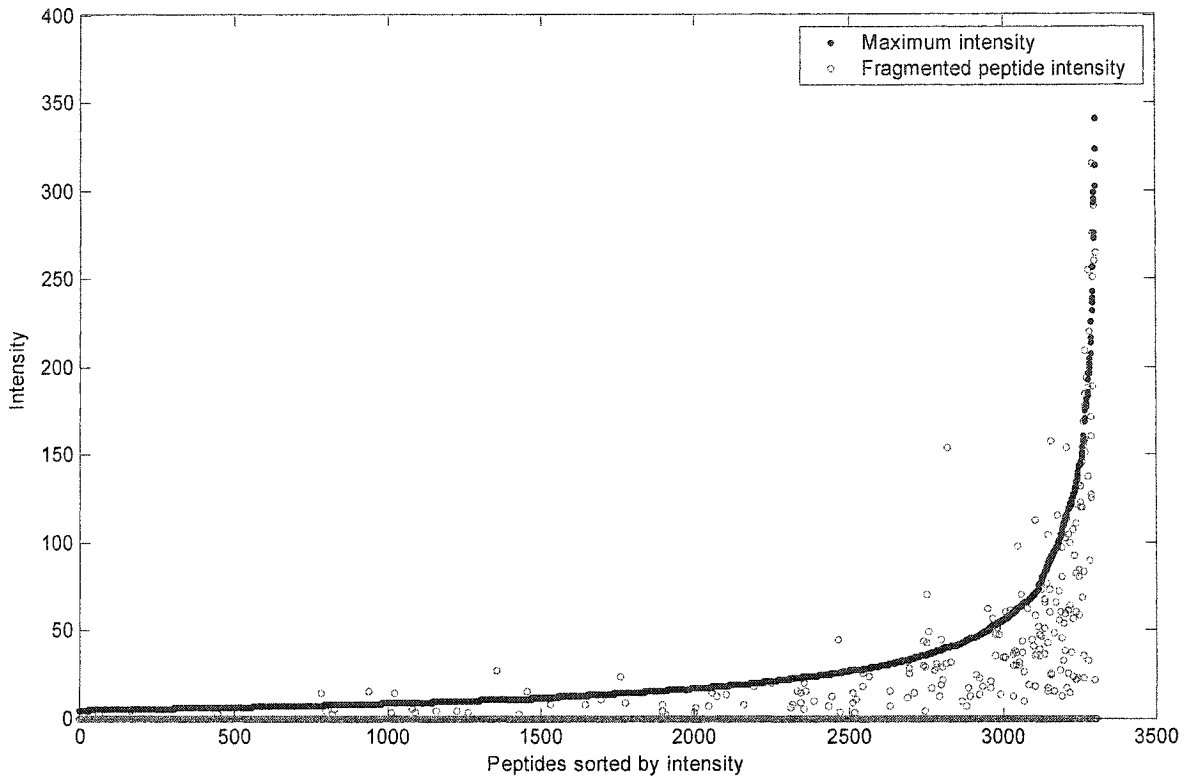

Figure 7-2: Peptide fragment and maximum intensity for a 5-4 switching parameter.

(231 peptides fragmented)

To compare the multiple operating parameter sets, the fragmentation intensities were normalized by the maximal peak intensity. Figure 7-3 compares various switching parameters with only one repetition. As expected, the number of peptides fragmented increases as the duty cycle increases, since less time is taken by MS scans, until a plateau of ~600 peptides is reached. The plateau suggests that, 1) there are only ~600 peptides available, or 2) that the gain from increasing the duty cycle is diminishing. The former is not possible since surface analysis has already determined that over 3000 peptides are available. High duty cycles are also prone to overlooking peptides, as illustrated by the string of zero intensity points along the x-axis for the 20-1 condition. This suggests that during the 10-20 seconds spent fragmenting peptides, the weaker intensity peptides have dropped below threshold.

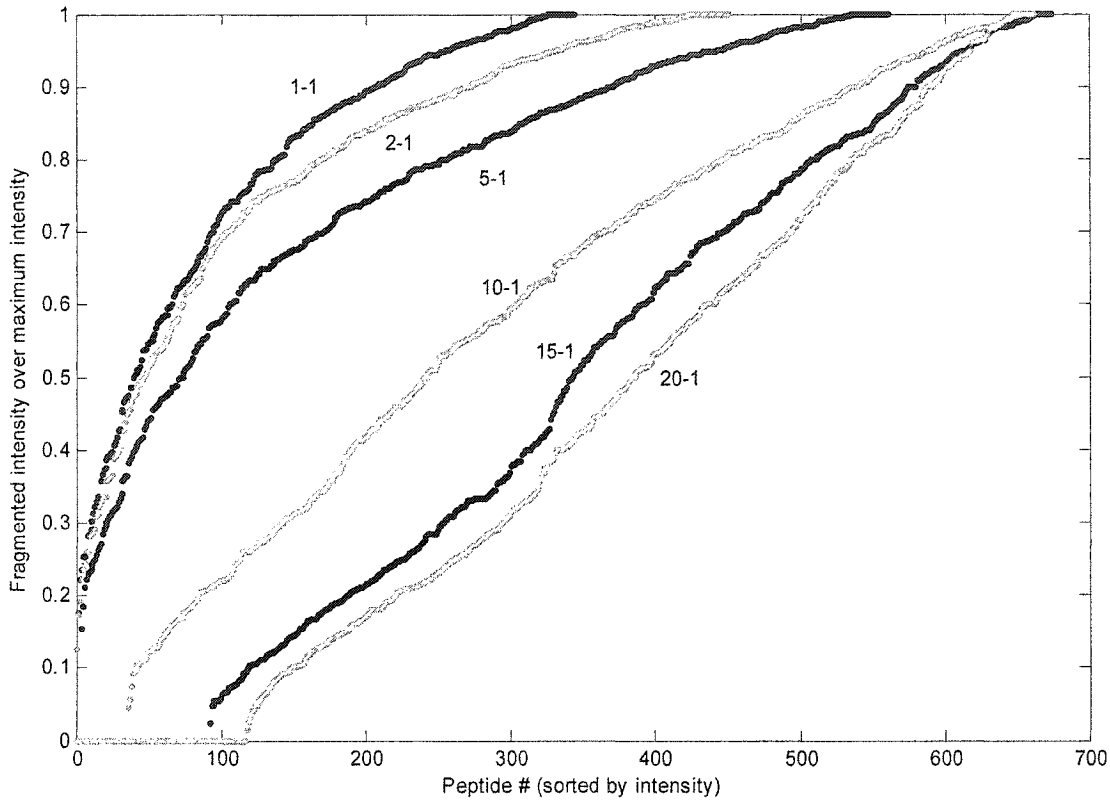

Figure 7-3: Normalized intensity scores for various duty cycle rates, at a repetition rate of 1.

A similar comparison of the various repetitions rates is illustrated in Figure 7-4. It is evident that raising the number of repetitions increases the intensity score of the peptides. However, the trade-off for increased intensity is decreased numbers of peptides fragmented. The increase in quality caused by higher repetitions is governed by the law of diminishing returns. The maximal intensity score achieved with two repetitions is approximately 1.5, while with five repetitions a maximum of 2.25 is reached. It is interesting to note that the percent gain in quality is similar to the percent loss in quantity.

