12-2013

DEMONSTRATION OF A TARGETED PROTEOME CHARACTERIZATION APPROACH FOR EXAMINING SPECIFIC METABOLIC PATHWAYS IN COMPLEX BACTERIAL SYSTEMS

Adam Justin Martin
University of Tennessee - Knoxville, amarti31@utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Adam Justin Martin entitled "DEMONSTRATION OF A TARGETED PROTEOME CHARACTERIZATION APPROACH FOR EXAMINING SPECIFIC METABOLIC PATHWAYS IN COMPLEX BACTERIAL SYSTEMS." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Chemistry.

Robert N. Compton, Major Professor

We have read this thesis and recommend its acceptance:

Robert L. Hettich, Shawn R. Campagna, Michael J. Sepaniak

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

DEMONSTRATION OF A TARGETED PROTEOME CHARACTERIZATION APPROACH FOR EXAMINING SPECIFIC METABOLIC PATHWAYS IN COMPLEX BACTERIAL SYSTEMS

A Thesis Presented for the
Master of Science Degree
The University of Tennessee, Knoxville

Adam Justin Martin

December 2013

DEDICATION

I owe the completion of this thesis to no one more than I owe it to my extraordinary wife, Keri. Every semester of my graduate experience brought new and seemingly insurmountable challenges, yet her steadfast faith in my success would always fend off my surrender. And when shadows were cast over our lives, she would not let them prevent me from seeing my research through to its end. Her patience for both my daily work and my career work was unbelievable, as was her loving sacrifice of sleep to meet me at the door with our beautiful daughter when I came home late, night after night. For all of her dedication to believing, supporting, and loving me, I dedicate this work to her.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to my co-advisors Dr. Robert Compton and Dr. Robert Hettich for their support and mentoring throughout my graduate career. I am thankful to both of them for their wisdom that fueled my growth as a chemist, and their patience throughout my scientific development. I would also like to thank Dr. Richard Giannone, Weili Xiong, Ritin Sharma, and Zhou Li in the BioEnergy Science Center's proteomics group at Oak Ridge National Laboratory for their expert advice and discussions on sample loading and separations. Specific appreciation goes to Dr. Paul Abraham and Dr. Rachel Adams in the BESC proteomics group at ORNL. Thanks to Paul for his tutelage on instruments and methods as well as his assistance with defining peptide abundance in Chapters 3 and 4; and thanks to Rachel for her uncannily patient guidance on operating bioinformatic software, and for her wealth of custom data scripts to bridge the important informational gaps. I also want to thank my committee members, Dr. Shawn Campagna and Dr. Michael Sepaniak, for their time reading my thesis and participating in its defense, and additional gratitude is due both of them for enriching my graduate career with collaborative projects not contained in this thesis.

Thanks to the U.S Department of Energy, the Bioenergy Research Program, and UT-Battelle for the funding, facilities, and additional resources that made my work possible.

I would also like to extend a special appreciation to Doug Hagemann for his exceptional maintenance of my scooter to provide safe and reliable transport to and from work every day.

ABSTRACT

Multiple Reaction Monitoring (MRM) is a powerful tandem mass spectrometry (MS/MS) tool frequently implemented in proteomic studies to provide targeted analysis of proteins and peptides. The selectivity that MRM delivers is so strong that it provides the quadrupole mass spectrometers (QQQ), on which it is commonly employed, with pertinence to proteomic studies that they would otherwise lack for their relatively low resolution. Additionally, this increased level of selectivity is sufficient to supplant complicated fractionation techniques, additional dimensions of chromatography, and 24 hour long MS/MS experiments in simplistic biological samples. But there is a deficiency of evidence to determine the applicability of MRM to complex samples such as those containing the entire proteome of single cellular organisms. These samples are often employed to profile entire metabolic pathways at a cellular level using the complete set of proteins involved in the pathway's characteristic enzyme driven reactions. This sweeping view of gene expression is vital to understand cellular response, and profiling these expressions would benefit greatly from the introduction of MRM as a viable approach for characterizing metabolic networks. This thesis takes two significant steps towards this viability by first demonstrating MRM reproducibility in complex samples, and characterizing degrees to which certain design related factors influence the quality of these MRM. The next step applies knowledge gained by the first to exhibit the MRM profiling of a vital metabolic pathway from a complex sample. This step also demonstrates the self-sufficient utility an *ab initio* method, based on proteins and peptides predicted from the genome sequence, for designing MRM. Combining the *ab initio* design approach with the MRM of complex samples represents substantially shorter experimental preparations for profiling metabolic networks, and renders the characterization of gene expression on a cellular level as a more widely accessible study within proteomics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction to Targeted Proteome Characterizations

1.1 Advent of Systems Biology

As a relatively new and emerging paradigm in bioscience, Systems Biology is a remarkably large and interdisciplinary field that combines computational modeling, high-throughput analytical methods, and state of the art data analysis in the holistic study of cellular metabolic pathways [1-4]. Rather than concentrating on isolated components, this integrated approach studies concerted biological activities as a unified network, which provides context for their interwoven causal nexus, and characterization of their mechanisms and purposes at the cellular level [5]. This new characterization insight provides a better understanding of cellular responses to environment, disease, and intercellular communications by offering a more direct connection between gene expressions within a cell and the stimuli that provoked them [3].

Molecular biology defines these gene expressions by the molecules and reactions they comprise. A cell's genes are contained in strands of deoxyribonucleic acid (DNA), and a gene expression begins when a cell transcribes a portion of its DNA to ribonucleic acid (RNA). Then genetic information is translated from RNA into functional macromolecules known as proteins. Some of these proteins are enzymatic which means that they catalyze reactions in life-sustaining metabolic pathways, such as glucose digestion and DNA synthesis[5, 6]; the latter example typifies the reticulate causal nexus mentioned earlier. Each protein catalyzes a specific reaction, each viable combination of reactions maps out a specific pathway, and each pathway is activated by expressing the genes for the proteins that drive its reactions. Due to their central role in this process, proteins provide a link between pathways and genes, and consequently they are a medium for monitoring gene expression[5]. This means that proteins can reveal detailed information about a cell's structure and metabolic activity; which for bacteria, a single-cell life form, this means they also provide a comprehensive molecular profile of the entire organism.

1.2 Proteomics

One component of the systems biology approach to molecular biology is proteomics, which is defined as profiling and characterizing entire metabolic pathways by identifying complete sets of proteins. Such large scale profiling requires using a protein's intrinsic properties to not only uniquely identify and accurately quantify it, but also to separate it from other proteins to assist these measurements. Applying this practice to multiple proteins with multiple properties, while providing meaningful qualitative and quantitative data in one experiment, requires high throughput analytical techniques and instruments possessing both a wide dynamic range and an elevated level of identification.

A staple analytical technique in proteomics is mass spectrometry (MS), which encompasses a multitude of strategies applied across a variety of instrumental configurations. Like high performance liquid chromatography (HPLC) and two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), MS is a high throughput technique that can analyze thousands of proteins in one experiment; however, it provides significant increase in identity specificity over HPLC and 2D-PAGE. Rather than using a protein's affinity to stationary phase (an external interaction), MS identifies proteins by their mass to charge ratio (*m/z*, an intrinsic chemical property). As an additional performance capability, HPLC can be coupled indirectly or directly to MS (offline or online LC-MS) to enhance both separation and subsequent MS identification, and can be augmented by adding a second chromatographic dimension before MS that operates on a different intrinsic property (2D-LC).

Further specificity in LC-MS identity is achieved through a very common strategy known as bottom-up proteomics. This strategy begins by digesting the proteins into peptides for MS analysis, then adding a second MS step that fragments (MS/MS). The fragments identify the peptides and the peptides identify the protein, which minimizes ambiguity at the protein and peptide level. Fragment based identifications are facilitated by a systematic nomenclature that identifies their portion of the peptide sequence. Each peptide has an amino-terminus and a carboxyl-terminus (N- and C- terminus respectively), and when a peptide ion breaks, the charge remains on either the N- or C-terminus leading to the detection of attached fragment. N-terminus ions are labeled a-,

b-, or c- type ions, while C-terminus are labeled x-, y-, and z- type ions, as seen in Figure 1.1. The three possible labels arise from three possible peptide bonds that could break to form a fragment, and ions of one terminus are complementary to ions of the other. Specifically, a- and x- type ions are complementary and originate from breaking the alpha carbon to carbonyl carbon bond, b- and y- ions are complementary and originate from breaking the carbonyl carbon to amide nitrogen bond, and c- and z- type ions are complementary and originate from breaking the alpha carbon to amide nitrogen. Since these bonds can occur multiple times in a peptide, with multiple possibilities of each ion type, the bonds are numbered in sequence from both termini, with each ion type numbered in turn. For example a peptide with three carbonyl-amide bonds, such as the one in Figure 1.1, will have three b-type and three y-type ions as seen in Figure 1.2, with the complementary ions being b-1 and y-3, b-2 and y-2, and b-3 and y-1.

As one of the preferred analytical methods in proteomics, bottom-up 2D-LC-MS/MS offers a very high level of specificity, yet such experiments can still run as long as 24 hours to accommodate the complexity of proteomic samples. Although individual run times can be shortened by performing the first LC separation offline (fractionation), this strategy requires multiple runs and time consuming sample preparation.

The various methods of proteomic MS can be divided into two categories: global and targeted, and each one presents advantages and challenges. The global approach is and enables the single experiment profiling of multiple metabolic pathways under that sample's conditions; this can also retroactively apply to proteins and pathways not yet linked. However, with so many peptides and fragments to separate and scan from a continuous flow, some less abundant proteins will not be identified, which in turn means that some metabolic pathways may be expressed but lack enough data for a profile.

The targeted approach looks for specific proteins in order to profile select metabolic pathways. Targeted experiments require prior knowledge of protein-pathway links, and they ignore a large majority of proteins, hence the data from these experiments will not be applicable outside of the chosen pathways. However, this focused scanning allows targeted experiments to measure proteins that would otherwise go undetected, and to

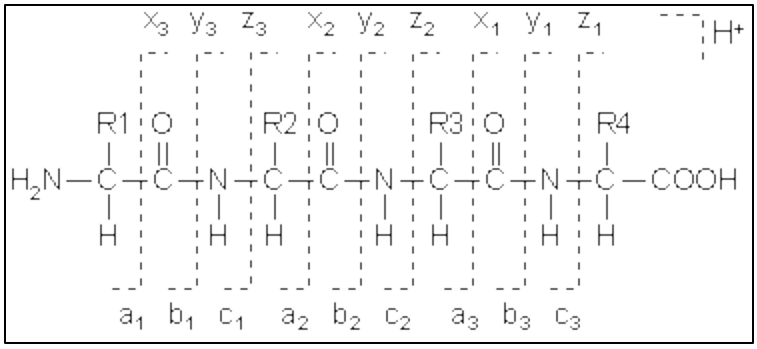


Figure 1.1 Diagram of peptide fragment labeling, courtesy of Matrix Science [7]

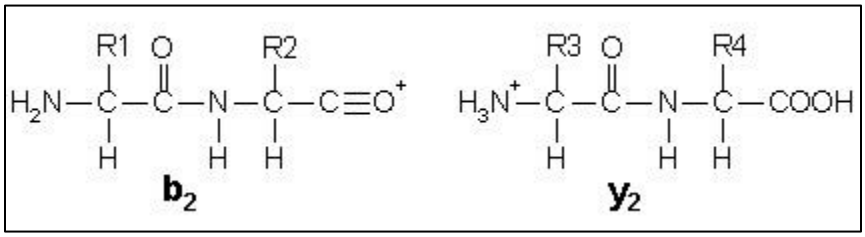


Figure 1.2 B- and y-type ions to show charge placement, courtesy of Matrix Science [7]

profile pathways of interest that would otherwise lack sufficient data. Additionally, this narrow view can facilitate the measurement of absolute protein concentrations, and thus better measure the scope of a gene's expression.

1.3 Targeted Proteomics

Although targeted detection of proteins and targeted MS have both existed for several decades, targeted MS is a relatively new addition in proteomics. Older targeted techniques, such as affinity chromatography and western blotting, exploit the selective binding between enzymes and substrates, or antigens and antibodies to achieve highly specific protein isolation and detection[8]. However, these techniques rely on the careful execution of complex and time consuming sample preparations and processes.

Targeted MS analysis is based on an MS/MS technique known as selected reaction monitoring (SRM), and similar to a previously discussed concept, it utilizes mass instead of interaction to achieve specific isolation and detection. Within MS/MS the select profiling of SRM differentiates it from global MS/MS which is designed to identify thousands of proteins and peptides in one experiment. Global analysis scans all peptide charge-state (precursor) *m/z*s and all fragment *m/z*s, while SRM scans for one specific precursor *m/z*, with one specific fragment *m/z*. Each pairing of specific *m/z*s, one precursor with one fragment, is called a *transition*, and when multiple transitions are targeted, the experiment is called a multiple reaction monitoring (MRM). Each protein contains a number of unique amino acid sequences found among some of the peptides it comprises, and MRM can identify such a peptide by detecting its unique transitions. These unique peptide identities provide protein identities as specific as those obtained by traditional targeted techniques.

Being both a targeted approach and an MS method, MRM offers a few unique advantages. First, compared to other MS approaches, it requires less scans per experiment, which affords it more time per scan while reducing overall time for the experiment. Next, MRM's selective detection allows it to identify low abundance proteins with less fractionation than global MS, and MRM can even be performed with only one dimension of chromatography (1D-LC). Also, because isolation and detection take

place inside the instrument, MRM requires far less sample preparation than other non-MS targeted approaches. Finally, MRM can be performed on relatively low cost, and readily available, instruments without requiring expensive antigen tags, making it one of the least expensive options of both targeted and MS methods.

Despite its youth, MRM is a widely accepted method, and as it is backed by substantial precedent, MRM related research continues to grow for both proteomics in general, as well as for its own development [9]. The instrumental backbone for MRM research and application has historically been the triple quadrupole mass analyzer (QQQ), but utilization of the recently developed quadrupole-Orbitrap (Q-Exactive) is steadily increasing as this instrument offers advanced resolution [10, 11]. Other recently proposed advancements include methods for single-transition-based MRM identities and standard protocols to evaluate MRM system performance [11, 12]. Currently, employment of MRM includes profiling protein networks, quantifying post-translational modifications, and disease biomarker verification[9]. Recent specific examples of MRM applications include research by Gall et al[6] in developing a sub-5-minute method for measuring blood plasma concentrations of rufinamide in low volume samples, and another study by Li et al [13] developed a method for quantitative measurement of arenobufagin in rat plasma, which was also a 5 min method. Representing samples outside of plasma, Vierikova et al [14] developed an MRM method for determining the natamycin content of cheese in a 14 min measurement.

The success of proteomic MRM is due in large part to advanced bioinformatics tools, such as the popular software application *Skyline* [15]. The applicability of this software is evident throughout every step of MRM development; specifically, it can be used to design, evaluate, and modify a method, as well analyze final MRM data. Skyline's versatility will be discussed further in chapter two.

1.4 Thesis Objectives

Thus far, MRM has been applied to relatively simplistic sample sets, such as blood, serum, urine, or even cheese. Yet it poses the potential to interrogate large unfractionated samples containing whole proteomes of multiple organisms. Placing

greater emphasis on tuning the MRM methods and instruments allows the sample to remain complex; this reduces the time, money, and mistakes risked in both planning and executing individual sample preparations for each study.

However, designing and tuning a specialized MRM for each study is a daunting task that presents its own potential errors and fruitless endeavors. Meaningful experimental design requires guidelines that point out subtle yet fatal flaws, and help maintain a focus on pertinent issues. The work presented in this thesis was devised to offer such guidance by first identifying factors that have a potentially significant impact on MRM design, then conducting comparative MRM's that focus on those factors, and finally evaluating both the quality of the resulting MRM data and the effects of each factor on that quality.