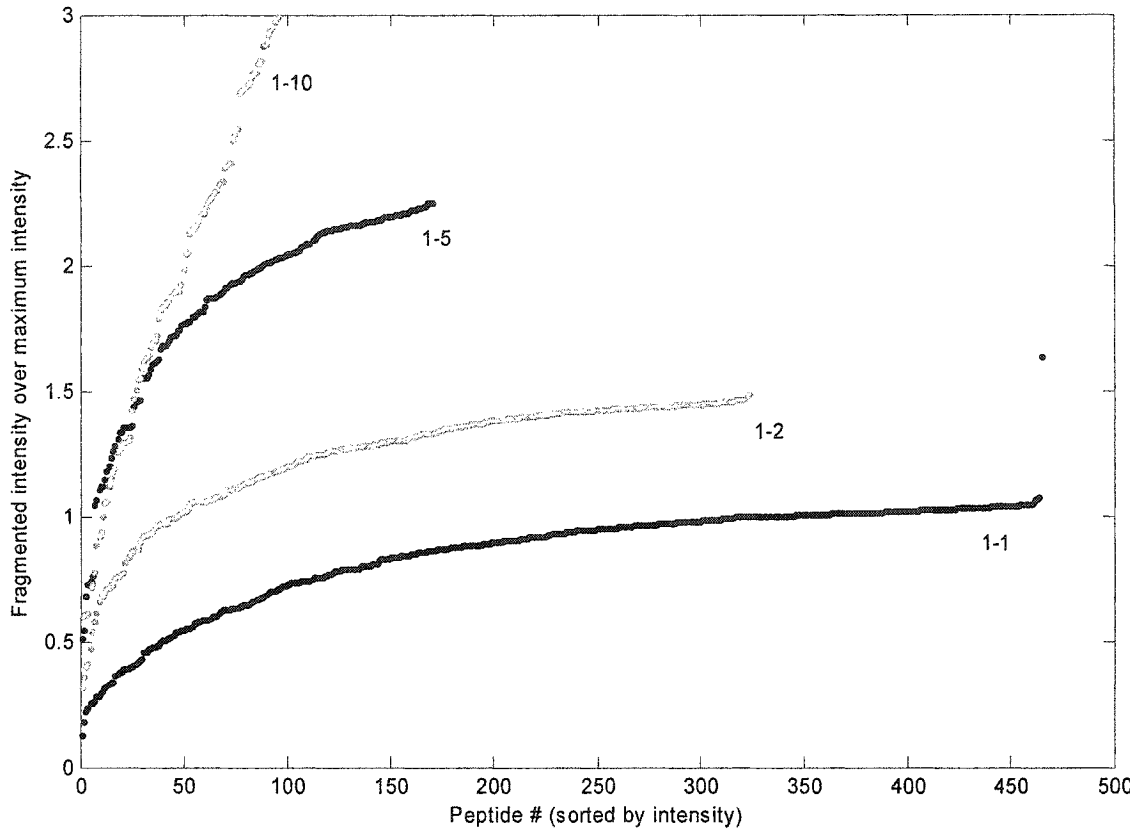

Figure 7-4: Normalized intensity scores for various repetitions rates, with a duty cycle of 1.

Figure 7-5 shows the data from Figure 7-3 reformatted as a histogram of normalized intensity. This histogram demonstrates that higher duty cycles (10-20 MS/MS per MS) result in fewer high scores and considerably more poor scores. The increased number of peptides fragmented in the simulations with high duty cycles is attributed to poorer quality spectra. However, the lowest duty cycle does not perform the best. From the selection of simulations shown, the 1-5 switching parameter performs best in the high quality range, is comparable to other parameters in the medium quality range, and has few poor quality spectra.

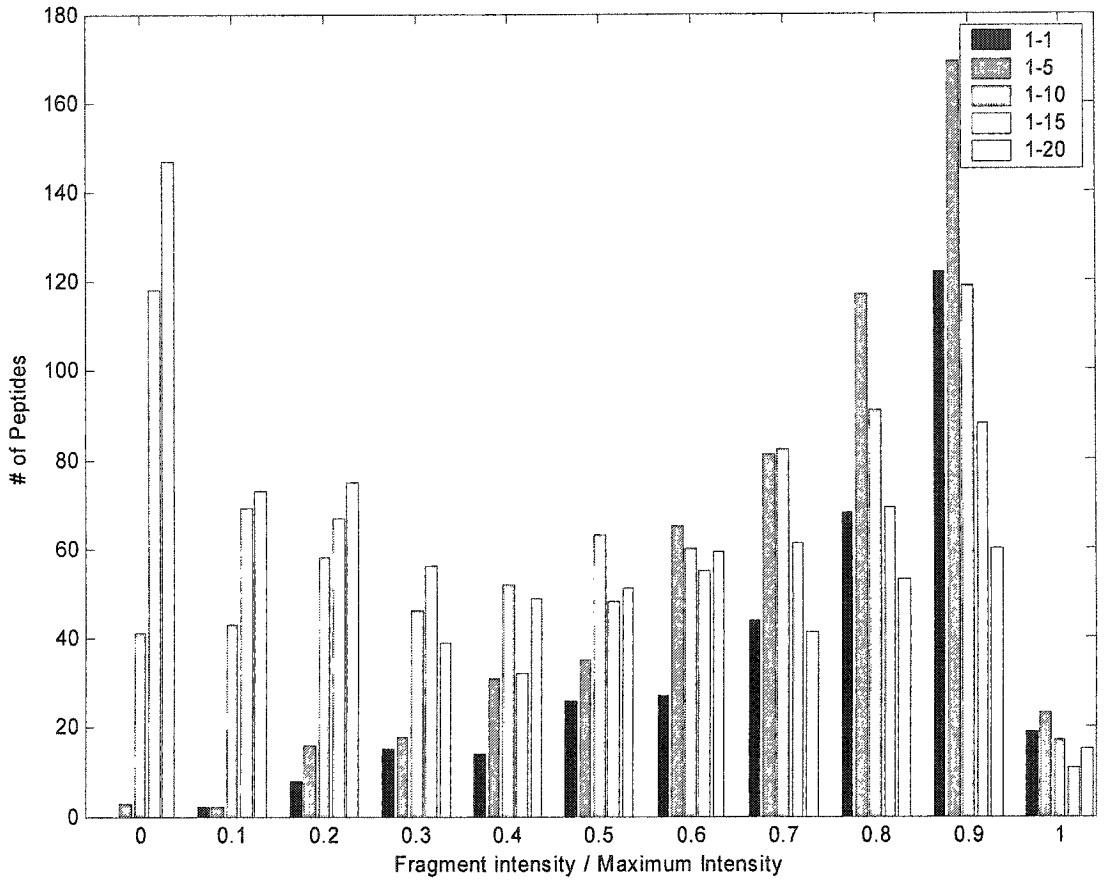

Figure 7-5: Fragment quality histogram for various duty cycles.

7.2 Comparative Quality

The foregoing comparisons were made relative to the intensity of the ideal peak. The peaks chosen in the various simulations are not necessarily the same so the comparison may be skewed. To better assess the quality of various switching parameters, we examined how quality varied from simulation to simulation to provide a better indication of optimal parameters.

Figure 7-6 shows the simulation result of parameter pairs for a 30 minute gradient. As expected, the number of fragmented peptides increases with an increasing duty cycle and decreases with repetitions. Also as expected, the fragmentation quality increases as repetitions increase and duty cycles decrease.