The research conducted in pursuit of this goal was split into two projects, with two of the following chapters of this thesis dedicated to discussing how each was designed, executed, and evaluated. The first project, as presented in chapter three, fields the fundamental question: "Can MRM of peptides in an unfractionated sample, containing whole proteomes of multiple organisms, be reproducibly accomplished?" Additionally, this project identifies and investigates chromatographic congestion and peptide abundance for their potential influence on MRM quality. Chapter four encompasses the expansion project, which responds to the challenge: "Can such an ambitious MRM be employed to profile an entire metabolic pathway?" In the course of facing this gauntlet, this project also provides a comparison of two competing methods for MRM design, empirical and *ab initio*. Each of these projects are complimentary to the other; the first is a proof of concept study that provides the basis for the second, which itself is an applicable demonstration that brings relevance to the first. Together, these symbiotic projects present a defendable method for designing an MRM to profile metabolic pathways in a complex sample of complete proteomes.

With the success of these projects in meeting both their individual goals and the overarching objective, this thesis will deliver three contributions to the advancement of proteomic MRM within systems biology. First, it will demonstrate that MRM's strength in eliminating a significant amount of sample preparation can be further applied to samples

of complete proteomes. Second, this thesis will explore the limits of how complex an MRM sample can be while still returning meaningful data. Finally, the work presented in this study will demonstrate the ability of MRM to characterize a specific metabolic pathway in a complex sample, containing the complete proteomes of several organisms, without fractionation, in 60 min.

Chapter 2: Experimental Approach for Targeted Proteomics

2.1 Overall Experimental Design

Fulfilling the goals of this thesis relied on carefully designing the experiments for each project, based on a significant understanding of biological samples, established methods, and instrumental mechanics. This chapter will outline and define these concepts through a detailed discussion of the experimental designs. First, the contents of the selected sample will be described to define its merit in testing MRM designs. This will be followed by a discussion on relevant sample preparation, sample loading, and LC methods. Next, a brief explanation of the ion source will connect LC to MS/MS, and explain the origin of multiple peptide precursors. The two sections that follow will cover the two types of mass analyzers employed by this thesis, and will describe how both instruments play a role in each project. These sections will also serve as a comparison for how each mass analyzer achieves MS/MS. Finally, this chapter will focus on the software employed in collecting, calculating, and evaluating data. This portion will also look closer at the specific contribution provided by Skyline.

2.2 Microbial Sample Selection

A model synthetic microbial consortium consisting of the microbes *Escherichia coli*, *Rhodopseudomonas palustris*, *Ignicoccus hospitalis*, and *Nanoarchaeum equitans* (*E. coli, R. pal, I. hos, N. equi*) was prepared to evaluate the design and execution of each MRM in a moderately complex but highly controlled biological system. For this four isolate system (4-iso) there are approximately 11,000 possible proteins, based on genome evaluations. Assuming that each microbe could express about one-half of its genome products under one growth condition would suggest a possible pool of about 5,500 protein products. This would yield over 80,000 possible tryptic peptides in the sample.

2.3 Sample Preparation and Separations

The samples were prepared and loaded according to in-house techniques and then separated by an in-house LC method that was modified for a 60 min 1D-LC run; the polar and nonpolar solvents used therein were also prepared in house[16]. The polar solvent (solvent A) is 5% acetonitrile (ACN), 95% HPLC-grade water, and 0.1% formic acid; the nonpolar solvent (solvent B) is 70% ACN, 30% HPLC-grade water, and 0.1% formic acid.

The microbial samples were pulse-sonicated for 2 min and boiled for 5 min in a sodium dodecyl sulfate (SDS) solution, resulting in cell lysis, and yielding proteins which were immediately precipitated with trichloroacetic acid (TCA). After being washed and re-solubilized, the proteins were then digested with sequencing-grade trypsin for approximately 12 hours, which cleaved the proteins into their constituent peptides at lysine and arginine residues. For each MS/MS run, 5 to 10 µg of peptides were bomb-loaded onto a fused silica back column packed to approximately 5 cm with Kinetix C18 reverse-phase (RP) resin, and washed offline with solvent A. Next, the back column was placed in line behind a nanospray emitter that was approximately 10 cm long and packed full length with C18 RP resin. The peptides were then separated by 1D-LC with a 60 min gradient from 2% solvent B to 90% solvent B, at a flow rate of 300 nl per minute.

2.4 Electrospray Ionization/Nano-Electrospray

Nano-flow electrospray ionization (Nanospray) was employed for interfacing the LC to the MS/MS. Figure 2.1 shows this atmospheric pressure interface, and Figure 2.2 illustrates the basic operation of nanospray in which the peptide laden solvent is passed through a charged needle to form charged droplets. These charged droplets accelerate through a potential gradient to the inlet port of the spectrometer; meanwhile the solvent evaporates until the droplets break into charged peptide ions. Solvents A and B both contain ACN, which is a volatile organic that aids in solvent evaporation; they also both contain formic acid to increase droplet conductivity. By utilizing nanoliter flow rates, nanospray decreases solvent volume and increases ion formation. For both projects,

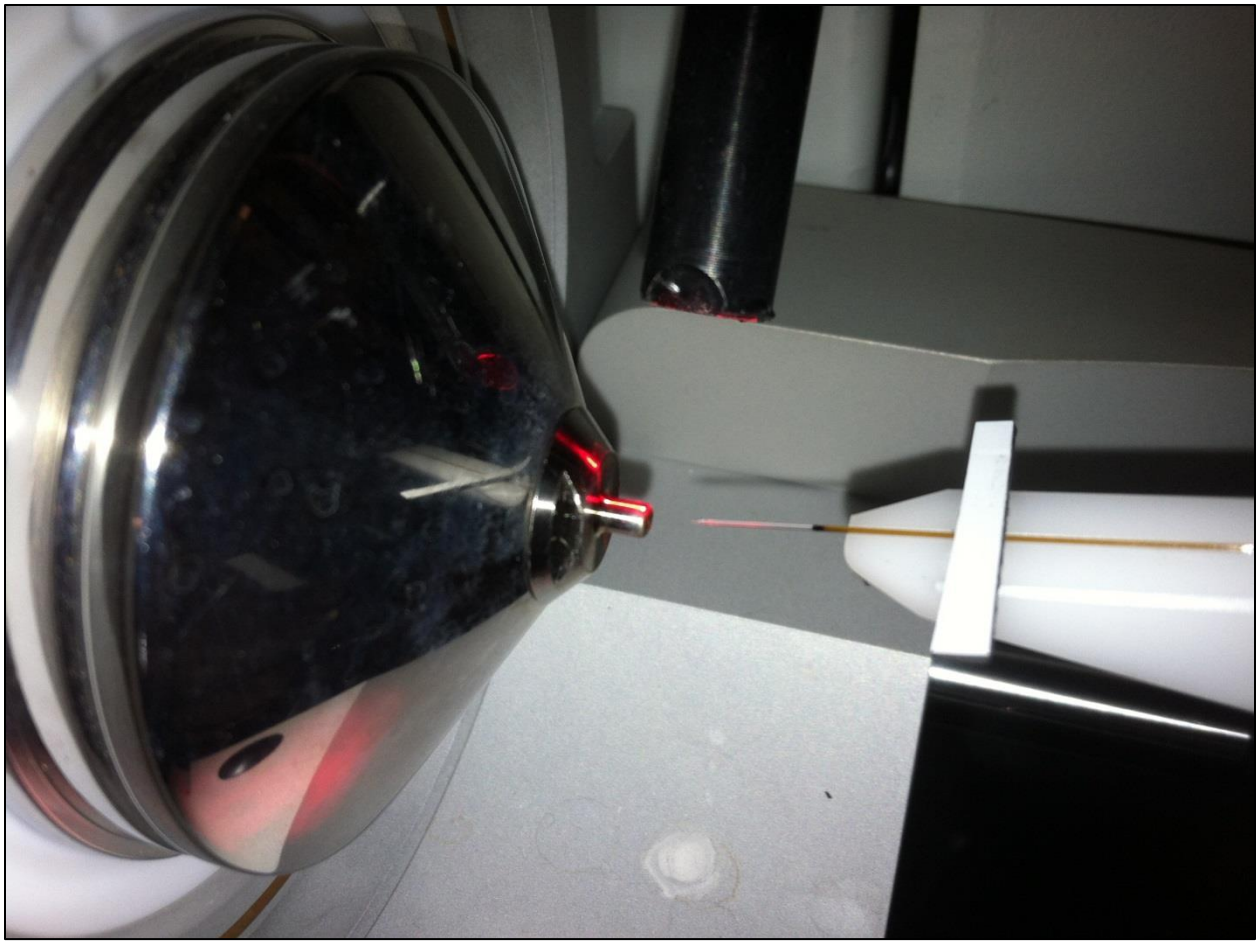


Figure 2.1 Nanospray Interface for LC-MS/MS. The pulled nanospray tip of the fused silica tubing is shown on the right; this connects the flow from the LC column to the ESI source of the MS and is open to atmosphere. This nanospray tip is localized a few mm. away from the heated metal capillary (shown on the left) of the ESI source of the MS.

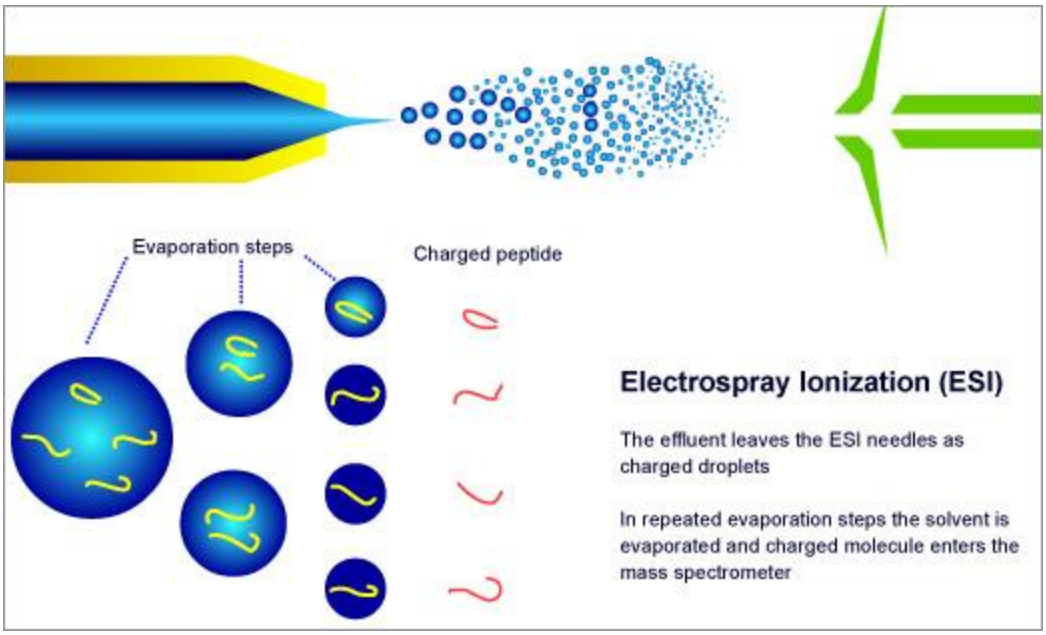


Figure 2.2 Basic illustration of nano-flow electrospray ionization, courtesy of Dionex[17]

nanospray flowed at 300 nL per minute, as set by the LC pump. A voltage between 2300 and 2700 volts was used to establish the electrospray ionization conditions.

The amount of charge a peptide ion receives can vary, which gives rise to multiple precursors ions for one peptide. Although they have a minimal difference in mass, these precursors will differ greatly in *m/z* according to their charge. Typically, nanospray ions range in charge from +1 to +4, but +2 ions are the most commonly detected by MS/MS. This is because +1 ions often lack enough charge repulsion to fragment, while the +3 ions and up have too much charge repulsion to be easily formed[18].

2.5 Global Proteome Characterization using 1D-LC-MS/MS

After peptide elution from 1D-LC and ionization by nanospray, the ions enter the MS and are targeted for tandem mass spectrometry (MS/MS). Within the confines of this study, all MS/MS analysis, both global and targeted, was solely performed in the cells of quadrupole mass analyzers. A quadrupole cell is a square array of four hyperbolic metal rods to which voltages are applied for the manipulation of ion trajectories. Any two rods that are opposite each other in the array are connected electrically, and as this occurs twice in a quadrupole, it is better to view the array as being two pairs of rods rather than four individual rods.

All global 1D-LC-MS/MS measurements were made using a Thermo-Fisher LTQ-Velos Pro linear trap quadrupole mass spectrometer (LTQ). An LTQ is one continuous cell, split into three array sections, designed to trap ions inside and then analyze them. A diagram of an LTQ is shown in Figure 2.3. An ac voltage at a constant radio frequency, known henceforth as RF, is applied to the rods to guide ions along the axis of the cell. Meanwhile three separate dc voltages are applied to the three sections to create potential wells that confine axial ion movement to be within the cell, and effectively trap the ions. Before each analysis, the trap opens for a discrete duration to accumulate ions, and a helium damping gas in the cell helps slow incoming ions to facilitate trapping by the voltages. During and after analysis, ions are ejected through slots in the center section, for which it is termed the exit rods.

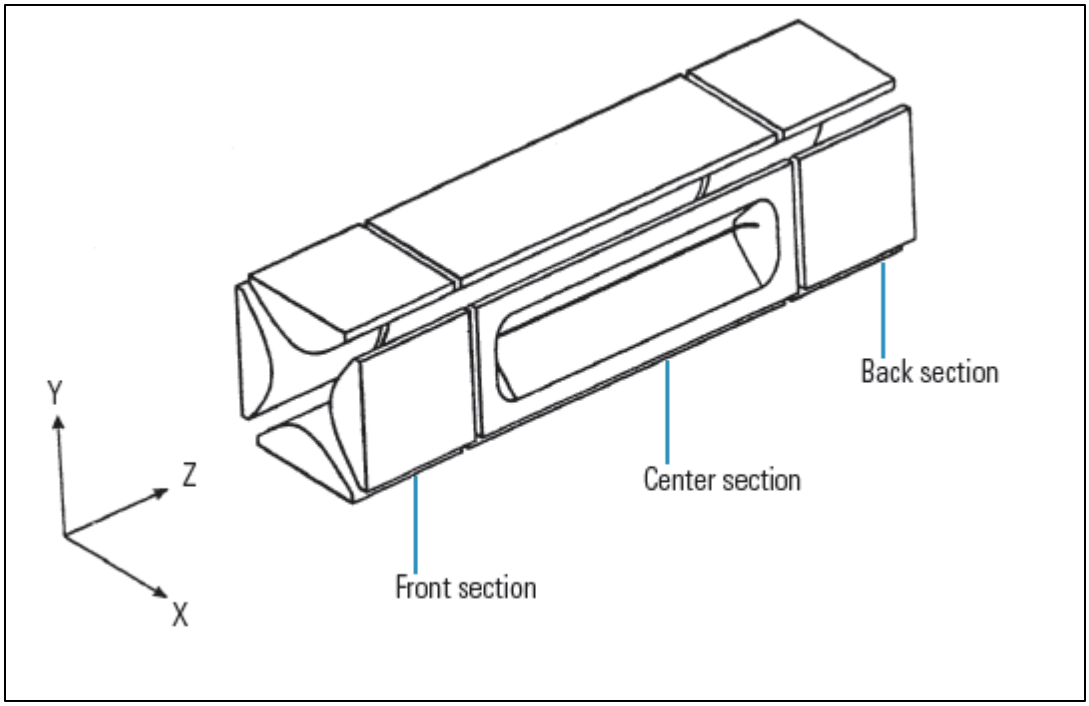


Figure 2.3 Basic diagram of linear ion trap quadrupole rod assembly, courtesy of Thermo Fisher Scientific[19]

To perform MS/MS, the LTQ accumulates ions and then conducts MS1 isolation, precursor fragmentation, and MS2 scan out. This MS/MS is termed tandem in time, as the three steps occur sequentially and in the same cell. The latter quality is enabled by applying, to the exit rods, an additional ac voltage that can vary in frequency or even comprise several at once. All ions in the trap oscillate at discrete frequencies as determined by both their *m/z*s and the RF amplitude; and when the ac's frequency matches an ion's, they resonate, and impart the ion with enough kinetic energy to cause its ejection from the trap or its collisions within.