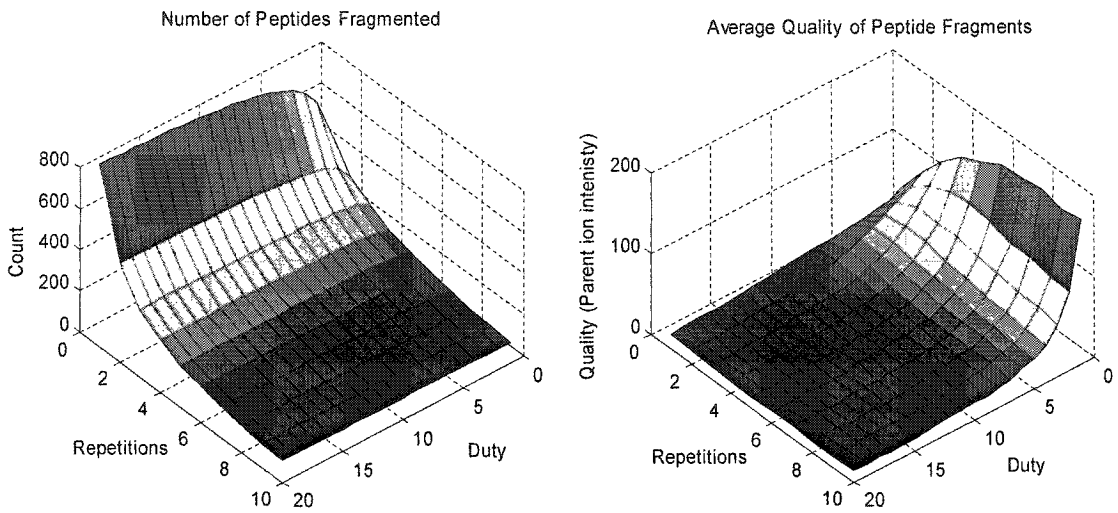

Figure 7-6: Quantity and quality of a 30 minute gradient

The mass spectrometer has a feature that stops repeating the fragmentation of peptides if any of the previous repetitions generate poor data. The simulations do not mimic this behaviour and thus there may be a bias against the score of high repetition switching parameters. Figure 7-7 illustrates the result of removing these poor quality spectra, i.e. fragmentations with intensities below the mass spectrometer's fragmentation threshold. This new graph shows the same plateau of approximately 600 peptides seen in Figure 7-3, and depicts a reduction in peptides fragmented at higher duty cycles.

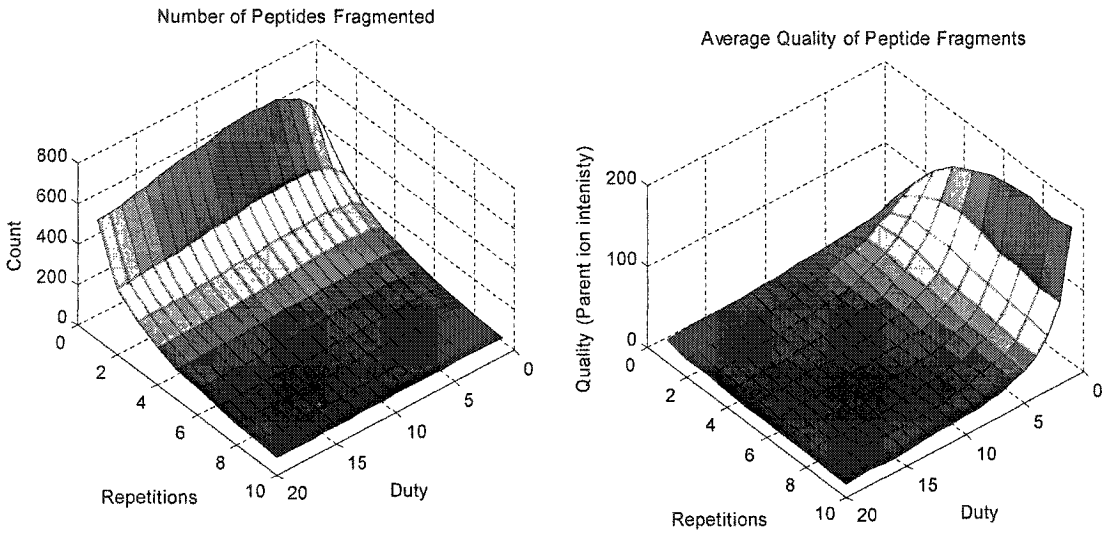

Figure 7-7: Quantity and quality of a 30 minute gradient (missed fragmentations removed)

Figure 7-8 illustrates the residuals from the subtraction of Figure 7-7 from Figure 7-6 and therefore represents the distribution of the fragmented peptides with poor quality spectra. The number of peptides removed by this process is rather drastic, but only in the high duty/low repetition area of the graph. This is contrary to expectations as the feature to stop repetitions is meant to assist the high repetition parameter sets. The average change in quality, however, is relatively minor and again only significant in the high duty/low repetition area of the graph. In the low repetition area of the graph, the non-repetition of poor spectra feature will have little to no effect on the simulation, therefore this area cannot benefit from this feature. The analysis of longer gradients (not shown) displayed similar trends. Thus, the simulations will not be biased and should provide valid results.

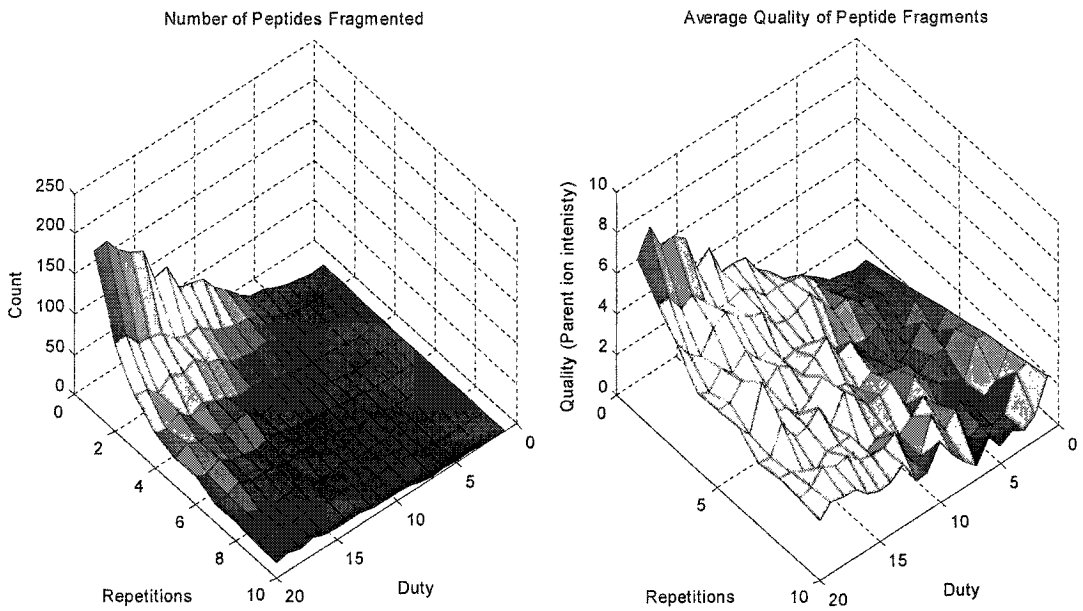

Figure 7-8: Residuals caused by subtracting Figure 7-7 from Figure 7-6

To find the optimal trade-off between the quality and quantity of peptides fragmented, a measure combining both criteria is required. To achieve this, the two data sets were multiplied together to generate a quality-quantity surface which is a measure of the total amount of information acquired during the various simulations.

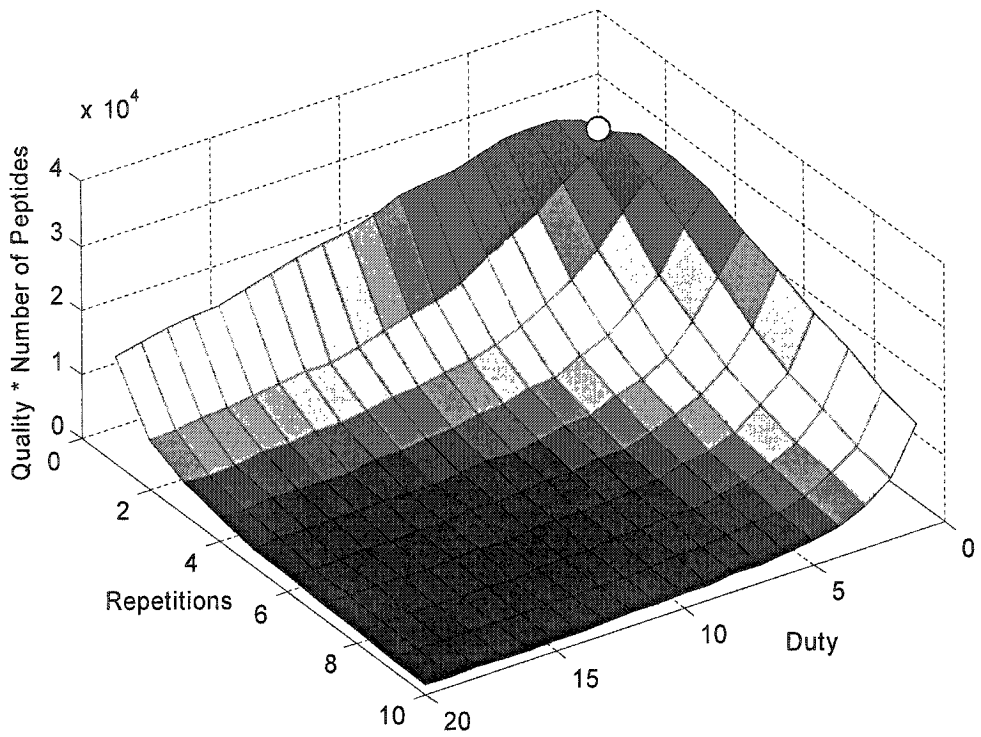

Figure 7-9: A 30 minute gradient with a circle indicating optimal operating point at 2-2.