During MS1, the ac is multi-frequency and can resonate with ions of a handful of *m/z*s at once. A ramp in the RF, from low to high amplitude, excites one handful after another into resonance to cause their ejection into waste. But a discrete gap in the ac-RF combination matches one specific *m/z*, which results in these precursor ions being isolated in the cell. The ac frequency then changes to resonate with the precursors, but without an RF ramp, so they stay in the cell and collide with the helium damping gas. The collisions convert the ions kinetic energy into internal energy, causing them to dissociate into fragments in a process known as collision induced dissociation (CID); the resulting fragments have different *m/z*s and thus different frequencies. Finally, MS2 scan out occurs by setting the ac to one frequency, and ramping the RF to move fragment ions into resonance, one at a time, according to their *m/z*. This results in the sequential ejection of fragment ions, but this time into the detector. It is important to specify at this point that CID predominantly produces b- and y-type ions [7, 20], and for the remainder of this thesis, all fragments can be assumed to be one of only these two types.

If just an MS1 scan is desired, as to detect any precursors present at a given time, the LTQ does not isolate or fragment, but only scans ions out as described above. This is important, as the LTQ can use MS1's to guide itself through global MS/MS, and consequently boost the efficiency therein. Specifically, the LTQ runs an MS1 scan and picks the most intense precursors detected, then it runs consecutive MS2 scans on each pick, starting with the most intense; user input predetermines the number of picks per MS1 scan, and number of MS2 scans per pick. This process is known as a *data-*

dependent scan, and it is repeated over the entire duration of a global MS/MS experiment. As the LTQ is picking the most intense precursors from an MS1, it temporarily ignores any precursors picked from prior MS1 scans, and this duration is also set by user input. This process is termed *dynamic exclusion*, and its purpose is to provide MS2 scans on precursors that may be less intense, but are no less important.

Global MS/MS experiments in this study were run with one MS1 scan followed by twenty MS2 scans, one each for twenty precursors, and a dynamic exclusion time of 1 minute. The resolution of a scan is determined by the mass of a detected ion, divided by the difference in mass between two adjacent distinct ions. In the case of peptides, with masses from 400 to 1700 Daltons, being able to separate a 0.5 Dalton difference requires a resolution of 800 to 3400. Both MS1 and MS2 have a resolution of 1000, which can separate peptides with an approximately 0.5 to 1.5 Dalton difference in mass, depending on the mass of the peptides being separated.

2.6 Operational Principles of the QQQ-MS System

A Thermo TSQ Quantum triple quadrupole mass spectrometer (QQQ) was used for all the targeted 1D-LC-MS/MS measurements in this study. A QQQ is three single-array cells that are isolated and arranged in sequence, as seen in Figure 2.4. There is no axial trapping, but rather a continuous ion stream travels from source to detector, through each cell along its axis. Ion oscillation stability is controlled through variable RF and dc voltages, where RF are applied to all three cells, but dc are applied only to the first and third. Again it is best to view the rods of each cell as two pairs, and the voltages applied to one pair of rods are of equal amplitude but opposite sign as those applied to the other pair, as seen in Figure 2.5. An additional dc voltage is used as an offset and will be referenced only as “offset” to prevent confusion. The offset is applied to all three cells, and with equal amplitude and sign for their two rod pairs. It controls ion acceleration, and thus controls ion translational kinetic energy (TKE).

MS/MS on a QQQ is termed tandem in space, because it is done by distributing the three steps, in sequence, among the QQQ’s three cells. The first cell, designated Q1, operates as the MS1 mass analyzer and isolates precursor ions, according to their

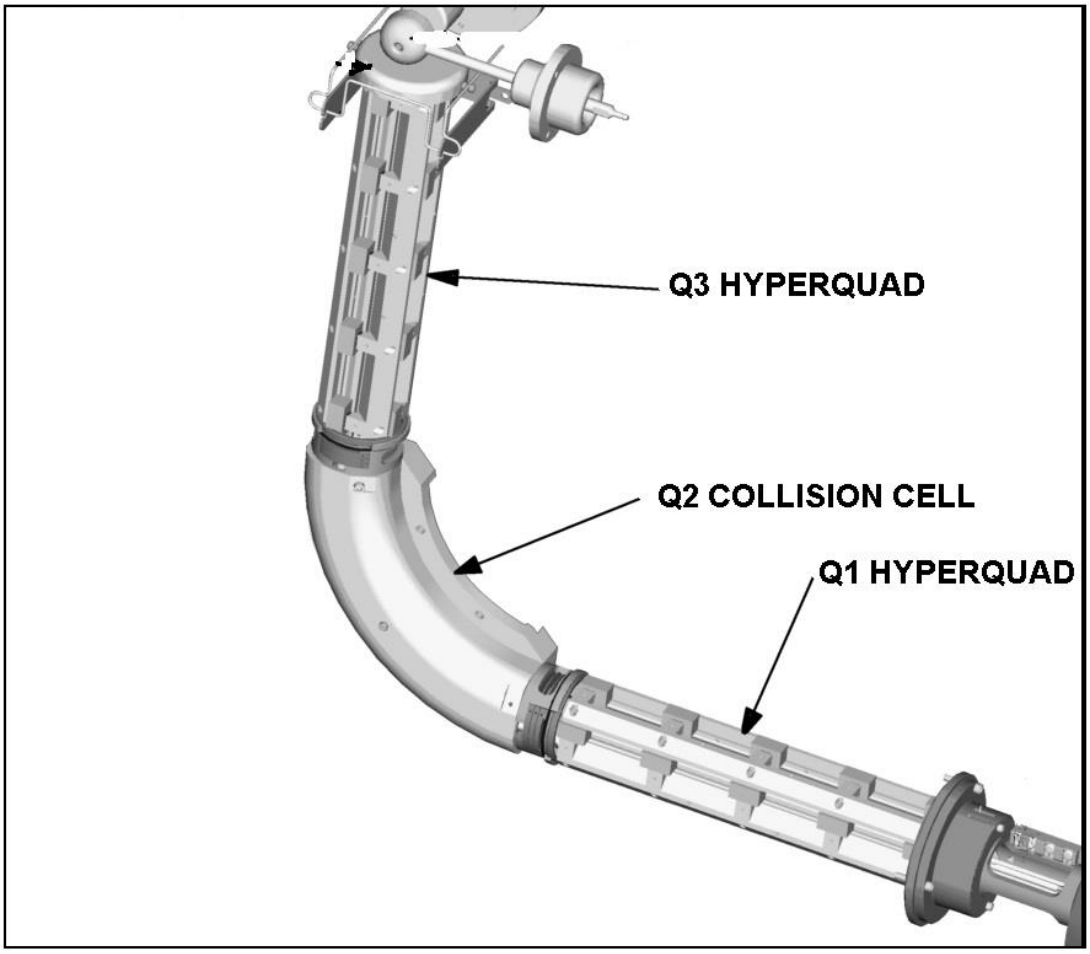


Figure 2.4 Basic diagram of a triple quadrupole mass analyzer, courtesy of Thermo Fisher Scientific[21]

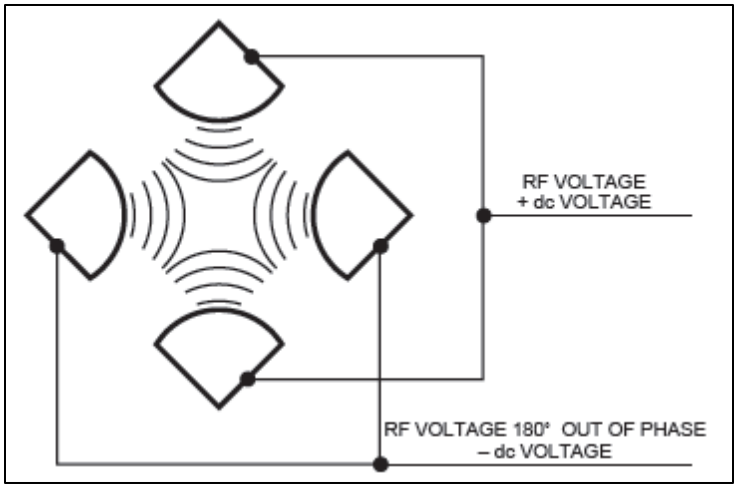


Figure 2.5 Diagram of electrical connection for the quadrupole rod assembly of the QQQ, courtesy of Thermo Fisher Scientific[21]

*m/z*s, by applying varying ratios of RF to dc. A given ratio of voltages will cause precursor ions of a specific *m/z* to have controlled oscillations and travel through Q1 to the next cell. Ions of any other *m/z* will oscillate out of control and either crash into the rods or be ejected from Q1. As the RF and dc are varied, their ratio is changed, and ions of a new *m/z* are brought into stable oscillation, while ions of the previous *m/z* join the others in instability.

Precursor ions, isolated in Q1, then enter the second cell, termed Q2, which functions as the collision cell. There is no dc applied to Q2, but there is a variable RF that provides stable oscillations to ions over a wide *m/z* range. Also, Q2 contains argon gas for the precursors to collide with and subsequently undergo CID; the energy for which is provided by an ion's TKE as determined by the offset. The ions follow a path, set by the voltages, through a curve in Q2 (Figure 2.4), while neutral molecules miss the turn and are ejected from the cell, thus this simple curve dramatically reduces noise in the spectrum. The fragment ions generated in Q2 then pass into the third cell, Q3, which is the MS2 mass analyzer. Q3 operates by the same mechanism as Q1, and isolates fragment ions one after another for passage to the detector.

There are a couple of important differences between MS/ MS on a QQQ and that on an LTQ. By using argon instead of helium as a collision gas, QQQ fragmentation is more extensive, and requires less kinetic energy, than that of LTQ. Because argon is bigger and heavier than helium, a collision with argon converts more of the ion's kinetic energy into internal energy than for a collision with helium. The other difference is MS/MS scans are faster on the QQQ than on the LTQ. Since the QQQ is tandem in space, there is no waiting for ions to accumulate before isolation begins, or waiting for one step to finish before starting the next. Precursor isolation, CID, and fragment isolation are simultaneously performed in their respective cells. However, the ion stream experiences them sequentially, and on the fly, as it passes from one cell to the next; and MS/MS takes only as long as ions going from source to detector. That is to say, tandem in space scans occur in real time in as little as 0.001s; this renders the QQQ ideal for targeted MS/MS, which in turn affirms its value to proteomics despite have unit resolution.

2.7 Targeted Proteomics with Multiple Reaction Monitoring (MRM)

The construction of a proteomic MRM method involves selecting several of a protein's unique peptides, then selecting several unique transitions for each peptide. This collection is called a transition list, and is used by the QQQ's operating software to guide the targeted detection of transitions. The resulting spectra can then be used for the selective identification and measurement of peptides. Naturally, as more transitions are detected with strong signals, confidence in the identity of the target peptide increases, and the possibility of an interference peptide being identified decreases. An interference peptide is a regularly occurring peptide with the same precursor *m/z* as the target, and is often called an isobaric peptide. Also, it can produce some of the same fragments, or fragment *m/z*s, as the target, which means that some transitions are shared among the target and interference peptides. It also means that as an interference peptide passes through the QQQ, it produces a signal for any shared transitions that are being monitored.

Interference can be remedied through careful transition selection, as exemplified by selecting larger peptides, selecting larger fragments, selecting more fragments, and using early MRM's as feedback for designing later ones. Larger peptides, between ~10 to ~25 amino acids long, make better selections for a couple of reasons. First, these peptides most often exist as +2 ions, and although they can form ions of +3 or higher, few of them have to in order fit the QQQ's mass range of 30 to 1500 Daltons. Second, a large peptide can produce large fragments, and a large number of fragments. Large fragments, such as b/y-6 to b/y-12 ions, have greater portions of their peptide's sequence, which gives the fragments a less replicable *m/z* and makes them more characteristic of the target peptide. And with a larger number of fragments, comes a lower chance that a significant number of them will be shared with any one isobaric peptide.

All of the above translates into large peptides having more transitions that are more unique and more likely to occur. Additionally, targeting numerous transitions for each peptide in an MRM provides more evidence on which to determine authenticity. Finally, the results of the MRM can be used to identify which transitions are shared, which are

unique, and how the unique compare in signal strength. With this information, the transition list can be purged of shared transitions followed by the weakest, until 3 to 6 of the strongest and most confident transitions per peptide remain for further MRMs. This dramatically improves confidence in peptide I.D. and subsequently protein I.D. It is with this high level of confidence exhibited in proteomic MRM, that the QQQ deflates the effects of low resolution.

As important as it is to have many peptides and transitions, there are limitations and performance tradeoffs. One explicit limitation in the QQQ software's operating parameters is that the transition list is limited to 320 transitions and cannot have duplicates. Notable performance tradeoffs exist between list size and duty cycle, as well as scan time and sensitivity. Duty cycle is a transition's scan time as a percentage of the list's scan time. Only one transition can be scanned at a time, and the entire list must be scanned before starting again; so as the list increases, the scan time increases, and transition duty cycle decreases. And although a transition can be scanned in 1ms, it comes at the expense of lost sensitivity; conversely, in order to increase sensitivity, scan time must increase as well. Thus, even if scans take only tens of milliseconds, having a large list scanned with high sensitivity means those milliseconds add up; meanwhile, the peptides continuously flow off the column, each for a limited amount of time, and many for as little as 30 seconds. With a list of 320 transitions, each transition has a duty cycle 0.31% of the scan time, and at 20ms per scan, each will be scanned once every 6.4s. If a peptide elutes for 30 seconds, a transition will only be scanned 4 or 5 times, which means at best 5 data points will be acquired to plot a transition's entire profile. But a list of 75 transitions provides each transition with a duty cycle of 1.3%, which translates to 20 data points. Therefore, it is best to limit the number of peptides per protein, and transitions per peptide to the top 3 to 6 each. And when an experiment requires a large number of transitions, there will need to be a compromise between the amount scanned in one run, and the quality of each. The targeted MS/MS experiments in this study were run with a scan time of 20ms, with the offset voltage individually set for each transition as will be explained below. The transition lists ranged in size from the mid 70's to the 310's depending on

the nature of the MRM experiment.

2.8 Bioinformatics Methods for Data Analysis

Proteomic MRM design requires prior knowledge on proteins, peptides, and transitions. Specifically, transition list selection requires data on which peptides are unique to protein, and which precursor and fragment *m/z*s represent the most unique transitions; this data must also provide the suitability of each peptide, precursor, and fragment for MS/MS detection. As will be extensively discussed in chapter 4, this data can be generated and evaluated through empirical measurements or theoretical calculations. However, neither route is possible without precise sample information that includes organism identity, protein sequence, protease identity, peptide length, peptide sequence, number of missed peptide cleavages, fragmentation mechanism, and fragment ion type. Reading, matching, calculating, and processing this information would be impossible without a work belt full of purpose built software tools.

As mentioned above, this study presents two methods to gauge which peptides will be best detected in MRM. The first will be called the empirical method, and employs measurements attained with traditional global MS, such as those made on the LTQ. The second method will be termed the *ab initio* method, and like its namesake, it relies solely on theoretically calculated probabilities. An exhaustive comparison of these two methods is offered in chapter 4, but what follows is a comprehensive list of the programs and scripts used by each method, along with a brief description of their functionality.

Empirically employed software extracts a variety of technical data on proteins and peptides from MS1 and MS2 measurements, which are generated by global MS and contained in a raw file. After an MS experiment, the raw file and a sample-specific protein database, called a FASTA, are loaded into MyriMatch[22], which is also given the protease I.D., fragmentation mechanism, and number of missed cleavages to allow. MyriMatch then predicts peptides from the FASTA according to the given settings and reads the raw file for matches. These matches are displayed in IDPicker[23] and provide identities for proteins and their peptides, along with showing the associated

peptide MS2 spectra. Having a list of each protein's peptides, and knowing how they fragment, is the first step in selecting peptides and transitions.

Picking the best fragments for each transition can be done straightforwardly by looking at a peptide's MS2 spectra. But selecting the best peptides to start with takes further navigation, and offers several paths that focus on different qualifiers. Two paths were followed in this study, one in chapter 3 and one in chapter 4, and each was laid out with different software. The characterization study in chapter 3 required software that assisted in defining discrete time bins. This same software was used to evaluate each peptide's abundance for their guided selection. Chapter 4 had no such time constraint, which allowed the use of software that provided a more comprehensive evaluation of peptide abundance.