The quality-quantity surface for a 30 minute gradient is illustrated in Figure 7-9. The graph displays the combined peptide quality-quantity index for the various duty cycles for a 30 minute gradient. The optimal switching parameter is at 2-2. All parameters near the low duty cycle - low repetition rate perform almost equally well; however the quality decreases rapidly when either duty cycle or repetition rate exceeds a value of three. For the 120 minute gradient of Figure 7-10, the optimal operating point is closer to 2-7, and the "sweet spot" corresponds to the high repetition – low duty cycle area. The shift of the optimal operating point towards more repetitions is expected; the longer gradient allows peptides to be available for a longer period of time; therefore, using more repetitions will not decrease the number of peptide fragmentations.

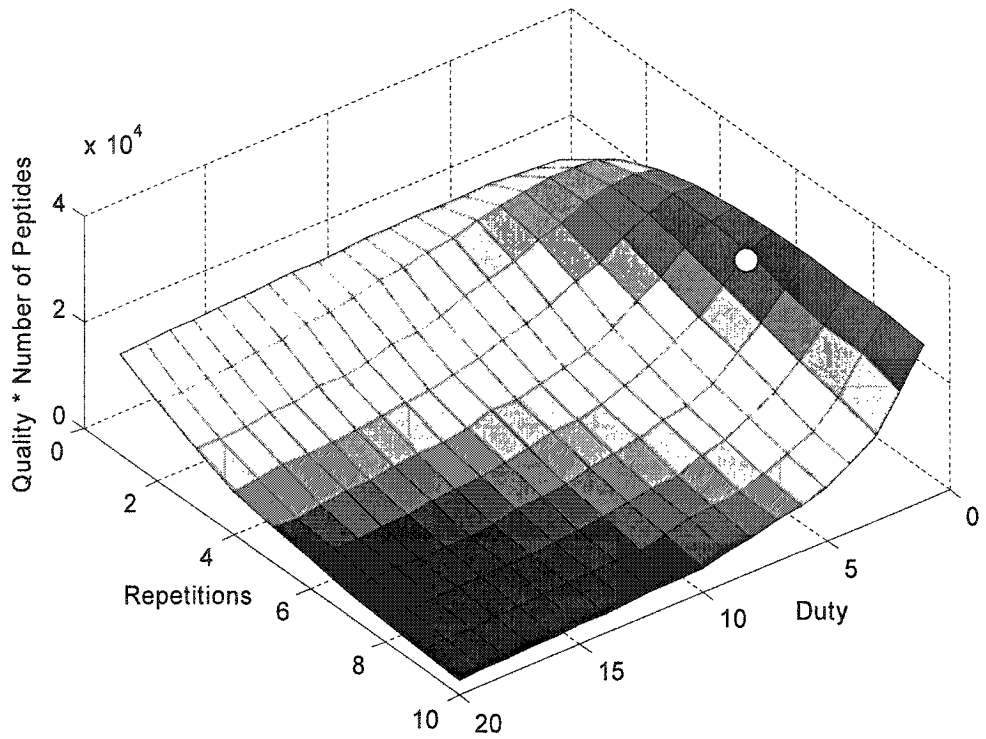

Figure 7-10: A 120 minute gradient with a circle indicating optimal operating point at 2-7.

The optimal operating point defines the best parameter set based on the quality-quantity scale. However, there are many ways to combine the two data sets; the quality-quantity scale is but one example. With this function comes an implicit assumption that one spectrum with quality N, is equal in value to N spectra of quality 1.

Therefore, it may be better to consider the optimal operating curves which define the best parameter set for a given operating point. For example, given a particular choice for average quality, the optimal operating curve will define the maximum number of peptides that can be fragmented, and give the switching parameter required to achieve it. Figures Figure 7-11 through Figure 7-14 illustrate the quality versus quantity curves for four different gradients. Each curve is associated with a single duty cycle. Points along the curve represent scores for different values of the repetition parameter. The optimal operating curve (heavy dashed line) was constructed by maximizing quality for all possible peptide counts.

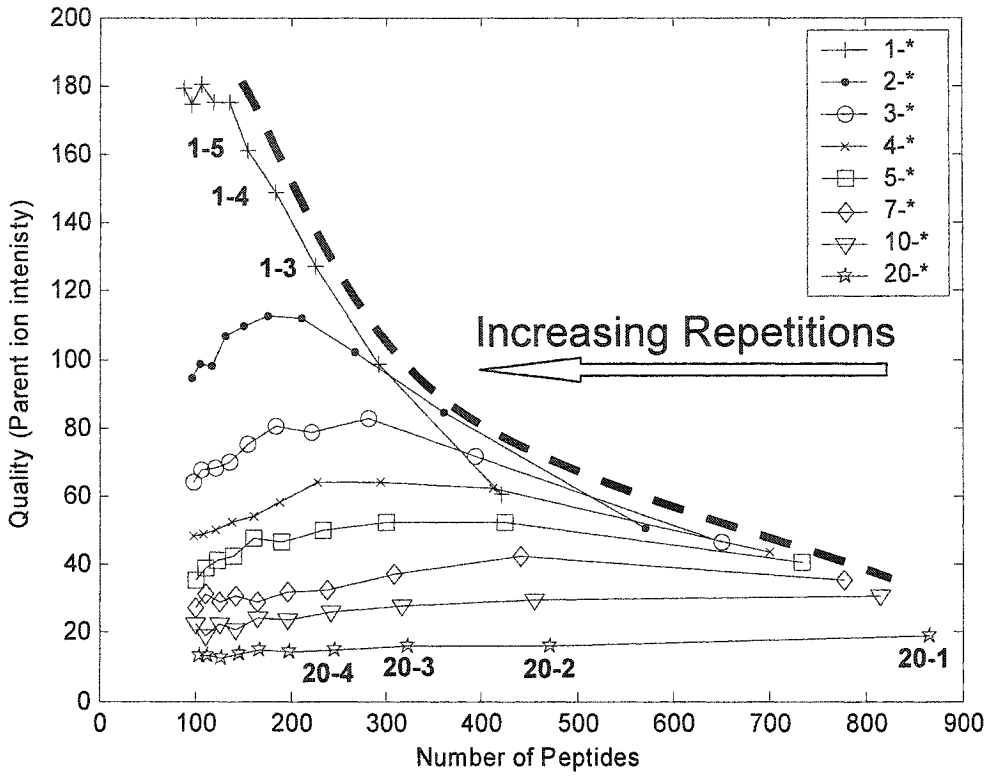

Figure 7-11: 30 minute gradient, including optimal operating curve (dashes)

The number of repetitions for each duty cycle curve increase from right to left. Several data points are labeled to show trends.

For the 30 minute gradient illustrated in Figure 7-11, the optimal operating curve is shown by the thick dashed line (shifted up slightly to avoid obscuring the data). Again, the low duty cycles dominate the optimal operating curve. The set of switching parameters with only one target selected dominates the high quality region of the curve, while the set of switching parameters with only one repetition dominates the high quantity region of the curve.

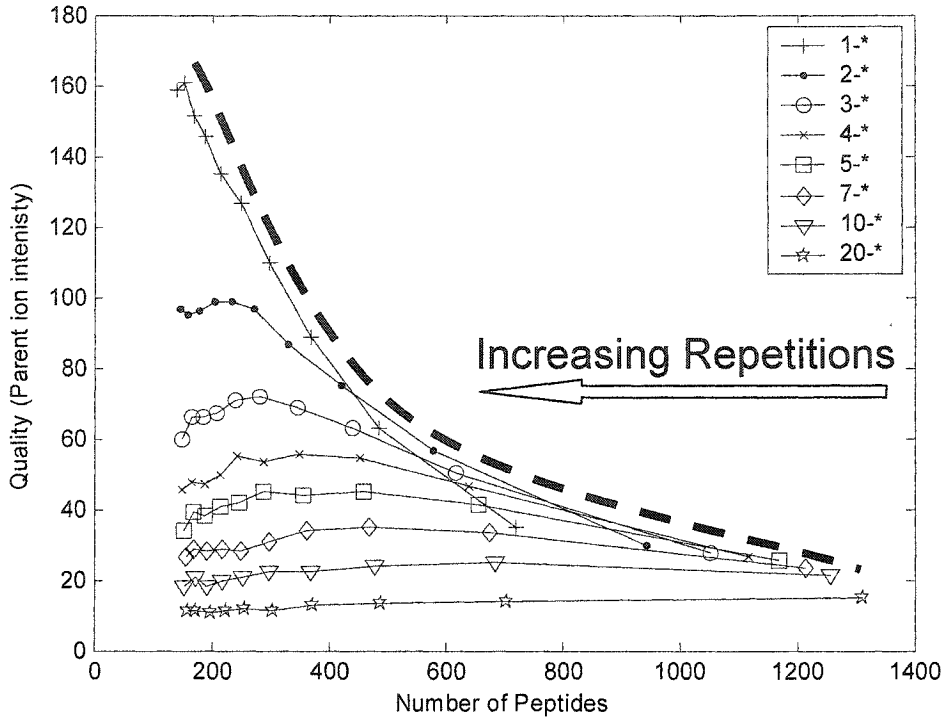

Figure 7-12: 60 minute gradient (repetitions increase from 1-10 right to left)

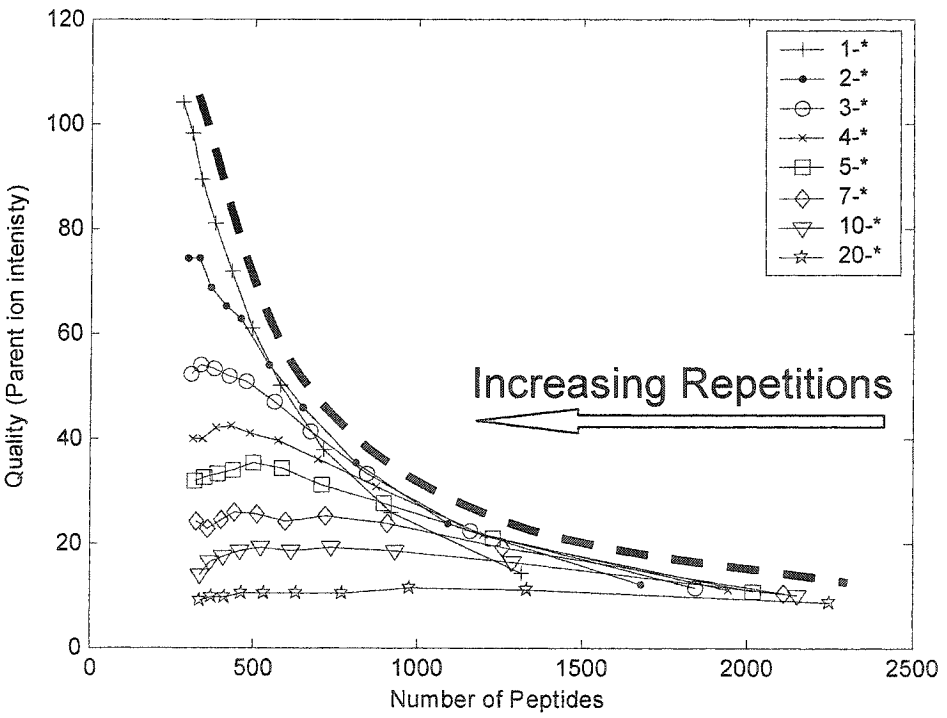

Figure 7-13: 120 minute gradient (repetitions increase from 1-10 right to left)

The 60 and 120 minute gradients of Figure 7-12 and Figure 7-13 respectively, show trends similar to that of the 30 minute gradient. There is, however, a decrease in the maximum quality and increase in the maximum quantity that accompanies the longer gradients. Also, as the gradients get longer the single target switching parameter set dominates less and less of the graph.

The 240 minute gradient of Figure 7-14 shows a considerable decrease in peptide quality accompanied by very little gain in quantity. This could indicate that the 240 minute gradient is too long, since peptides must be eluting over long periods causing a drop in intensity.

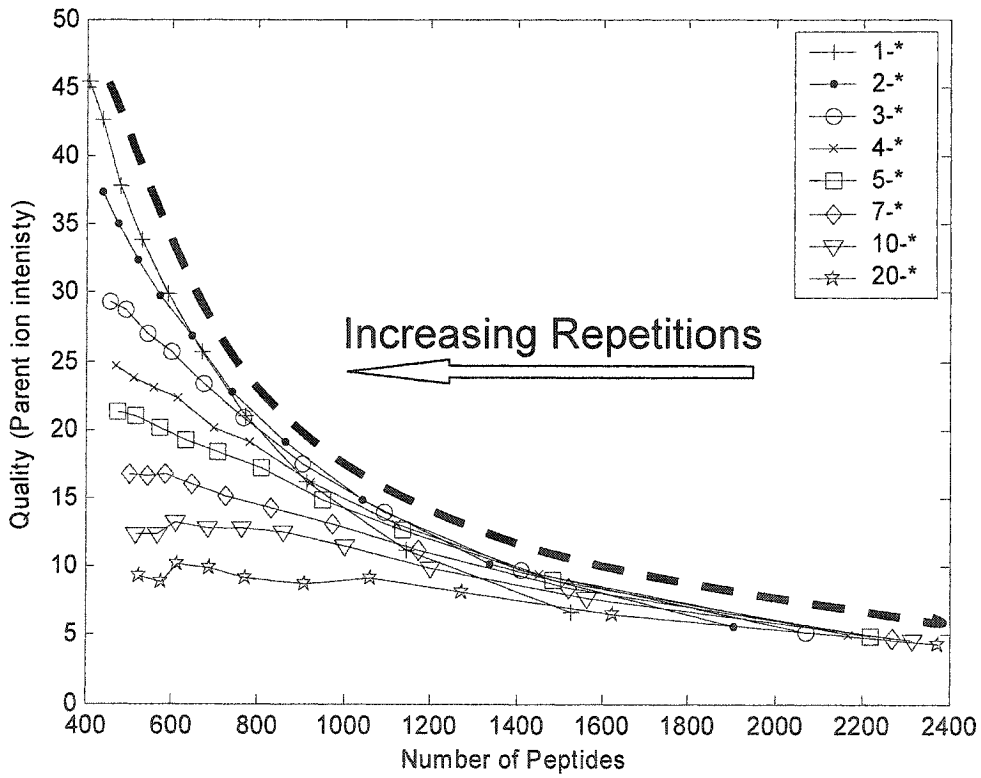

Figure 7-14: 240 minute gradient (repetitions increase from 1-10 right to left)

Figure 7-15 illustrates the optimal operating curves extracted from the four gradients superimposed on the same axes (the optimal operating curves are the heavy dashed lines in figures Figure 7-11 through Figure 7-14). These optimal operating curves overlap one

another significantly, with the different gradients (with the exception of the 240 minute) tracing various sections of a "total" optimal curve. The 240 minute gradient is sub-optimal in all cases except where a very high number of poorer quality spectra are required.