The first path used an in-house script called gitR_MS that reads IDPicker results and matches peptides to their MS1 scan number. Then, MASIC[24] reads the raw file mentioned above to match MS1 scan numbers with their scan times and the ion current for each *m/z* in a given MS1scan, and it calculates the total ion current (TIC) for that scan. The peptides are matched to their MS1 scan times and ion currents to simultaneously provide the time stamp and abundance. Here, peptide abundance is defined by a peptide's percent contribution to the TIC of the MS1 scan in which it was found.

The second path is laid out by another in-house script named POSI, which uploads the IDPicker files and the FASTA to match peptides to their multiple MS2 scan intensities, a combination termed matched ion intensity (MII). The MIIs for each peptide were summed and averaged across the number of MS2 scans for that peptide; peptide abundance in this path is defined by the log of the product of a peptide's summed MII and averaged MII. This means that abundance here considers both the quantity and quality of MS2 scans for a peptide.

The software employed by the *ab initio* method provides probabilities and scores to assist in transition selection, and operates by internal calculations done on user provided input, according to user defined parameters. Peptide selection was guided by PeptideSieve[25], which uploads a FASTA and is given the protease I.D., peptide

lengths to allow, number of missed cleavages to allow, and ionization method. It generates a list of predicted peptide sequences, and assigns them detection probabilities based on their amino acid composition and physiochemical properties. After picking peptides, fragment selection begins with assistance from PepNovo[26]. PepNovo is given a list of peptide sequences and charge states, and instructed on what fragmentation method, protease I.D., and post translational modifications (PTM) to consider. It gives the top fragments for each peptide, the number of which is user defined, and provides each fragment's score based on length and amino acid composition.

It's important to note here, that the outputs from programs and scripts mentioned above were loaded into excel for further extensive processing. Data processing tasks performed in Microsoft Excel included, but were not limited to, searching, matching, filtering, sorting, grouping, summing, averaging, ranking, highlighting, and graphing.

Although reams of informative input on MRM design came from several programs and scripts, the actual MRM design and evaluation was performed solely with Skyline[15]. Like some of the programs above, Skyline uses a FASTA and protease I.D., and with these it generates a hierarchical transition list. The highest rank list is of the proteins to be searched for and characterized, and each protein can be expanded to display a list of its possible peptides. Each peptide can then be expanded to show its possible precursor *m/z*s, and finally each precursor can be expanded into a list of its possible fragment ions. This tiered approach can be seen in Figure 2.6. In Skyline, any proteins, peptides, precursors, or fragments can be manually deleted from the list; doing so, as guided by the scores and probabilities mentioned above, can provide a short list of transitions with high probability and high duty cycle. Additionally, it can predict each precursor's optimum fragmentation energy and provide the corresponding offset voltage. Once finished, Skyline can export a transition list file, which contains for each transition: the precursor *m/z* to monitor in MS1, the offset voltage for optimum fragmentation, and the fragment *m/z* to monitor MS2. Finally this file is uploaded to the QQQ and guides it through an MRM experiment.

The use of Skyline does not end at exporting a transition list; when the MRM is

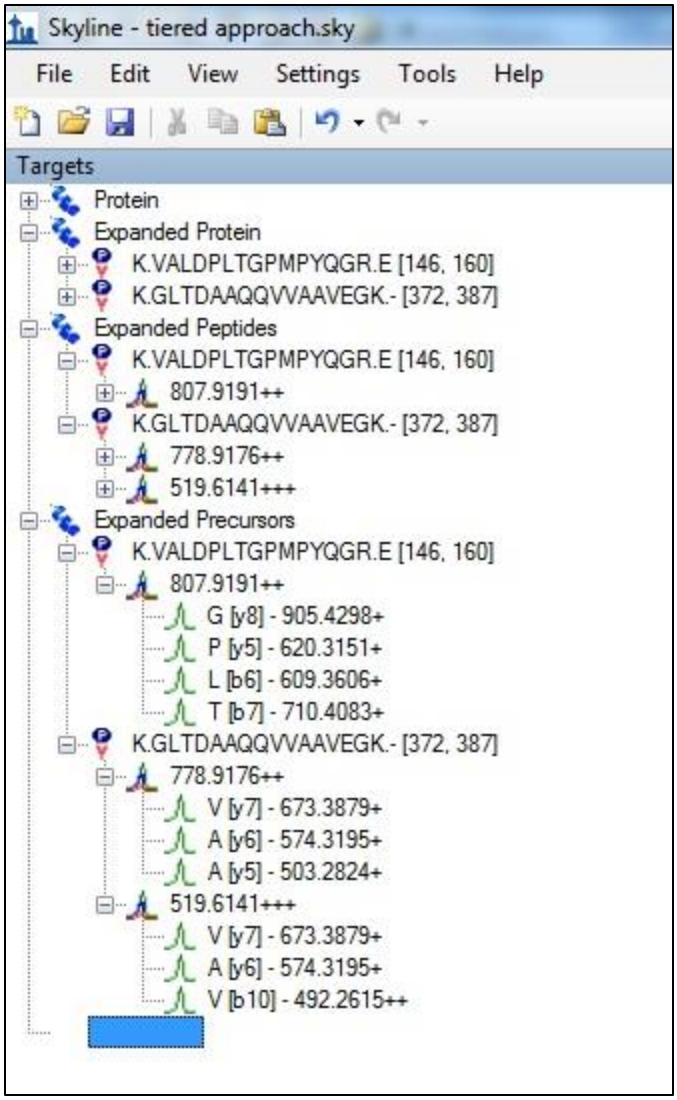


Figure 2.6 Tiered display of proteins, peptides, precursors, and transitions as seen in Skyline[15]

done, the resulting raw file is uploaded into Skyline to evaluate the MRM design based on each transition's performance. Figure 2.7 demonstrates how the results are presented in Skyline; it displays an overlay of the chromatograms of all transitions for one peptide precursor at a time, and the intensities are shown relative to the highest available for a given zoom. This overlay allows the user to quickly see which transitions produced the highest peaks, which transitions have the least noise, and where on the chromatogram transition peaks overlap. The latter is the most important in authenticating that peaks originated from the desired peptide, and this will be discussed further in chapter 3. Beyond displaying the chromatograms, Skyline also evaluates their signals at a selected peak and offers individual ranks for each transition present in a given selection, as shown on the left of Figure 2.7. Utilizing the chromatograms and ranks, facilitates transition list improvement by showing which to keep and which to discard or replace depending on the number.

Once the transition list is optimized, it can be used to run multiple technical replicates of an MRM experiment, the results of which will again be loaded into Skyline. The replicate chromatograms provide a ready visual for comparing retention times, peak intensities, and peak area ratios. To facilitate this visual evaluation, Skyline can produce bar charts and scatter plots of the transitions peak areas and retention times as comparisons across peptides or across replicates. It can also provide numerical data for exact comparisons of each peptide's reproducibility by exporting a report on the MRM results. Among a multitude of other metrics, a report can contain each transition's retention time and peak area, and each peptide's retention time as a function of its best transition, and peak area as a sum of its transitions. The report automatically displays the data by replicate, but it can provide averages over all the replicates, along with ranges, standard deviations, and coefficients of variability. By generating chromatograms, charts, and reports, Skyline offers a comprehensive view of the results, and provides a self-sufficient platform for evaluating MRM design.

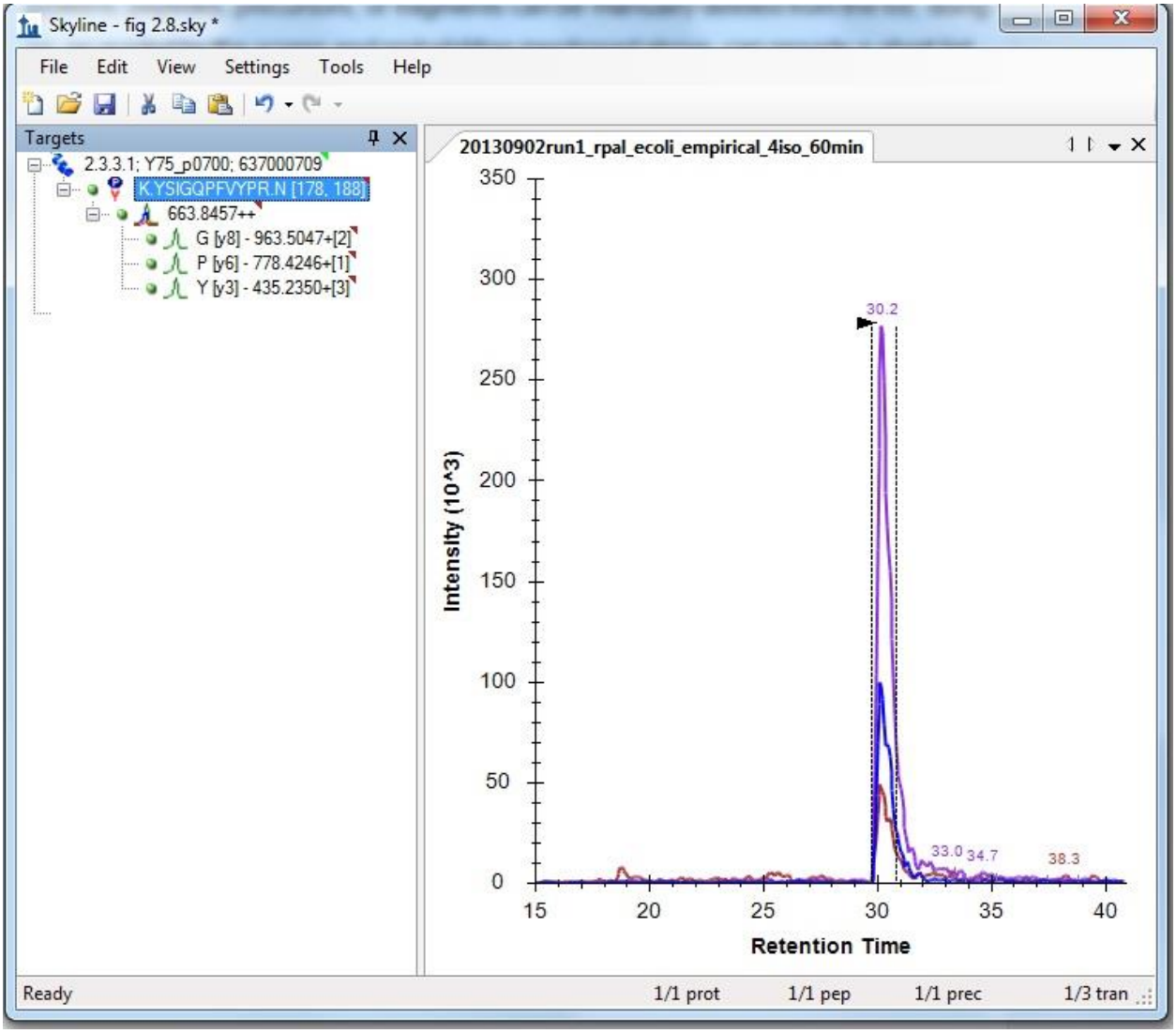


Figure 2.7 MRM results with transition ranks as displayed in Skyline[15]. Upper left portion displays tiered approach, while different colored lines on the right indicate separate transitions.

2.9 Summary

This chapter laid out the experimental design of each project to provide a context for the explanation of the materials and methods presented in this study. Within this context was a description of the content of the 4-iso sample used for both projects, offered as an argument of how this sample would challenge the scope, and characterize the effects, of the complexity of a sample analyzed by MRM. The techniques used for sample preparation and loading, as well as the LC method employed, were consistent throughout each project; and as they lack offline fractionation or secondary chromatography, they further emphasize the difficulty present in the analysis of this sample. By explaining the operation of the spray source, it was shown how ionization generates multiple peptide precursors. Operations of the both LTQ and QQQ were covered and comparisons were made as to how each approaches tandem MS. These comparisons showed how LTQ data functioned for both defining congestion and empirically designing MRM, while the QQQ was the workhorse for all MRM experiments and provided data characterizing and evaluating MRM performance. Finally, the software was laid out to show how data was processed to provide definitive measurements on MRM reproducibility in a complex sample.

Chapter 3: Peptide Peaks Growing Up in a Tough Chromatographic Neighborhood: Characterizing MRM Reproducibility and Robustness in Complex Microbial Samples

3.1 Design of MRM Experiment

MRM is typically performed using one dimensional chromatography, which can be problematic for very complex samples due to challenges from incomplete separations and ion suppression leading to diminished reproducibility. The first project of this thesis was tailored to evaluating such reproducibility under these conditions so that a systematic method of selecting target peptides could be developed for MRM in complex samples with 60 minute 1D-LC. This chapter chronicles the design and execution of this project and begins with discussions on identifying congestion and abundance as factors to be characterized for their influence on MRM quality, as well as the reason for limiting the number of target peptides. Then, the mechanics of filtering the pool of peptides down to the desired number will be detailed to explain how such filtering defines areas of different chromatographic congestion. This will be followed by a description on how three levels were established for peptide abundance, and how peptides were selected to represent each level. Next, preliminary MRM will be presented and justify the removal weak peptides from final MRM consideration. The results of these MRMs will be the subject of the last sections of this chapter, which will comment on the viability of MRM in complex samples, the influence of predicted factors, and the identification of any unpredicted factors shown to have influence.

MS/MS analysis of peptides often employs complicated fractionation and secondary separation techniques to minimize congestion in chromatographic flows to make low abundance peptides more detectable. This is true even for simple samples such as blood or urine. But with these simple samples, MRM is a regularly selected approach to bypass additional separations as its targeted nature excels at ignoring congested areas of the chromatogram to detect peptides of low abundance. Since complicated samples present a substantial increase in congestion, and thus further bury low abundance peptides, it was decided that these two factors, chromatographic congestion and

peptide abundance warranted characterization during the evaluation of MRM reproducibility in complex samples.

Only 12 to 36 peptides would be utilized for MRM evaluation, as this is a comparable amount to what is targeted for MRM exploration of a metabolic process, and the resultant MRM design would be tested on how reproducibly it measured this handful of peptides from among more than 80,000 others in a 60 minute 1D-LC-MS/MS. These targets would be selected from a pool of peptides identified by a series of global MS/MS measurements. The peptides selected from this pool would have to possess the traits necessary to characterize the effects of chromatographic congestion and peptide abundance on MRM performance. The peptides would also undergo preliminary MRM to determine which, if any, of their transitions were suitable for testing reproducibility.

3.2 Characterizing Viable Transitions from Selected Peptides

Four technical replicates, of a global measurement on the LTQ, yielded 3,827 non-redundant peptide identifications. To be considered candidates for the MRM evaluation, identified peptides had to be reliably detected and time-specific. Table 3.1 contains a partial list of these 3,827 peptides and their associated metrics. Five peptides on this table, along with their scan times and replicate I.D. counts, are highlighted to exemplify how a peptide was, or wasn't, deemed reliable and specific.

Peptides were considered reliable if they appeared in all four replicates, as shown on the right of Table 3.1 in green. This criterion is adjustable and could be relaxed to provide more candidates if needed; for example, investigating a specific biological process may require more options than what stricter criteria allow. However, selecting peptides with lower reliability, such as those shown in yellow, risks diminishing MRM reproducibility.

A peptide's time-specificity was defined as the peptide having appeared in only one 5 min window per replicate, and the same five minute window across each replicate; these five minute windows are referred to henceforth as bins. Examples of peptides that were specific to one bin are shown on the left of Table 3.1 in yellow, while peptides identified in more than one bin are in orange. It was considered essential that

Table 3.1 Excerpt of peptide sequences with first and last scan times, and replicate count. Qualities in green are favorable, qualities in yellow are unfavorable. Peptides in blue qualify as candidates; peptides in red do not qualify

Sequence	scan time$_{earliest}$ (min)	scan time$_{latest}$ (min)	Replicate Count
SKEHTTEHLR	0.00965	0.00965	1
SFSHQAGASSK	0.08764	0.08764	1
AQASTHGIGK	0.17481	4.74361	4
KQLDHGQK	0.42847	0.42847	1
TGRNPQTGK	0.51986	0.51986	1
TQDATHGNSLSHR	0.53575	3.65567	2
KLKDEAAK	0.7896	0.7896	1
APAAAAPAAK	1.4402	8.07177	4
SHALNATKR	1.46481	1.46481	1
VYVNKDDTTK	2.51166	6.18079	2
KVHPNDDVNK	1.55398	1.55398	1
MEQELHHR	1.71655	3.65567	2
APHVSEK	1.73009	1.73009	1
DAGGTAEAVR	6.52622	8.95852	4

candidates be specific to a bin so that the chromatogram could be reliably divided into segments of equal time but differing congestion.