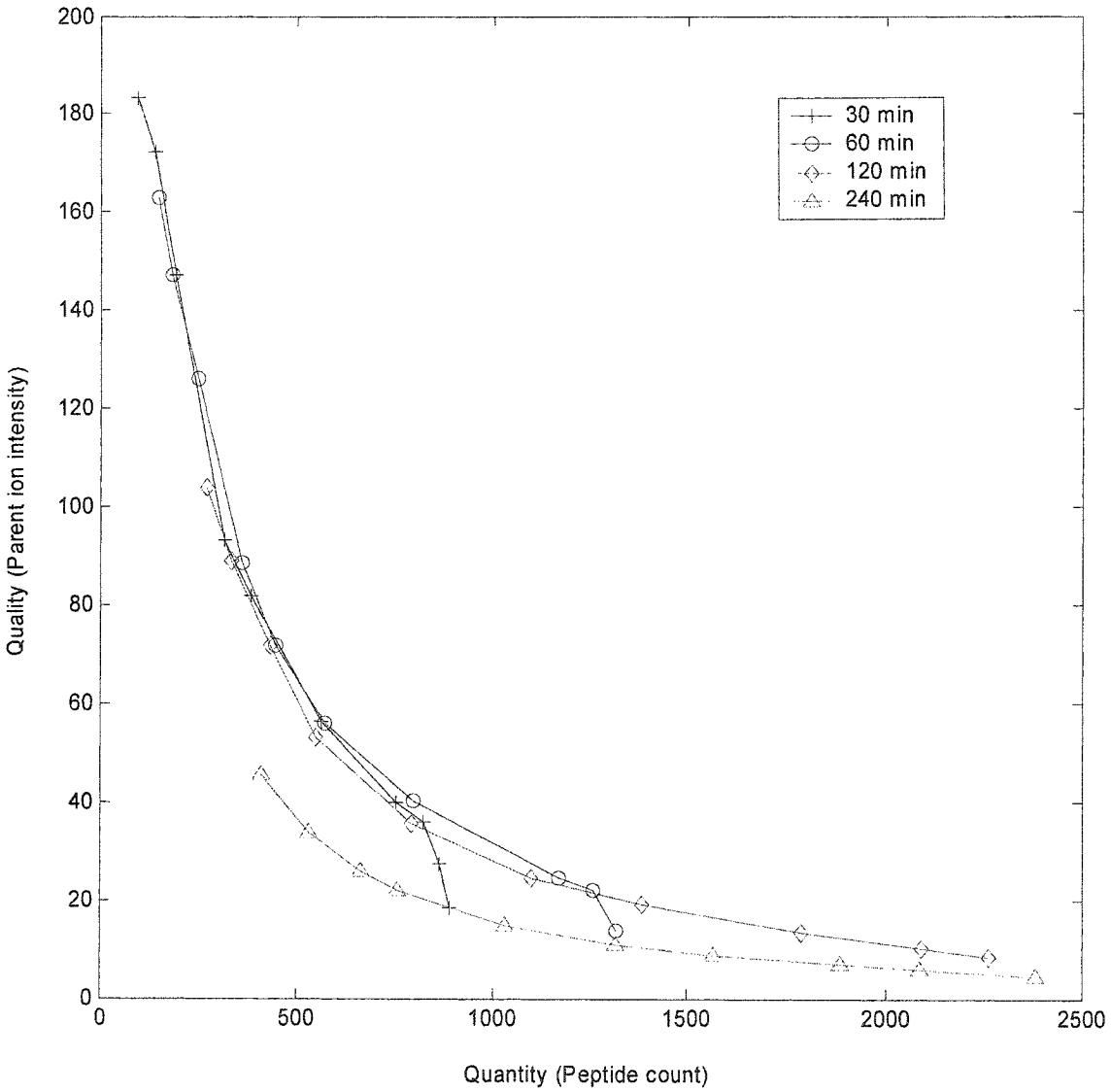

Figure 7-15: Optimal operating curves for various HPLC gradients

These combined optimal operating curves illustrate the best possible switching parameter sets for each gradient. The graph allows for the selection of the most efficient gradient for specific peptide quality-quantity operating points. The graph suggests that the longer gradients do not provide any gain in quality; the longer gradients do however provide

gains in quantity, but at the expense of quality and obviously time. This curve defines the relationship between peptide quality and quantity in tandem mass spectrometry. It can be used to direct parameter choices for experiments. For example, if the minimum acceptable peptide quality is 60, then the graph indicates that 500 peptides will be fragmented and that this can be achieved with a 30, 60, or 120 minute gradient. The operating curves for the chosen gradient will then dictate which switching parameters should be selected.

8 Summary and Future Work

This thesis has described the design, implementation and evaluation of algorithms for peak picking and deisotoping of mass spectra. These algorithms were used to mimic the operation of the mass spectrometer, by making the decision on when and where to perform tandem scans. A new type of analysis, surface intensity analysis, which uses the information contained in all the scans, is introduced. The surface intensity approach better localizes the mass of a peptide by taking advantage of its elution profile. Surface intensity analysis is used to verify that peptide locations selected for fragmentation in the simulation of the mass spectrometer are valid targets and not simply misidentified peptides.

In order to evaluate the efficiency of the operating parameters, two criteria, peptide quantity and peptide fragmentation spectra quantity, were examined. Two expected trends were verified: 1) peptide quantity increased and peptide quality decreased with increasing duty cycles and 2) peptide quality increased and peptide quantity decreased with increasing repetitions.

The simulations were compared to the 'ideal' results obtained via intensity surface analysis to determine the relative quality of the operating parameters. Unfortunately, results show that for all parameters the large majority of the peptides were missed completely. They also showed that the parameter sets with smaller duty cycles were more likely to fragment peptides closer to their maximum.

By comparing simulations it was evident that some parameters were clearly superior to others. However, some simulations were exceptional in only one of the criteria indicating that there was a negative correlation between quality and quantity. Plots of the simulations on axes of quality and quantity illustrated this correlation and defined the optimal operating curves for the mass spectrometer.

The analysis of the optimal operating curves for various length gradients led to another expected trend, a decrease in peptide intensity with increased gradient lengths; however, the longer time period allowed more peptides to be fragmented. When the optimal operating curves were superimposed on the same graph, the curves overlapped considerably. This overlapping trend suggests that there is no gain achieved by operating the mass spectrometer with a longer gradient unless large numbers of low quality peptide spectra are desired.

The most striking result of this research is that in all data sets, the majority of available peptides (~80%) were not fragmented. Thus, it is important to note that no matter which switching protocol is implemented, only a small subset of the peptides will be fragmented. This suggests that peptide mixtures of this complexity are not good candidates for this type of tandem mass spectrometry. Complex peptide mixtures should be separated even further to allow for better coverage. However, the separation should not be performed in the time domain, as longer HPLC gradients have been shown to lead to lower intensities.

8.1 Future work

Looking ahead to more advanced investigations, there are several improvements to the outlined process that may be considered to aid in both speed and accuracy.

The first and simplest improvement is to manipulate data in the time domain rather than the *m/z* domain. Doing so resolves many of the problems associated with resampling the data, and leads to fewer data points to process. Conversion back to the time domain creates regularly sampled data which allows for increased processing speed, and results in less data to manipulate. The drawback is that the time domain is less intuitive.

Surface intensity analysis can be improved by taking into consideration the peptide profile distribution rather than just the location of maximum intensity. For example, peptide spots that are not symmetric might indicate that more than one peptide is co-

eluting. Algorithms can be created with the goal of unraveling these overlapping peptides.