Peptides such as those in blue were both reliable and specific, and were thus candidates for MRM analysis. Peptides such as those in red lacked reliability, time-specificity, or both, and thus were not MRM candidates. 415 of the 3,827 identified peptides were selected as candidates for MRM analysis.

The congestion of each bin was determined by how many of the 415 candidate peptides were unique to it; Figure 3.1 and Figure 3.2 provide a visualization of how each one differed in congestion. These figures show that the most congested bin is at 25 to 30 minutes, and that the majority of the peptides were identified between 10 and 45 minutes. This is because most peptides elute when the mobile phase is between 5% and 40% ACN[27], and the gradient employed for this project increased from 5% to 60% over 55 min. With chromatographic congestion adequately defined, its effect on MRM quality would be characterized by examining how a peptide from one bin performed relative to peptides from other bins. Performance would be measured in terms of a peptide generating viable transitions that produced large sharp peaks with consistent areas and retention times.

In order to evaluate how peptide abundance impacts MRM quality, different levels of abundance were established, as displayed in Figure 3.1 and Figure 3.2, but it is important to note that because abundances were often based on different MS1 scans, comparing two ranks is not a comparison of absolute abundances, but of relative abundances. That is to say, if one peptide were scanned among less abundant peptides, while another was scanned among more abundant peptides, the former will have a higher rank for its bin, even if the latter is more absolutely abundant.

Peptides with an average percentile rank of 75% to 100% were considered high abundance level, peptides with 50% to 75% were medium level, and peptides with 25% to 50% were low level; Figure 3.2 gives the abundance breakdown by bin. The effect of peptide abundance on MRM quality would be characterized by comparing how high, medium, and low abundance peptides performed relative to each other. Depending on each bin’s availability, 3 peptides would be selected; the highest of the high abundant,

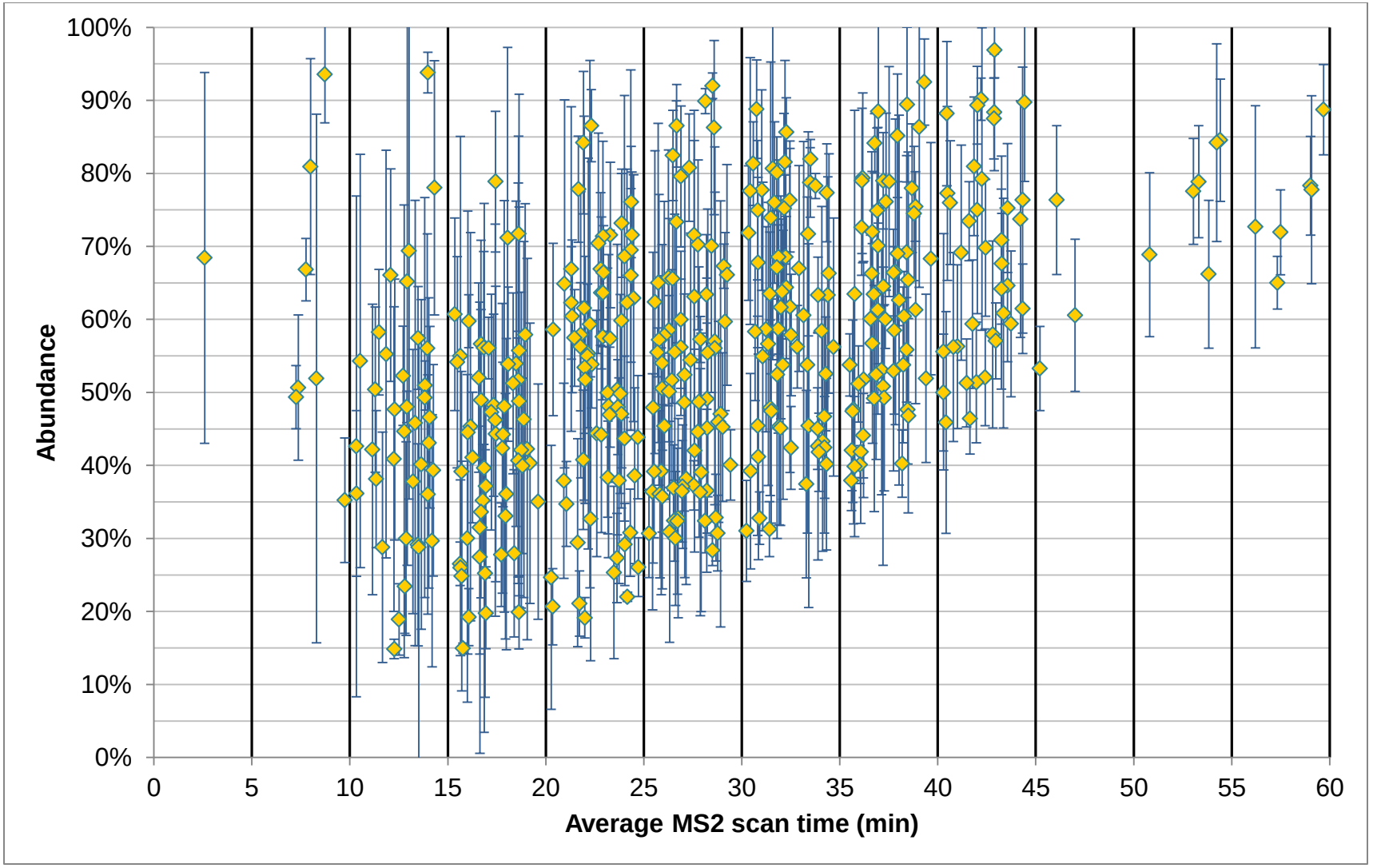


Figure 3.1 Scatter-plot of peptide scan time versus abundance. Error bars are provided for abundance.

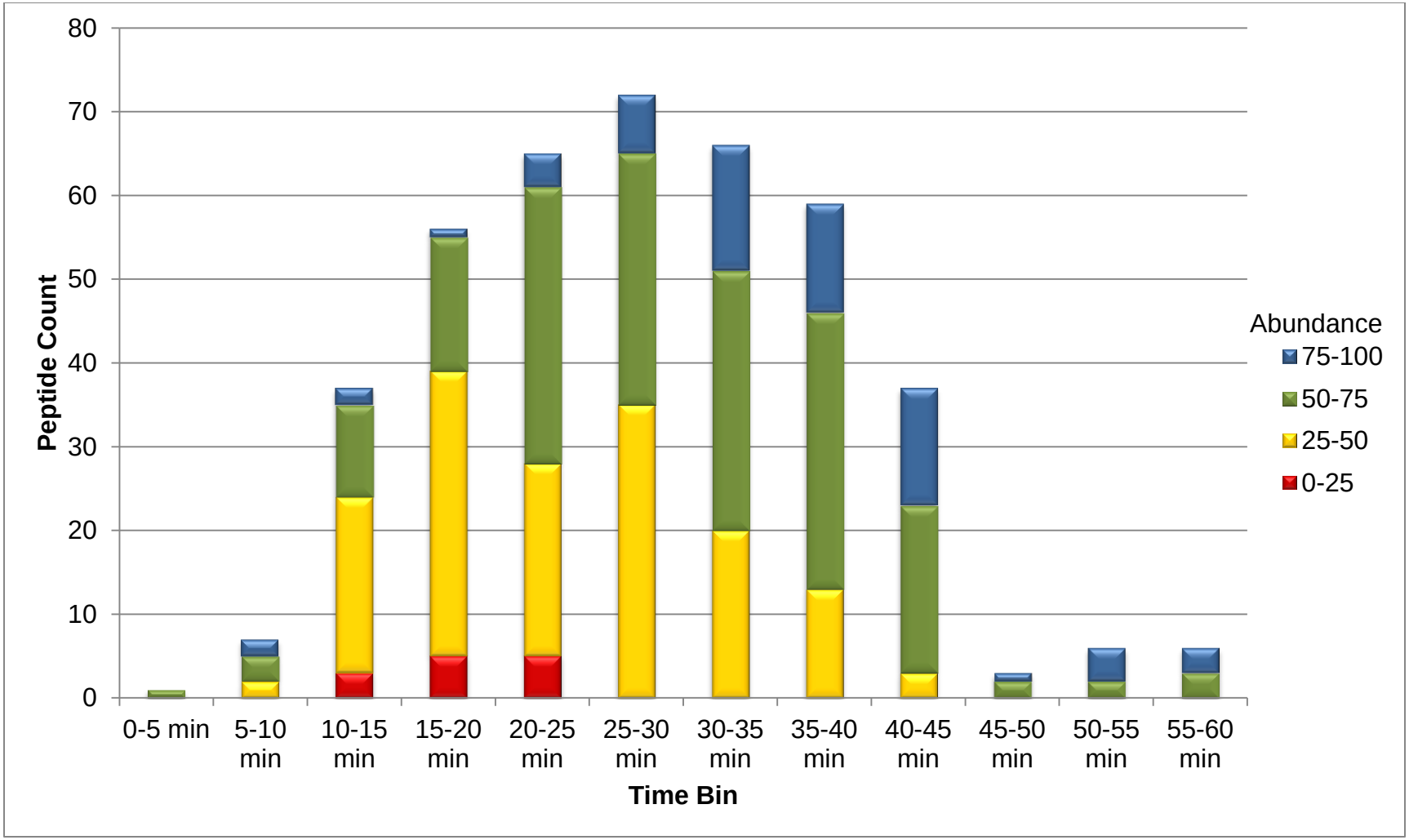


Figure 3.2 Bar chart of total peptide counts per bin with breakdown by abundance

the median of the medium abundant, and the lowest of the low abundant. The limited availability of peptides from each abundance level within some bins led to a total of 31 selected peptides. Table 3.2 lists these 31 peptides, with their respective percentile ranks, and grouped into their respective bins. It can be seen in the table that 4 bins did not have all three possibilities. As there were only 13 of the 415 peptides with a rank below 25%, and considering that these were only present in 3 of the 12 bins, it was decided that none of them would be used to evaluate the effect of peptide abundance.

3.3 Evaluation of 31 Selected Peptides by MRM

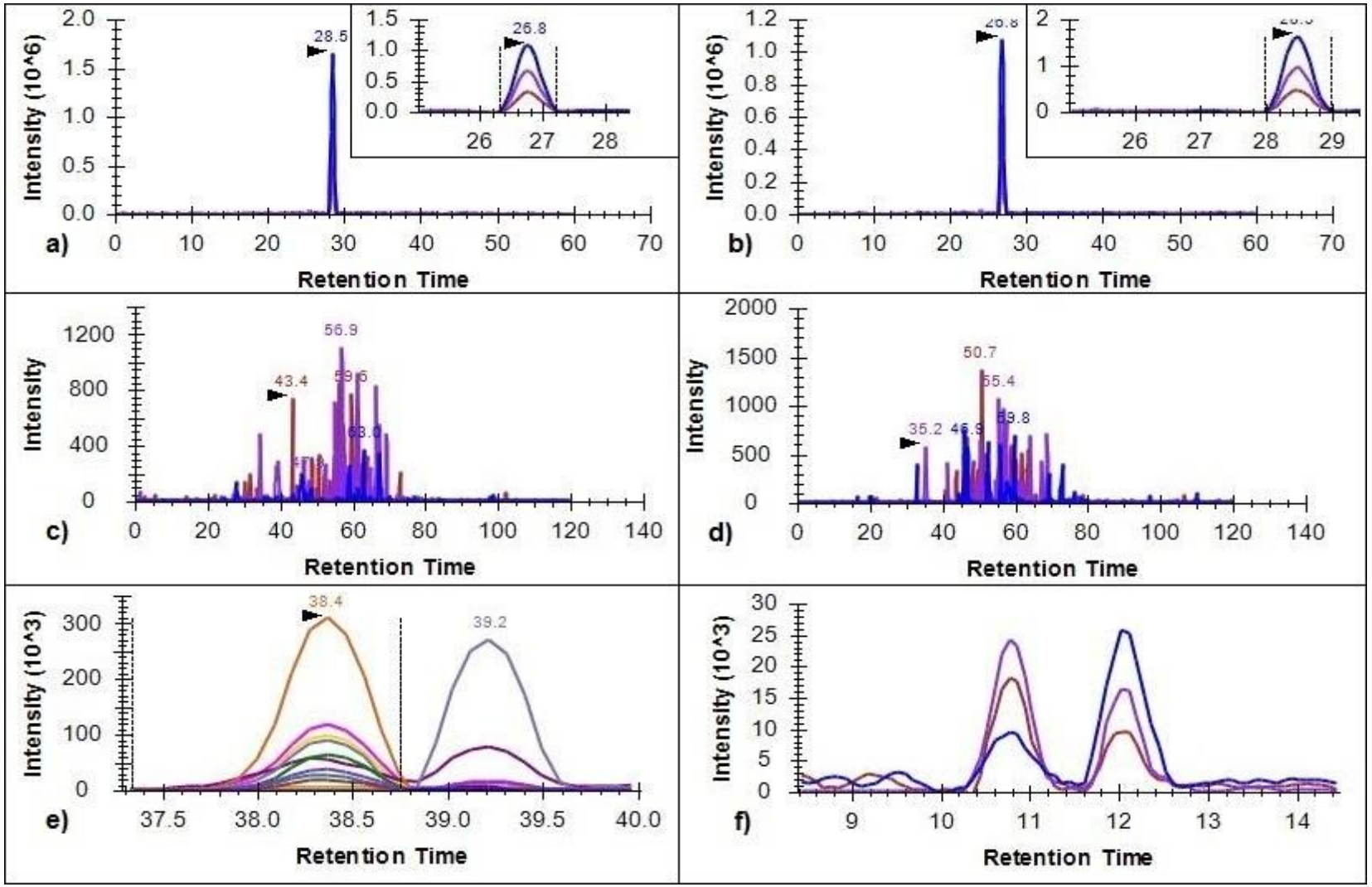
The 31 representative peptides were subjected to preliminary MRM scans on the QQQ to determine which, if any of their transitions were suitable for a reproducibility study. Skyline software was used to predict the possible transitions for each of the 31 peptides, and to generate a target list for use in the preliminary MRM's. Due to the complexity in examining all possible transitions in a single MRM measurement, multiple MRM's were done to scan for the 720 possible transitions. The results were then analyzed in Skyline to determine which transitions were suitable for MRM evaluation, and which peptides had a sufficient number of transitions. Suitability was based on consistency, signal strength, evidence of peak origin, and lack of ambiguity. Figure 3.3 shows 6 results of the preliminary MRM measurements. Each is a measurement of transitions predicted for a target peptide, and illustrates how the best transitions were selected for the final comprehensive MRM's.