Before performing surface intensity analysis, the constructed 2D images are filtered. The filter, however, has fixed parameters. Lekpor et al. [17] has shown that the peak frequency content varies along *m/z*. A varying filter, such as that developed by Lekpor et al., can be expanded for use in this 2D case in order to improve noise removal.

In this analysis, all data sets were generated for identical samples. In order to better characterize the switching behaviour of the mass spectrometer, samples of different composition should be analyzed to determine if the optimal operating curves change from sample to sample. Samples of various complexities can also be analyzed to identify trends in operating curve shape and position caused by sample richness.

This analysis focused on duty cycle and repetition rate parameters. However, the peptide detection threshold parameter may play an important role in the mass spectrometer switching behaviour. The entire analyses should be re-run varying the detection threshold to discover its role in the quality-quantity curves. More important, however, is the need to determine the minimum acceptable threshold that can still provide adequate quality spectra for sequence identification.

The comparison of the simulations to surface intensity analysis showed that peptides were not always fragmented when their intensities were maximized. This is inefficient, as stronger signals will be obtained when intensity is maximized. This suggests that the current method of data-directed acquisition is far from optimal. Methods that can follow the contour of peaks and fragment them at maximum intensity would give better results.

This thesis shows that peptide peak localized is improved considerably with off line analysis. If the LC separation of peptides can be made to be reproducible, then this suggests that a map of peptide locations can be made with an initial LC-MS run. Then

additional LC-MS/MS runs on identical samples can be improved by targeting the peptides localized by surface intensity analysis.

Finally, the results of these simulations need to be tested and verified experimentally. Following the completion of the experimental verification process, the implementation of these optimizations within the proteomics pipeline can substantially increase information output.

9 Appendices

9.1 Peak picking Matlab® Code

```
function peaks = peak_picking(in, mz);
%% FUNCTION PEAK PICKING, TAKES TWO ARGUMENTS,
%% LIST OF MZ VALUES, AND THE CORRESPONDING INTENSITIES
%% PEAKS IS AN ARRAY OF MZ, VALUES AND INTENSITY OF THE PEAKS

%% BRIAN CARRILLO, MAY-6 2004
%% COPYRIGHT (C) 2004 BRIAN CARRILLO

minPeak = 2;

%% ENSURE THERE IS DATA AVA. FOR P.
if isempty(mz)
  in = [0; 0];
  mz = [0; 0];
end;

%% PAD WITH INTENSITY WITH ZEROS , PAD MZ AXIS ACCORDINGLY
in = [0; 0; in; 0; 0];
mz = [0; 0.1; mz; (mz(end)+1); (mz(end)+2)];

%% RUN RECURSIVE PEAK FINDING (BELOW)
[peak_i, peak_m] = recursive_part(in, mz, minPeak);

%% SORT PEAKS SO THEY DISPLAY CORRECTLY
[peak_m, I] = sort(peak_m);
peak_i = peak_i(I);

%% RESULTS
peaks = [peak_i, peak_m];


%% RECURSIVE PORTION OF ALGORITHM.
function [pi, pm] = recursive_part(in, mz, minPeak)
%% INSTRUCTION TO PERFORM RECURSIVE PEAK PICKING

%% CLEAR OUTPUT
pi = [];
pm = [];

%% CONTINUE ONLY IF SPECTRUM HAS MORE THAN ONE DATA POINT
if (length(in) > 2)
  %% FIND BIGGEST DATA POINT
  [big, I] = max(in);

  %% STOP WHEN BIGGEST PEAK HAS INTENSITY = 2
  if (big <= minPeak)
    return;
  end;

  idx = find(in == big);
```

```
  %% TOO MANY EQUAL POINTS, FIND CENTRAL POINT (TO SPEED UP DIVISION)
  if (length(idx) > 5)
    I = idx(round(length(idx)/2));
  end;

  %% CALCULATE PEAK WIDTH BY EMPIRICALLY DETERMINED EQUATION
  half_peak_width = 0.00017656 * mz(I) + 0.022296;

  %% CALCULATE INDEX OF REMAINING 'HALF' SPECTRA
  [temp, lo] = min(abs(mz(mz < mz(I)) - (mz(I) - half_peak_width)));
  [temp, hi] = min(abs(mz(mz > mz(I)) - (mz(I) + half_peak_width)));
  %% SHIFT NECESSARY TO CORRECT INDICES
  hi = hi + I;

  %% ALLOW FOR ENOUGH POINTS IN A PEAK, SMALL PEAKS ARE DISCARDED
  if ((hi-lo) > 3)
    %% GET APPROPRIATE INTENSITY AND MASS RANGE
    in_t = in(lo:hi);
    mz_t = mz(lo:hi);

    %% SORT THE PEAKS VIA INTENSITY
    [Y, I] = sort(in_t);
    %% TAKE THE MOST INTENSE PEAK
    I = I(end);

    %% PEAK INTENSITY IS THE MAXIMAL POINT
    pi = [in_t(I)];
    %% PEAK M/Z IS THE M/Z OF THE TOP PEAK
    pm = [mz_t(I)];
  end;

  %% CLEAR OUTPUT DATA
  ihi = [];
  ilow = [];
  mhi = [];
  mlow = [];

  %% WORK RECURSIVELY ON HALF SPECTRA IF DATA AVAILABLE
  if (lo == 0)
    [ihi,  mhi]  = recursive_part( in(hi:end), mz(hi:end), minPeak);
  elseif (hi == length(mz))
    [ilow, mlow] = recursive_part( in(1:lo),   mz(1:lo), minPeak);
  elseif (hi == lo)
    a = 1;
  else
    [ilow, mlow] = recursive_part( in(1:lo),   mz(1:lo), minPeak);
    [ihi,  mhi]  = recursive_part( in(hi:end), mz(hi:end), minPeak);
  end;

  %% RETURN CURRENT PEAKLIST, AND RECURSIVE PEAKLISTS
  pi = [ilow, pi, ihi];
  pm = [mlow, pm, mhi];
end;
```

9.2 Deisotoping Matlab® Code

```
function [deiso] = deisotope(in, mz);
%% FUNCTION DEISOTOPE TAKES TWO ARGUMENTS
%% INTENSITY LIST, AND CORRESPONDING M/Z LIST
%% DEISO IS AN ARRAY OF INTENSITY AND M/Z CELLS
%% EACH CELL REPRESENTS ONE CHARGE STATE

%% BRIAN CARRILLO, MAY-6-2004
%% COPYRIGHT (C) 2004 BRIAN CARRILLO


%% CLEAR OUTPUT
cell_out = {};

%% MZ ERROR
Err       = 0.05;
%% PERCENT INTENSITY ERROR (50%)
IntErr    = 0.75;
%% PERCENT INTENSITY ERROR (30%)
IntExtra = 1.25;
%% THRESHOLD FOR MONOISOTOPIC PEAK
thresh    = 1.0;
%% MASS OF HYDROGEN
Hydrogen = 1.007825;

%% SPACING FOR THE VARIOUS CHARGES;
sext_spc = 1/6;
quint_spc = 0.20;
quad_spc = 0.25;
trip_spc = 1/3;
doub_spc = 0.5;
sing_spc = 1.0;


%% HAVE CLEAR OUTPUT
clear mz_o in_o;
mz_o{6} = [];
in_o{6} = [];

for i = 1:length(mz)

  %% SKIP IF THE MONO-ISOTOPIC PEAK HAS A LOW AMPLITUDE
  if(in(i) < thresh)
    continue;
  else
    %% CHECK TO SEE IF THERE IS A PEAK THE APPROPRIATE SPACING AWAY.
    sext = find( abs( mz - (mz(i) + sext_spc)) < Err);
    quint= find( abs( mz - (mz(i) + quint_spc)) < Err);
    quad = find( abs( mz - (mz(i) + quad_spc)) < Err);
    trip = find( abs( mz - (mz(i) + trip_spc)) < Err);
    doub = find( abs( mz - (mz(i) + doub_spc)) < Err);
    sing = find( abs( mz - (mz(i) + sing_spc)) < Err);

    for l = 6:-1:1
      %% SET APPROPRIATE CONSTANTS FOR LOOP ITTERATION
```

```
      switch(l)
        case 1
          charge = sing;
          spc = sing_spc;
        case 2
          charge = doub;
          spc = doub_spc;
        case 3
          charge = trip;
          spc = trip_spc;
        case 4
          charge = quad;
          spc = quad_spc;
        case 5
          charge = quint;
          spc = quint_spc;
        case 6
          charge = sext;
          spc = sext_spc;
      end;


      %% ERROR CHECKING TO ENSURE ONLY ONE PEAK FOUND!
      if ( length(charge) > 1)
        [temp, idx] = min(abs(mz(charge) - (mz(i) + spc)));
        charge = charge(idx);
      end;


      %% CHECK FOR CHARGED IONS
      if ((isempty(charge) == 0) & in(i) > 0)
        %% COMPUTE ISOTOPIC PEAK HEIGHTS FOR THE NEXT FEW
        %% ISOTOPES (UNTIL PEAK IS TOO SMALL < 0.5)
        dist = iso_dist2(l * mz(i),in(i),0.5);

        %% BREAK IF EXPECTED PEAK TOO SMALL...
        if (length(dist) < 2)
          continue;
        end;

        %% CHECK IF 2ND ISOTOPIC PEAK IS PRESENT @ APPROPRIATE
INTENSITY
        if (in(charge) >= dist(2) * IntErr)

          %% ADD PEAK TO PEAKLIST, SUM PEAK INTENSITIES
          mz_o{l} = [mz_o{l}; mz(i)];
          in_o{l} = [in_o{l}; in(i) + dist(2)];

          %% SUBTRACT IDEAL PEAK (PLUS 20%) FROM THE REAL PEAK
          in(charge) = in(charge) - dist(2) * IntExtra;

          %% IF THE 2ND. PEAK WAS SMALLER THAN EXPECTED,
          %% REMOVE IT (NO NEGATIVE PEAKS)
          if (in(charge) < 0)
            in(charge) = 0;
          end;
```

```
            %% CHECK FOR FURTHER ISOTOPES
            j = 2;
            while j <= (length(dist) - 1)
              %% FIND NTH ISOTOPE
              charge = find( abs( mz - (mz(i) + spc * j)) < Err);

              %% BREAK IF ISOTOPE NOT FOUND
              if (isempty(charge))
                break;
              end;

              %% CHECK IF NTH ISOTOPE IS PRESENT @ RIGHT INTENSITY
              if (in(charge) >= dist(j+1) * IntErr)

                %% SUBTRACT IDEAL PEAK (PLUS 20%) FROM THE REAL PEAK
                in(charge) = in(charge) - dist(j+1) * IntExtra;

                %% IF THE NTH PEAK WAS SMALLER THAN EXPECTED,
                %% REMOVE IT (NO NEGATIVE PEAKS)
                if (in(charge) < 0)
                  in(charge) = 0;
                end;

                %% THESE EXTRA PEAKS SUPPORT MONO-ISOTOPIC PEAK,
                %% SO ADD INTENSITY TO MONO-ISOTOPIC PEAK
                in_o{l}(end) = in_o{l}(end) + dist(j+1);
              end;

              j = j+1;
            end;
            % WHILE J <= LENGTH(DIST)

            %% DELETE THE CURRENT PEAK
            in(i) = 0;
            continue;
          end;
          %(IN(SYNC) >= IN(I) * DIST(1) * (1 - INTERR))
        end
        % (ISEMPTY(SYNC) ~= 0)
      end
      % FOR L = 4: 1:
    end
    % IF(IN( ) < THRESH)
end
% WHILE I = 1:LENGTH(MZ)


%% SORT ALL MASS
for i = 1:6
  [mz_o{i}, I] = sort(mz_o{i});
  in_o{i} = in_o{i}(I);
end;
deiso = [in_o, mz_o];
```

10 References

[1] . About the Human Genome Project, 29 10 2003, http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

[2] Karp, Gerald. (1999) Cell and Molecular Biology. Concepts and Experiments 2nd Edition. New York: John Wiley & Sons.

[3] Marte, Barbara. (2003) Proteomics, Nature Insight 13: 6928.

[4] Cahill, D. J., et al. (2001) Bridging Genomics and Proteomics. In: Pennington SR, Dunn MJ, (eds.) Proteomics: from protein sequence to function, 1-22 Oxford: BIOS.

[5] Wilkins, M.R., (2001) The Automation of Proteomics: Technical and Informatic Solutions for High-Throughput Protein Analysis. In: Pennington SR, Dunn MJ, (eds.) Proteomics: from protein sequence to function, 171-192 Oxford: BIOS.

[6] Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. (1999) De novo peptide sequencing via tandem mass spectrometry, J Comput Biol. 6(3-4):327-42.

[7] Aebersold R, Mann M. (2003) Mass spectrometry-based proteomics, Nature 422(6928):198-207.

[8] Dunn MJ, Gorg A. (2001) Two-dimensional polyacrylamide gel electrophoresis for proteome analysis, In: Pennington SR, Dunn MJ, (eds.) Proteomics: From Protein Sequence to Function. 43–64 Oxford: BIOS.

[9] Rybicki Ed, and Purves M, SDS POLYACRYLAMIDE GEL ELECTROPHORESIS (SDS-PAGE), 12 02 2004, http://www.mcb.uct.ac.za/sdspage.html

[10] SWISS-2DPAGE Map Selection: NUCLEOLI_HELA_1D_HUMAN (small, no highlight), 12 02 2004, http://ca.expasy.org/cgi-bin/map2/noid?NUCLEOLI_HELA_1D_HUMAN

[11] Swiss-Prot Release 42.9 statistics, 02 02 2004, http://ca.expasy.org/sprot/relnotes/relstat.html

[12] Harris, Daniel C., (1999) Quantitative chemical Analysis, 5th Edition. New York: W.H. Freeman and Company.

[13] De Hoffman Edmond, Stroobant V, (2001) Mass Spectrometry, Principles and Applications (2nd ed.). Chichester: Wiley.

[14] Baker Peter, Clauser Karl, MS-Isotope 11 12 2002, http://prospector.ucsf.edu/ucsfhtml4.0/msiso.htm

[15] Breen EJ, Hopwood FG, Williams KL, Wilkins MR. (2000) Automatic poisson peak harvesting for high throughput protein identification, Electrophoresis 21(11): 2243-51.

[16]Gay S, Binz PA, Hochstrasser DF, Appel RD. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides, Electrophoresis. 20(18): 3527-34.

[17] Lekpor et al. (2003) Development of a robust software-based method for noise filtering in time-of-flight mass spectra, Mol Cell Proteomics 2: 606-641.

[18] Curtis A. Hastings, Scott M. Norton, Sushmita Roy (2002) New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data, Rapid Commun. in Mass Spectrom. 16(5): 462-467

[19] Zhang, Z, and Marshall, A.G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra, J. Am. Soc. Mass Spectrom. 9: 225–233.

[20] A. Almudaris, D.S. Ashton, C.R. Beddell, D.J. Cooper, S.J.Craig and R.W.A. Oliver (1996) The assignment of charge states in complex electrospray mass spectra, Eur. J. Mass Spectrom. 2: 57-67.

[21] C. Lavoie, et. al. (1999) Roles for α2p24 and COPI in Endoplasmic Reticulum Cargo Exit Site Formation, J. Cell Biol. 146: 285-300.

[22] Sylwia Wasiak, et. al. (2002) Enthoprotin: a novel clathrin-associated protein identified through subcellular proteomics, J. Cell Biol. 158: 855-862.

[23] M. Wehofsky, et. al. (2001) Isotopic Deconvolution of MALDI Mass Spectra for Substance-class Specific Analysis of Complex Samples, Eur. Mass Spectrometry, 7: 39-46.

[24] D. N. Perkins, et. al. (1999) Probability-based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data, Electrophoresis, 20: 3551-3567.