Transitions demonstrated consistency and signal strength through regular retention times and a peak intensity of at least 10^4, with minimal noise. If a predicted transition only produced random, low intensity peaks, then there would be too much difficulty in distinguishing if the signal came from a fragmenting target peptide or just random noise. Figure 3.3a and Figure 3.3b each show a spectra generated by the same three transitions. The peaks displayed are of high intensity and regular occurrence, which evidences fragmentation of the target peptide; hence these three transitions provide useful information on their peptide. Figure 3.3c and Figure 3.3d also each show a spectra generated by another same three transitions, but the peaks displayed are

Table 3.2 The 31 selected peptides showing abundance and time bin. Shadowing is to assist in time bin differentiation

Sequence	Abundance	Time Bin
N/A	N/A	0-5 min
AQASTHGIGK	68.43%	
N/A	N/A	
TPALAAK	93.56%	5-10 min
NGGVAGNTTVNQK	66.82%	
LVEGSAQVK	35.23%	
TATEYGVVR	93.80%	10-15 min
KVVVEYPK	65.19%	
VMQAQGSQLTNK	28.79%	
STC[57]TGVEMFRK	78.87%	15-20 min
VDDGGTLDVR	60.72%	
LGSHNDMTFGEGTSSR	25.20%	
LGQMGELVR	86.53%	20-25 min
TVSENEVPLYK	62.27%	
EFNVEANVGKPQVAYR	25.32%	
IEIPGC[57]SLC[57]MGNQAR	91.99%	25-30 min
GPASVTNEQIEQVVR	62.37%	
FDGNAC[57]VLLNNNSEQPIGTR	28.36%	
LAATIAQLPDQIGAK	88.78%	30-35 min
GIVDSNLGLSPATEGQVIR	61.73%	
VRELTVQATTGTNSESDLSSIQDEIK	31.04%	
VALYGIDYLMK	92.50%	35-40 min
VIDLMC[57]PFAK	62.64%	
AVIFAGELLK	37.94%	
ANQVPQQVLSLLQG	96.89%	40-45 min
VPDIGADEVEITEILVK	61.45%	
DIQLATPPQVGAPATEYAALAELK	45.87%	
VGAGPFPTELFDETGEFLC[57]K	76.35%	45-50 min
GITLPETELR	60.57%	
N/A	N/A	
ALLNSMVIGVTEGFTK	84.54%	50-55 min
QSIAVSALSLANQSQQGVLQLLR	66.18%	
N/A	N/A	
VLLPVPFALINDPFGK	88.72%	55-60 min
LMEQITTSDELIDFLTLPGYR	65.03%	
N/A	N/A	

Figure 3.3 Excerpts from Preliminary MRM Results. Different line colors indicate different transitions. a & b) Show two replicate measurements of the same transitions for the peptide sequence AQASTHGIGK, which have high intensity and minimal noise, and the insets show the transitions coincide, which makes these ideal for final MRM. c & d) Show two replicates of the same transitions for QSIAVSALSLANQSQQQGVLQLLR with low intensities and considerable noise, making the unsuitable for final MRM. e) Shows that some transitions for GITLPETELR also belong to an interference peptide of similar *m/z*, but there is sufficient evidence to determine the leftmost peaks are from the target peptide, while the others are from noise. f) Shows transitions for TPALAAK that also belong to an interference peptide, but there is insufficient evidence to determine target from noise.


Intenslty (10^6)
28.5
26.8
a)
Retention Time
Intenslty (10^6)
b)
Retention Time
Intenslty
56.9
43.4
c)
Retention Time
Intenslty
50.7
35.2
d)
Retention Time
Intenslty (10^3)
38.4
39.2
e)
Retention Time
Intenslty (10^3)
f)
Retention Time

randomly occurring and low intensity. It is impossible to determine which, if any, of these peaks are from the target peptide fragmentation, which makes these transitions useless.

Suitable transitions also needed evidence that their peaks originated from the target peptide, and not an interference peptide. This evidence of origination is provided by a transition peak being part of a large set of coinciding transition peaks. Figure 3.3e shows two sets of coinciding peaks. Both sets appeared because the transitions of each are predicted for the target peptide. But only one set came from the target; the other set comprises shared transitions and, came from an interference peptide. Since Skyline predictions were tailored to target peptides, there should be more unique transitions than shared, and so the target set should have more peaks. Thus, when determining which peak set in Figure 3.3e originated from the target peptide, there is more evidence for set on the left than the set on the right.

Finally, suitable transitions were required to be free of ambiguity, which was demonstrated by generating only one coinciding set of high intensity peaks at a regular retention time. Again, Figure 3.3a and Figure 3.3b show three predicted transitions, whose high intensity peaks only coincide once, and so could only come from one peptide, the target peptide. In Figure 3.3f there are three predicted transitions producing high intensity peaks that coincide twice, and thus match to two different peptides. The target peptide could be represented by either set, at least one set is from an interference peptide, and it is possible that both sets are from interference peptides, thus these transitions must be thrown out altogether.

Some transitions met many of the above criteria, produced strong spectra, and appeared to be reliable candidates by initial inspections, but careful scrutiny yielded subtle, yet fatal flaws. The insets in Figure 3.4a show two multi-transition sets, either of which could be from the desired peptide or an interference peptide. The predicted retention time favors the latter set, suggesting it is the desired peptide. However, Figure 3.4b shows only one transition set, yet the predicted retention time fails to match. This set is reliably believed to be from the target peptide, which would indicate that these complicated 1D-LC separations interfere with the reliability of current retention time

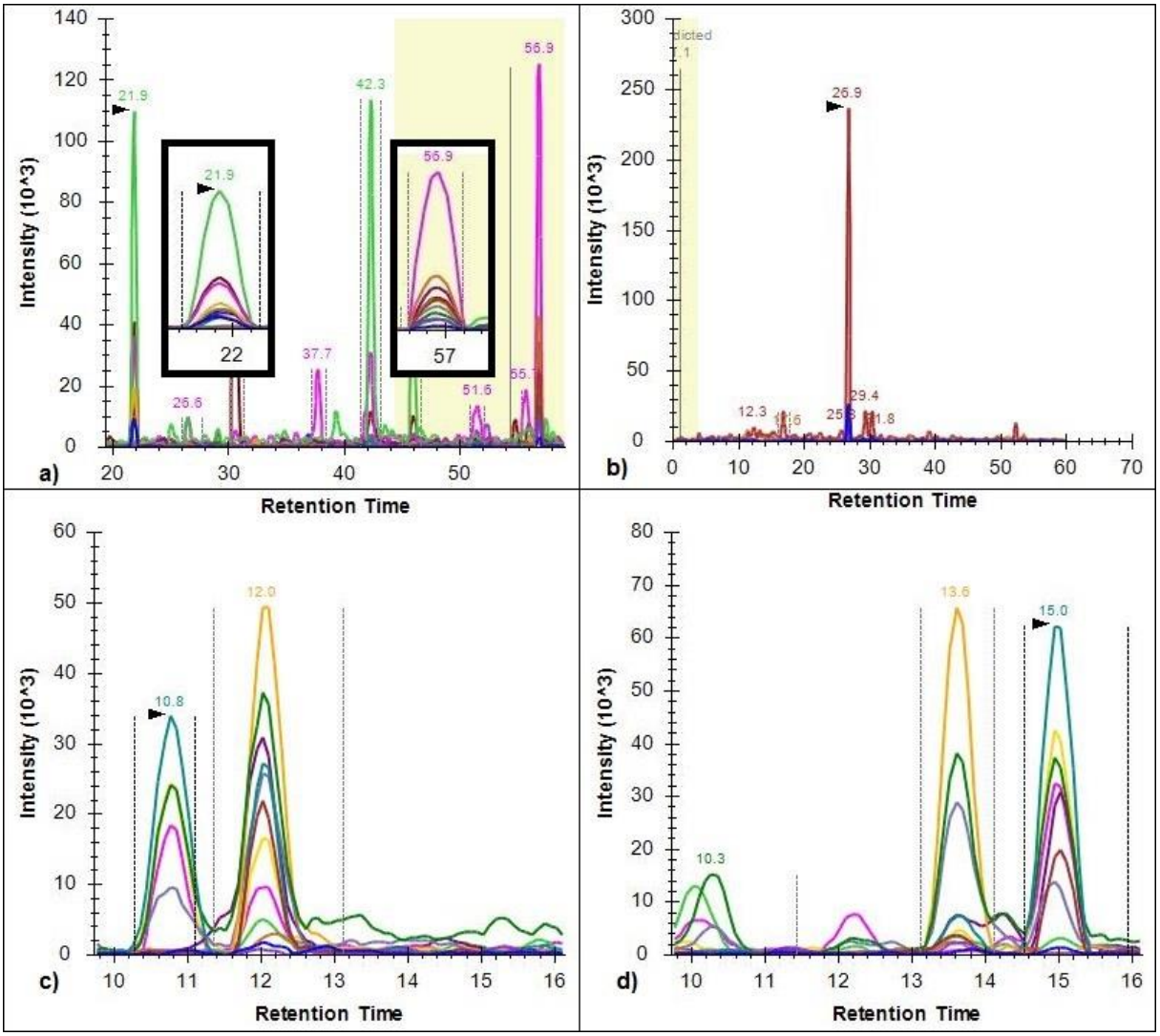


Figure 3.4 Spectral anomalies resulting in MRM ambiguity. As before, different colored lines indicate different transitions. a &b) Show inconsistency in predicted retention time. a) Transitions for VALYGIDYLMK also match to interference with equivalent evidence, but retention prediction favors the rightmost peak. b) However, transitions for AQASTHGIGK only match the target peptide while retention prediction fails to match. c & d) Show two replicates of the same transitions for TPALAAK. These display further ambiguity as peptide evidence and retention times switch between replicates.

predictions. Thus, the ambiguity in Figure 3.4a remains unresolved, and such peptides must be discarded. Figures 3.4c and 3.4d show transition sets that are too close in time, and even switch order. This makes it impossible to pick which transitions represent the selected peptide, even if predicted times were reliable. Further complication comes from the sets sharing at least four of their top transitions, thus peptides like the one in Figures 3.4c and 3.4d must be discarded.

These target peptides could be salvaged as viable candidates with newer and more discriminating methods for identifying the representative transitions. Calibrating the predicted retention time calculator to better reflect unfractionated samples could help the problem in Figure 3.4a, while peptides like the one in Figures 3.4c and 3.4d would benefit from reliable transition score predictions that accurately show the rank of the fragments for the target peptide.

Of the 31 selected peptides, 10 lacked suitable transitions and had to be discarded; 4 of these suffered from ambiguity created by interference peptides, and 6 produced poor signal to noise ratios. Table 3.3 shows the ten discarded peptides and highlights a few important observations. First, peptide abundance appears to have no effect in determining which peptides are discarded. Also, chromatographic location shows no apparent trend with a peptide having good transitions, but it does show a trend of early eluting peptides suffering more from ambiguity, while later eluting peptides suffered more from poor signal to noise.

The remaining 21 peptides were trimmed to 3 or 4 of their best transitions, and a list was created to generate MRM results to evaluate reproducibility in a complex sample. Figure 3.5 is an example of the MRM results analysis from Skyline. The best spectrum (Figure 3.5a) came from a medium abundance peptide and show a peak with very high intensity and no noise, but the worst spectrum (Figure 3.5b) is of greater interest because it was obtained from a low abundance peptide and demonstrates how bad a spectrum can look and still be useful. In spite of the contrast between Figures 3.5a and 3.5b, the latter's set is still of sufficient intensity and the only one with coinciding peaks (Figure 3.5b inset). This means that although Figure 3.5a is cleaner and 100 times more intense than Figure 3.5b, both spectra yield valuable qualitative and quantitative

Table 3.3 The 10 discarded peptides showing abundance, time bin, and reason for discarding. Peptides in white were too ambiguous, while peptides in gray had too much noise.

Sequence	Abundance	Time Bin	Reason
TPALAAK	93.56%	5-10 min	Ambiguous
NGGVAGNTTVNQK	66.82%		Ambiguous
LGQMGELVR	86.53%	20-25 min	Ambiguous
VRELTVQATTGTNSESDLSSIQDEIK	31.04%	30-35 min	Low Signal:Noise
VALYGIDYLMK	92.50%	35-40 min	Ambiguous
DIQLATPPQVGAPATEYAALAELK	45.87%	40-45 min	Low Signal:Noise
ALLNSMVIGVTEGFTK	84.54%	50-55 min	Low Signal:Noise
QSIAVSALSLANQSQQGVLQLLR	66.18%		Low Signal:Noise
VLLPVPFALINDPFGK	88.72%	55-60 min	Low Signal:Noise
LMEQITTSDELIDFLTLPGYR	65.03%		Low Signal:Noise

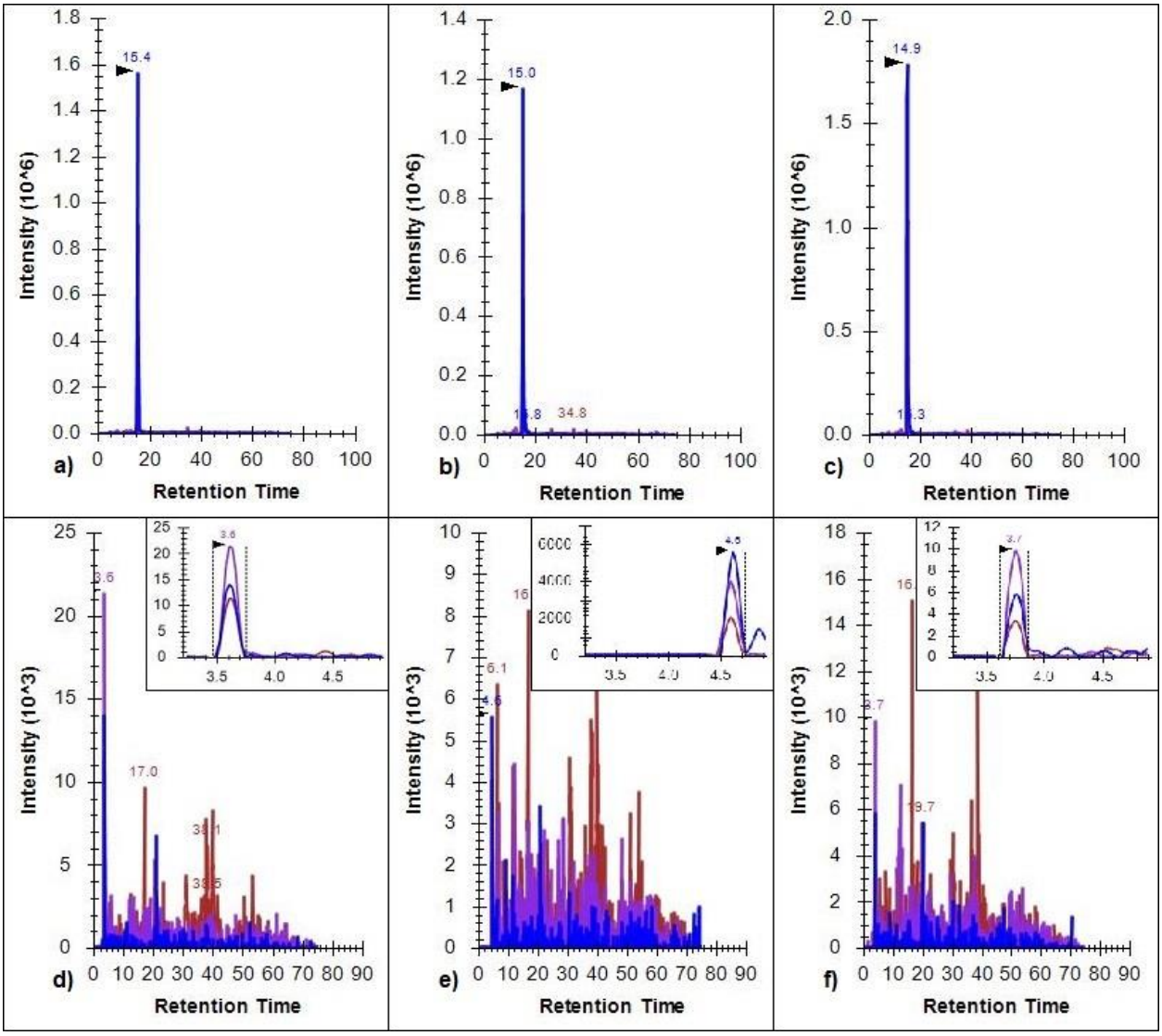


Figure 3.5 Excerpts of final MRM results. As before different colored lines indicate different transitions. a-c) Represent the strongest of the final MRM spectra and show three replicates of the same transitions for TVSENEVPLYK having high intensity, coinciding peaks, and consistent retention times. d-f) Represent the weakest of the final MRM spectra and show three replicates of the same transitions for VMQAQGSQLTNK, but insets show them as still having adequate intensity, the only coinciding peaks in their respective spectra, and consistent retention time.

information, and thus both are viable spectra. All 21 peptides produced viable spectra with their selected transitions, regardless of their peptide abundance or chromatographic congestion. This evidence supports conclusion that even in a significantly complex sample, low abundance and high congestion do not prohibit reliable MRM measurement in 60 minutes.

However, this conclusion and the results on which it is based, require that the congestions and abundances, as established by global experiments, be similarly maintained in targeted experiments. Although both global and targeted measurements used the same LC flow rate, chromatographic issues could stem from the LTQ having used a split-flow microflow pump while the QQQ employed a split-less nanoflow pump. Similarly, the uniform use of quadrupoles and CID fragmentation could not ensure abundance continuity since the LTQ and QQQ differ in mass analyzer configuration, collision gas, and other intrinsic properties. Yet this inconsistency is less of a dilemma, and more of a showcase for MRM software, as Skyline's comprehensive data report provides evidence for congestion and abundance being preserved across MS/MS platforms.

A portion of data from Skyline's report was used to generate Table 3.4, which demonstrates the sustainment of congestion across MS platforms. This table shows that although only 6 peptides eluted in their original bins, the order of elution was maintained for 18 of the 21 peptides. It also shows that peptides originating from the same bin remained close to or fewer than five minutes apart during MRM, with the exception of only the last bin. Finally, this table displays that all retention time standard deviations remained well below one minute across all three replicates.

As support for the preservation of peptide abundance, another selection of Skyline's report is employed by Table 3.5. Acknowledging that bin location was linked to both determining abundance and selecting peptides by it, this table groups peptides accordingly and then lists from lowest abundance to highest. In each bin, peptide order by LTQ abundance is reflected in QQQ peak area, again with the exception of the last bin. Table 3.5 also provides direct support for the above conclusion that MRM quality in a complex sample is not measurably influenced by abundance or congestion.

Table 3.4 The remaining 21 peptides comparing global MS/MS time to MRM time, and predicted order to detected order. Blue values are more favorable, green values are favorable, yellow values are unfavorable

		MRM Retention Time (min)			Order	
PeptideSequence	Global Bin	μ	σ	Range	Pedicted	Detected
AQASTHGIGK	0-5	15.24	0.20	N/A	1st	9th
LVEGSAQVK	5-10	8.42	0.23	N/A	2nd	7th
VMQAQGSQLTNK	10-15	3.99	0.55	1.8	3rd, 4th, 5th	1st, 2nd, 3rd
KVVVEYPK		5.05	0.69			
TATEYGVVR		5.81	0.51			
VDDGGTLDVR	15-20	6.79	0.31	1.4	6th, 7th, 8th	4th, 5th, 6th
LGSHNDMTFGEGTSSR		7.97	0.23			
STCTGVEMFRK		8.19	0.26			
TVSENEVPLYK	20-25	15.06	0.23	5.5	9th, 10th	8th
EFNVEANVGKPQVAYR		20.53	0.51			10th
GPASVTNEQIEQVVR	25-30	23.07	0.47	6.4	11th, 12th, 13th	11th
IEIPGCSLCMGNQAR		28.79	0.61			13th, 14th
FDGNACVLLNNNSEQPIGTR		29.49	0.64			
LAATIAQLPDQIGAK	30-35	32.22	0.65	3.4	14th, 15th	15th
GIVDSNLGLSPATEGQVIR		35.63	0.61			16th
AVIFAGELLK	35-40	39.27	0.48	4.1	16th, 17th	17th
VIDLMCPFAK		43.37	0.66			18th
ANQVPQQVLSLLQG	40-45	52.56	0.60	2.0	18th, 19th	19th, 20th
VPDIGADEVEITEILVK		54.52	0.34			
GITLPETELR	45-50	26.54	0.54	29.4	20th, 21st	12th
VGAGPFPTELFDETGEFLCK		55.92	0.15			21st

Table 3.5 The 21 remaining peptides showing factors of possible influence and associated MRM performance. Blue values are more favorable, green values are favorable, yellow values are neutrally acceptable, orange values are unfavorable, red values are more unfavorable

	Global			MRM Peak Area	
PeptideSequence	Bin	Abundance	Length	μ	Cv
AQASTHGIGK	0-5	68.43%	10	4.9E+06	0.07
LVEGSAQVK	5-10	35.23%	9	8.3E+05	0.73
VMQAQGSQLTNK	10-15	28.79%	12	2.0E+05	0.73
KVVVEYPK		65.19%	8	9.8E+05	0.09
TATEYGVVR		93.80%	9	4.6E+06	0.87
LGSHNDMTFGEGTSSR	15-20	25.20%	16	9.2E+05	0.60
VDDGGTLDVR		60.72%	10	5.4E+06	0.27
STCTGVEMFRK		78.87%	11	6.9E+06	0.43
EFNVEANVGKPQVAYR	20-25	25.32%	16	2.7E+06	0.55
TVSENEVPLYK		62.27%	11	4.6E+07	0.26
FDGNACVLLNNNSEQPIGTR	25-30	28.36%	20	1.5E+06	0.22
GPASVTNEQIEQVVR		62.37%	15	1.0E+07	0.34
IEIPGCSLCMGNQAR		91.99%	15	1.1E+07	0.17
GIVDSNLGLSPATEGQVIR	30-35	61.73%	19	2.7E+06	0.02
LAATIAQLPDQIGAK		88.78%	15	1.3E+07	0.07
AVIFAGELLK	35-40	37.94%	10	1.5E+06	0.33
VIDLMCPFAK		62.64%	10	8.2E+06	0.19
VPDIGADEVEITEILVK	40-45	61.45%	17	1.2E+06	0.24
ANQVPQQVLSLLQG		96.89%	14	1.2E+07	0.16
GITLPETELR	45-50	60.57%	10	5.4E+06	0.29
VGAGPFPTELFDETGEFLCK		76.35%	20	6.2E+05	0.10

Comparing global abundances to MRM peak areas yields no clear trend, and the tightest cluster of large peak areas is generated by the most congested part of the chromatogram.

Aside from experimentally distancing complex sample MRM from the influence of the predicted factors, this project revealed one influential factor that was not predicted during its inception. Referring to Table 3.6 as an illustration, it was noticed that while peptides of 10 amino acids or longer produced strong results with random exceptions, peptides that were less than 10 amino acids consistently displayed less reproducibility if any at all. Also referring back to Table 3.5, it would appear that lengths above 20 amino acids are questionable targets. This information will prove to be relevant in chapter 4.

3.4 Summary

In this chapter a workflow was developed to select robust peptides for evaluating the reproducibility of an MRM in a complex sample, containing more than 80,000 peptides, with a 60 min 1D-LC-MS/MS. A discussion of MRM's abeyance of peptide congestion and abundance in small samples, as well as MRM's typical target size, explained why the workflow was also guided to pick approximately 2 dozen peptides for the evaluation of each factor's influence on MRM quality in a complex sample. Following this discussion, the mechanics of the workflow were laid out to validate the characterization of each factor's influence. This discourse began by detailing how a pool of 3827 peptides was generated by global MS/MS, and then filtered down to 415 candidates in order to define areas of differing chromatographic congestion and characterize congestion effects on MRM reproducibility. The next section was dedicated to explaining how the selection 31 peptides provided a range of abundances on which to profile the influence of abundance on MRM quality. Preliminary MRMs of the 31 peptides were then presented and reviewed to demonstrate how the selected peptides were themselves evaluated for suitability in testing both MRM quality and factor influence; further explanation was offered as to why 10 peptides were rejected, leaving 21 for final MRM evaluations. After depicting the workflow, the final MRM results were explored to reveal that all 21 peptides generated

Table 3.6 MRM performances of peptides less than 10 amino acids long compared to peptides 10 amino acids long. Blue values are more favorable, green values are favorable, yellow values are neutrally acceptable, orange values are unfavorable, red values are more unfavorable

	Global			MRM Peak Area	
PeptideSequence	Bin	Abundance	Length	μ	Cv
TPALAAK	5-10	93.56%	7	Discard	-
KVVVEYPK	10-15	65.19%	8	9.8E+05	0.09
LVEGSAQVK	5-10	35.23%	9	8.3E+05	0.73
TATEYGVVR	10-15	93.80%	9	4.6E+06	0.87
LGQMGELVR	20-25	86.53%	9	Discard	-
AQASTHGIGK	00-05	68.43%	10	4.9E+06	0.07
VDDGGTLDVR	15-20	60.72%	10	5.4E+06	0.27
AVIFAGELLK	35-40	37.94%	10	1.5E+06	0.33
VIDLMCPFAK	35-40	62.64%	10	8.2E+06	0.19
GITLPETELR	45-50	60.57%	10	5.4E+06	0.29

reproducible spectra, which supported the verdict that MRM in a complex sample is robust and minimally affected by congestion and abundance. Finally, excerpts from the MRM data report were inspected to confirm the preservation of each factor, bolster the final verdict, and also identify peptide length as unpredicted factor with a noticeable influence on MRM quality.

Chapter 4: Designing and Demonstrating Possible Experimental MRM-MS Approaches for Characterizing Specific Metabolic Pathways of a Controlled Bacterial Mixture

4.1 Experimental Design for MRM of a Metabolic Pathway

This chapter contains the second project of this thesis, which was developed to profile a metabolic pathway in three organisms using MRM designed for complex samples. It will begin with a quick discussion on the merits of selecting the TCA cycle for this study, the origin for the comparison of the *ab initio* method to the empirical method, and the application of conclusions from the previous chapter. Following this discussion, will be descriptions on how MRM is employed to characterize a metabolic pathway, precautions for designing such an MRM, and common criteria for both design methods. Then, fundamental differences between the *ab initio* and empirical will be laid out before beginning three comprehensive comparisons. The first will differentiate how each method generates possibilities for building MRM transition lists, and it will contrast the numbers generated by each. The second comparison will explain the uneven protein selection as well as focus on how *ab initio* and empirical differed in ranking precursors and fragments for peptide transition selection. The last comparison will assess the MRM performance of each method, and provide a conclusion for the ability of MRM to profile a pathway in a complex sample. This section will also include a discussion of how the total numbers to tie back to availability, and how the results for shared proteins point to improvements for the *ab initio* method.

4.2 Selection of a Biological Pathway

The TCA cycle was selected for this study based partially on its large number of enzyme driven reactions and its universal availability, but it was chosen primarily for its vital role in cell life. As seen in the lower part of Figure 4.1, it is a 10 to 12 step process, and three organisms in the 4iso sample have well documented TCA cycles, *E. coli*, *R. pal*, and *I .hos*; as *N. equi* relies on the TCA cycle of *I. hos*, it contains no cycle of its own[16]. As for the criticality of this metabolic pathway to sustaining cellular life, the

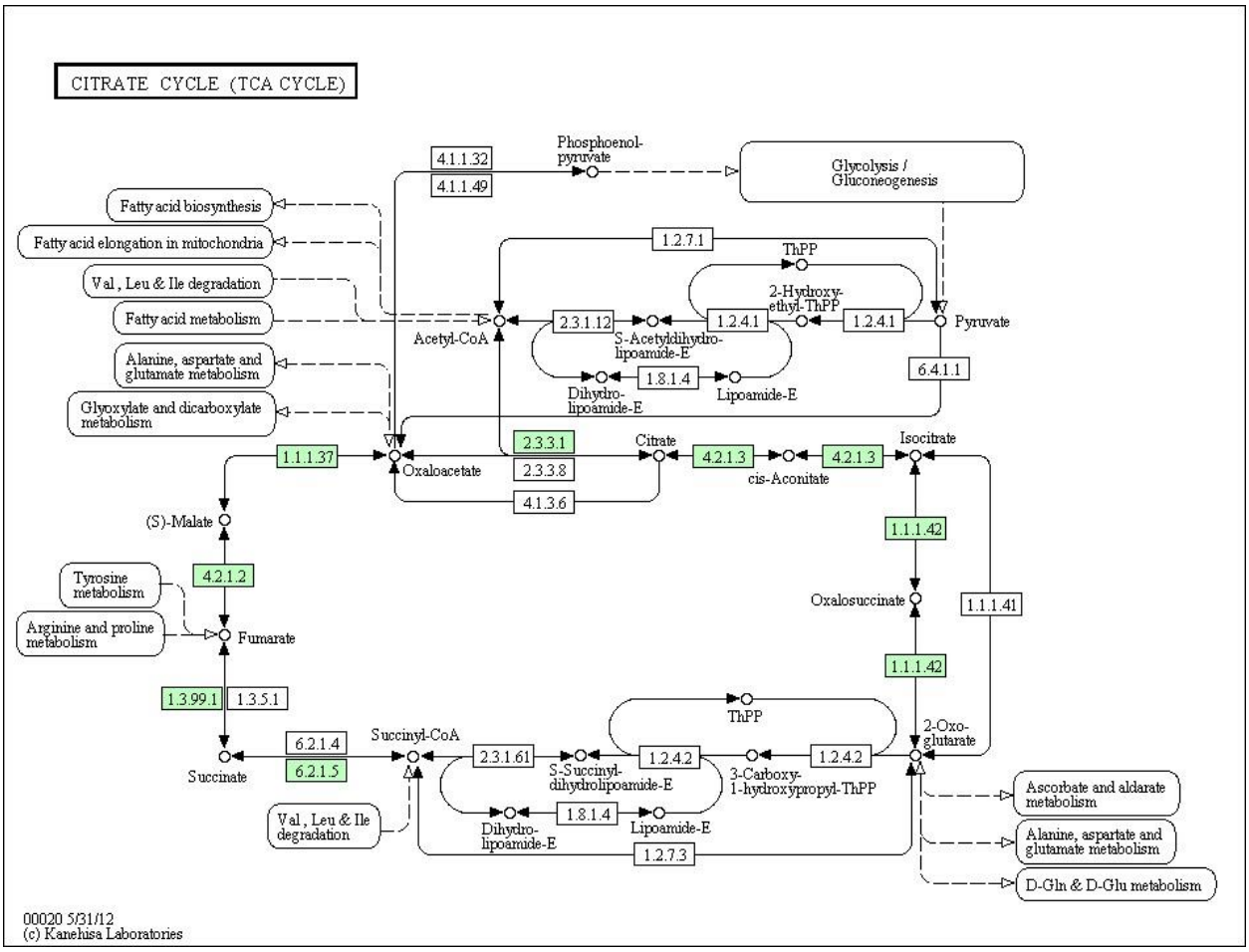


Figure 4.1 KEGG Cycle for TCA. The nine steps are highlighted to show the enzyme commission numbers that were represented in all three organisms, and to provide an understanding for the encompassing nature of the characterization of this cycle[28]

TCA cycle possesses a multifaceted role which is further increased by an ability to drive its reactions in reverse order. The reverse TCA cycle is utilized by some bacteria, such as *I. hos*, for carbon fixation and biosynthesis of amino acid precursors [16, 29], while the forward process is central to the catabolism of carbohydrates, fats, and proteins for all aerobic organisms. In the latter instance, the TCA cycle oxidizes the acetyl CoA obtained from these biomolecules and generates adenosine triphosphate (ATP), which is the primary form of intracellular chemical energy. For every 1 molecule of acetyl CoA, TCA can generate 1 molecule of ATP directly and another 10 ATP molecules indirectly through incident coenzymes [30]. This dexterous indispensability of the TCA cycle will provide considerable gravity to demonstrating MRM profiling of a metabolic pathway.

However, the conventionally employed empirical approach to MRM design proved inadequate, as the preliminary data did not identify enough proteins to characterize the cycle for *I. hos*. This represents a significant oversight in that *I. hos* provides an uncommon profiling opportunity by possessing a uniquely augmentation to facilitate a reversed TCA cycle [16]. The desire to characterize this rare pathway necessitated the utilization of an *ab initio* method to guide the selection of proteins and peptides that were missed by empirical means. Consequently, this project was enriched with the secondary objective of comparing these two methods on availability and selection of proteins and peptides, and on the performance of their respective MRMs. These two design approaches also link the current project to the previous one by employing conclusions from chapter 3; the *ab initio* method would avoid inclusion of peptides less than 10 amino acids long, while the empirical method would use a redefined measure of peptide abundance.

MRM characterizes a metabolic pathway by profiling the enzymatic proteins that catalyze its major reactions; targeting a few unique proteins across the beginning, middle, and end of a pathway can suffice for characterization. But as larger pathways are considered, more proteins are needed; and naturally, studying more proteins for any given pathway provides a more comprehensive understanding of that pathway as a whole. Along with identifying these proteins by their peptides, MRM can profiles the

proteins with the spectra generated by the peptide's transitions. Under certain circumstances, detecting one peptide precursor is sufficient for protein identity and profiling, but identities and profiles are not independently reliable unless based on at least two peptides with a combined three precursors. Thus an entire pathway should not be characterized solely on proteins with single-peptide profiles, and even doing so on dual-peptide profiles could be questionable. These profiles, and their underlying identities, are more reliable when corroborated by an independent counterpart.

Designing an MRM to characterize a biological pathway involves careful evaluations of proteins, peptides, and transitions in order to make good selections, but designing that MRM to be done in 60 min, without fractionation, demands accurate knowledge on the best selections at each level. It is important to remember that even the most carefully selected peptides may evade detection, and some precursors may not be generated in the ion source. Hence there is no certainty that a selected protein will be identified, and proteins with just one peptide precursor are considered unreliable for profiling. But each additional peptide renders its protein more reliable, and each reliable protein provides more evidence for a metabolic pathway. So naturally, designing a quality MRM for pathway characterization in a complex sample commands substantial availabilities of proteins, peptides, and transitions.

4.3 Two Possible MRM Approaches

The *ab initio* and empirical methods shared three criteria for designing MRM. First, TCA proteins were considered for profiling only if their enzyme commission number (ECN) was represented in all the three organisms (highlighted in Figure 4.1) [16, 28]. This provided continuity on which to directly compare MRM performance in each organism, and it naturally defined which proteins would be considered. One exception was made for *I. hos* to include the four additional proteins that were critical to its augmented cycle [16]. The final result included 7 ECNs in all three organisms, plus an additional 4 for *I. hos*, for a total of 46 proteins among them. The second criterion was the exclusion of peptides that contained missed cleavages, as this is an irregular occurrence. Finally, the precursor ions for each peptide had to have *m/z*s less than

1500 to stay within the QQQ's mass range. The numbers available to each method, for each organism, and at each level, are presented in Table 4.1; here gray numbers signify a one to one relationship with the subordinate level. As stated before, MRM peptide attrition rates rendered proteins with only one peptide as too unreliable for profiling. And if a given TCA cycle had less than three reliable proteins with each representing a different ECN, or if there were no independent proteins in the entire cycle, then an MRM would not be designed for that cycle.

From this point on, fundamental differences between the *ab initio* approach and the empirical approach would lay out separate paths for MRM design. The *ab initio* approach is relatively new to MRM, and true to its namesake[31], it relies wholly upon calculated models to score peptides, precursors, and fragments for transition selections. This enables the *ab initio* method to consider all possibilities, but requires the application of logical restrictions to prevent the unnecessary use of time and resource in evaluating unlikely candidates. Also, as there are no preliminary MS/MS experiments, this method carries an inherently higher degree of uncertainty. The empirical method is the more traditional approach, and as stated before, takes guidance from preliminary measurements. Although this provides proof that the target protein or peptide is available, it is not a guarantee. Also, this method is limited to what was discovered in the preliminary experiments, as any other options would lack the means to be scored.

The first comparison of MRM design by *ab initio* and empirical will focus on what each method provided for selection, and will begin with empirical method as its selection gave rise to this comparison study. The possibilities for empirical based MRM comprised all proteins, peptides, precursors, and fragments that were identified in the global MS/MS; this allowed any peptide length, and any precursor or fragment charge state. The global data provided the empirical method with 7 *E. coli* proteins, and 5 *R. pal* proteins, with all 12 proteins representing a different ECN. All proteins had at least two precursors, and only one protein from each organism had only two precursors. As for *I. hos*, the data provided only one protein that bore only two peptides, and regardless of its lack of independence, this was clearly insufficient to support an empirical MRM of *I. hos*, as was previously stated.

Table 4.1 Number of available selections at each level. Gray numbers signify a one to one relationship with the subordinate level.

		Possible Reaction	Protein	Peptide	Precursor
Empirical	e.coli	7	7	41	54
	r.pal	5	5	21	23
	i.hos	1	1	2	2
	Total	13	13	64	79
Ab Initio	e.coli	7	19	222	222
	r.pal	7	13	146	146
	i.hos	11	14	141	141
	Total	25	46	509	509

The *ab initio* based MRM could consider any protein meeting the shared criteria, and as this method was not limited to experimentally identified components, it could have considered all possibilities of peptides, precursors, and fragments. However, this could mean approximately 60 peptides for each of the 46 proteins; and each peptide would average 6 precursors and 9 fragments, which translates into 378 transitions per peptide. Aside from snowballing into an avalanche of options that would bury both Skyline and its user in endless processing, most of these possibilities are not realistically feasible under normal operating conditions; thus the majority of them were weeded out by restricting the prediction parameters to only include the properties that were most commonly observed. Specifically, peptide predictions were kept between 10 and 25 amino acids long, and transition predictions were limited to a +2 precursor charge with a +1 fragment charge. Despite these restrictions, there were still two or more peptides available to each of the 46 proteins, and any given TCA cycle could be characterized solely by proteins that had 9 or more peptides available. This meant that every protein was available its cycle, and each cycle had the potential for all of its proteins to be independently profiled. Comparing the total numbers in Table 4.1 clearly demonstrates that the *ab initio* method has more options at each level.

Table 4.2 displays direct comparisons of the availabilities for each method across the shared organisms; but more importantly, it provides an understanding of the amount that *ab initio* missed at the peptide and precursor level on account of its restrictions. Comparing the shared numbers to the empirical numbers shows that *ab initio* predictions failed to include a total of 7 peptides and 28 precursors that were found in the global data. The lost peptides were filtered out by their lengths, which consequently filtered out the 10 precursors found among them. However, 7 of those 10 would have been eventually filtered out by their charge; such was the root of *ab initio* exclusion for the remaining 18 precursors, as these belonged to shared peptides. Notably, 6 of these 18 were the only precursors to be empirically available to their respective peptides; this explains why the number of shared peptides is 6 more than the number of shared precursors. The importance of identifying the number and cause of these *ab initio* exclusions will be covered in the following discussions of which peptide precursors were

Table 4.2 Direct comparison by shared organism of number of available selections at each level. Gray highlight emphasizes discrepancy between number of shared peptides and number of shared precursors.

		Possible Reaction	Protein	Peptide	Precursor
e.coli	Ab Initio	7	19	222	222
	Empirical	7	7	41	54
	Shared	7	7	35	29
r.pal	Ab Initio	7	13	146	146
	Empirical	5	5	21	23
	Shared	5	5	20	20
Both	Ab Initio	14	32	368	368
	Empirical	12	12	62	77
	Shared	12	12	55	49

selected and detected.

4.4 The MRM Selection Process

With the possibilities defined, the scoring and selection process could begin. This brings the second comparison of *ab initio* to empirical, which focuses on how each one scored and selected peptide transitions at both the precursor and fragment levels; but there is no comparison of protein level selections for three reasons. First, both methods scored proteins with the average score of their top three precursors; ergo a comparison of protein selections is merely a comparison of the underlying precursor selections. Next, the empirical method provided no alternative proteins for any given ECN; thus a protein level comparison would be made not between different methods of selection, but between *ab initio* selection and empirical availability. Finally, the general lack of proteins by empirical method was what led to the inclusion of the *ab initio* method, and as the latter provided the only means for complete characterization in all three organisms, it was decided to guide protein selections strictly primarily by their merits for characterizing the *ab initio* method.

The protein selection numbers are summarized in Table 4.3 and show that *ab initio*, selected 25 proteins; 15 were by default as no alternatives were available, and 10 were by top score among alternatives. To test the effect of selecting a non-default *ab initio* protein with an inferior probability average, the highest averaging protein was passed for the next highest alternative in its ECN; this was *E. coli*'s p0115[28], and was substituted with p0744. Despite having the inferior average in its own ECN, p0744 had a higher average than 19 of the 24 proteins from the other ECNs. Checking for which proteins were shared revealed that all 5 *R. pal* in empirical had been selected by *ab initio*, and could therefore be used for comparison; this was a hardly coincidence though, as four of these proteins were defaults. *Ab initio* also selected 5 of the 7 *E. coli* proteins in empirical, with three by default and two by score. As for empirical's two *E. coli* proteins that were not selected, one was the above mentioned p0115. The other protein was p0703, which was ranked by *ab initio* as second to p4043 for their ECN. The impact of which proteins were selected and which were skipped will be addressed

Table 4.3 Number selected at each level

		Selected Protein	Precursor	Fragment
Empirical	e.coli	5	14	56
	r.pal	5	13	52
	Total	10	27	108
Ab Initio	e.coli	7	21	84
	r.pal	7	21	84
	i.hos	11	32	128
	Total	25	74	296

in the discussions on what was detected.

It could be argued that it would have been better to use p0703 to test inferiority in *ab initio* ranks, which would allow p0115 to be selected, and thus all 7 empirical *E. coli* proteins would be used. However, selecting these peptides presents two drawbacks. First, the presence of p0115 is already known and so its profile would only characterize selections of peptides and transitions; this in turn would mask the impact of *ab initio* scores on protein selections. Second, using all 7 *E. coli* proteins for the comparison would further disrupt empirical continuity between *E. coli* and *R. pal*.

With the proteins chosen, the selection process then moved to peptide and transition levels, but since *ab initio* and empirical shared more peptides than precursors, comparisons are better explained by currently referring to these levels as precursor and fragment selections. Precursor selection required two to three precursors, representing at least two different peptides in empirical's case, based on availability and score. This selection size maintained protein reliability and independence while minimizing MRM duty cycle and the precursor selection numbers are given in Table 4.3. The empirical method scored each precursor by its MII and had multiple precursors available to many of its peptides, but it only selected the +2 precursor for all but one of its peptides; for this peptide it selected the +2 and +3 precursors. For two of its proteins, empirical could only select two precursors, and it selected none of the 10 precursors that were excluded by *ab initio* restrictions; rendering these exclusions inconsequential. As for *Ab initio* selection, a precursor's calculated probability was used for its score, and only one precursor was available per peptide, but *ab initio* selected three precursors for every protein to have the maximum number of precursors for each organism. It should be noted that a secondary check of peptide suitability, following this project's data analysis, resulted in the ex post facto removal of one *I. hos* precursor from the *ab initio* selection, and one *R. pal* precursor from the empirical selection. These omissions are reflected in all tables, but affected neither *ab initio* evaluation nor method comparison.

The final step in the selection process was selecting the top four fragments for each precursor, where only fragments with *m/z*s less than 1500 were considered, and the selection size was again chosen to maintain confidence in precursor identity while

minimizing MRM duty cycle. Empirical selection was guided by a visual inspection of each precursor's fragment spectra, and the four fragments with most intense peaks were selected. Naturally, the *ab initio* method offered no such spectra, but rather it guided fragment selection with calculated scores similar to those it used for peptide selection. Again, the total selection numbers are given in Table 4.3, and they show that each method selected the maximum number of fragments for their precursors.

Table 4.4 focuses on shared *E. coli* and *R. pal* proteins to offer a direct comparison of the *ab initio* method to the empirical. Although both methods effectively filled the ranks, the ratios in this table make it is easier to see that each did so with mostly distinct selections. The widest gaps exist at the fragment level, yet stem from differences in precursor selections, more than fragment selections. Table 4.5 better illustrates this point by drawing from both methods only the precursor numbers from shared proteins, and only the fragment numbers from shared precursors. Looking at both organisms together, a total 48 precursors were selected from the 10 shared proteins, but only 9 precursors were shared among both methods, and they came from only 6 proteins. As a comparison, a total of 46 fragments were selected from those 9 precursors, and 26 of those were shared and came from all 9 precursors. Hence, fragment selection was more similar between the two methods than precursor selection.

4.5 Experimental Evaluation of the Two MRM Approaches

After the selections were made and the transition lists built, the MRM's designed by the *ab initio* and empirical methods were executed on the QQQ. The results were analyzed with Skyline, where any transitions with weak or irregular signals were discarded, and any precursor of less than three viable transitions was also discarded. But as mentioned before, single precursor detections were sufficient for identification and profiling, provided that an independent protein was present in the same method-organism scenario. Table 4.6 lists the targeted proteins by their ECNs and its numbers affirm the successful MRM profiling of the TCA cycle in a complex sample with all five scenarios providing no less than 4 profiled proteins, and at least one of which was independent. As neither method consistently produced better MRM peak intensities,

Table 4.4 Direct comparison by shared protein of number selected at each level, also showing number shared for each organism over total for each organism

		Selected Protein	Precursor	Fragment
e.coli	Ab Initio	5	15	60
	Empirical	5	14	56
	Sha/Tot	5/5	4/25	10/106
r.pal	Ab Initio	5	15	60
	Empirical	5	13	52
	Sha/Tot	5/5	5/23	16/96
Both	Ab Initio	10	30	120
	Empirical	10	27	108
	Sha/Tot	10/10	9/48	26/202

Table 4.5 Number of shared precursors per protein, and number of shared fragments per precursor by organism

	Unit:	Precursor	Fragment
	Base:	Protein	Precursor
e.coli	Shared Bases	5	4
	Bases with Shared Units	3	4
	Units from Shared Bases	25	22
	Shared Units	4	10
r.pal	Shared Bases	5	5
	Bases with Shared Units	3	5
	Units from Shared Bases	23	24
	Shared Units	5	16
Both	Shared Bases	10	9
	Bases with Shared Units	6	9
	Units from Shared Bases	48	46
	Shared Units	9	26

Table 4.6 Targeted proteins as identified by their ECNs and showing number of precursors detected per protein by organism and method. N/A is no protein available for that ECN; N/E is *ab initio* selected protein not available to empirical. Blue highlighting shows *ab initio* exclusive protein profiles.

Enzyme	e.coli		r.pal		i.hos	
Commision	*Ab initio*	Empirical	*Ab initio*	Empirical	*Ab initio*	Em
2.3.3.1	1	3	3	3	2	N/A
4.2.1.3	0	N/E	1	2	1	
1.1.1.42	3	2	1	2	1	N/E
6.2.1.5	3	3	2	2	2	
1.3.99.1	1	N/E	1	N/A	3	
4.2.1.2	2	0	1	N/A	0	
1.1.1.37	2	2	2	3	3	N/A
1.2.1.11	reactions unique to i.hos				2	
6.2.1					1	
4.2.1.120					2	
4.2.1.55					2	

the final comparison of the *ab initio* and empirical was left to the numbers of detected proteins, precursors, and fragments as shown in Tables 4.6 and 4.7. The general impression from both tables is that *ab initio* and empirical performances were comparable, with both methods missing one protein each in *E. coli*, profiling all of their target *R. pal* proteins, and detecting similar numbers of peptides in each organism. However, there are some obvious differences and some subtle differences, which together point to important conclusions, all of which are discussed below.

The numbers in Table 4.7 reveal that *ab initio* generated 14 more protein profiles than empirical, and more than 71% of the total protein profiles. The highlighting in Table 4.6 shows that these additional profiles arose from 10 *I. hos* proteins, 2 *R. pal* proteins, and 1 *E. coli* protein that were not discovered by empirical's global MS/MS. The remaining additional profile arose from *ab initio* selecting two of its exclusive peptides to detect a protein that empirical selections missed. These detection statistics point back to the beginning of this project and affirm that *ab initio* is stronger than empirical in regards to availability of proteins and peptides. The advantage of this strength is exemplified by *ab initio*'s exclusive profiling of the *I. hos* reverse TCA cycle, and is further bolstered by full inclusion of the critical augmentation to this cycle as highlighted in Figure 4.2.

Connecting to another earlier discussion, the *E. coli* protein that was missed by *ab initio* was p0744, which was selected over p0115 to the test a second rank selection. This loss was despite p0744 having a higher average than many of the first ranked proteins, and it is made even more unique by being the only non-default protein to be missed by either method. Although it does not conclusively measure the effect of *ab initio* score order on MRM selection, this loss does invite further investigation on the topic.

Another interesting aspect about p0744 is that it was one of only two proteins that were not available to empirical while representing ECNs that were; the other was p0703, which was ranked first place by *ab initio* and detected by MRM. This means p0744 was the only targeted protein that was deemed inferior by both methods, and it was the only portion of the *E. coli* TCA cycle to be missed by both methods. The latter

Table 4.7 Number of detected at each level

		Detected Protein	Precursor	Fragment
Empirical	e.coli	4	10	35
	r.pal	5	12	44
	Total	9	22	79
Ab Initio	e.coli	6	12	45
	r.pal	7	11	38
	i.hos	10	19	76
	Total	23	42	159

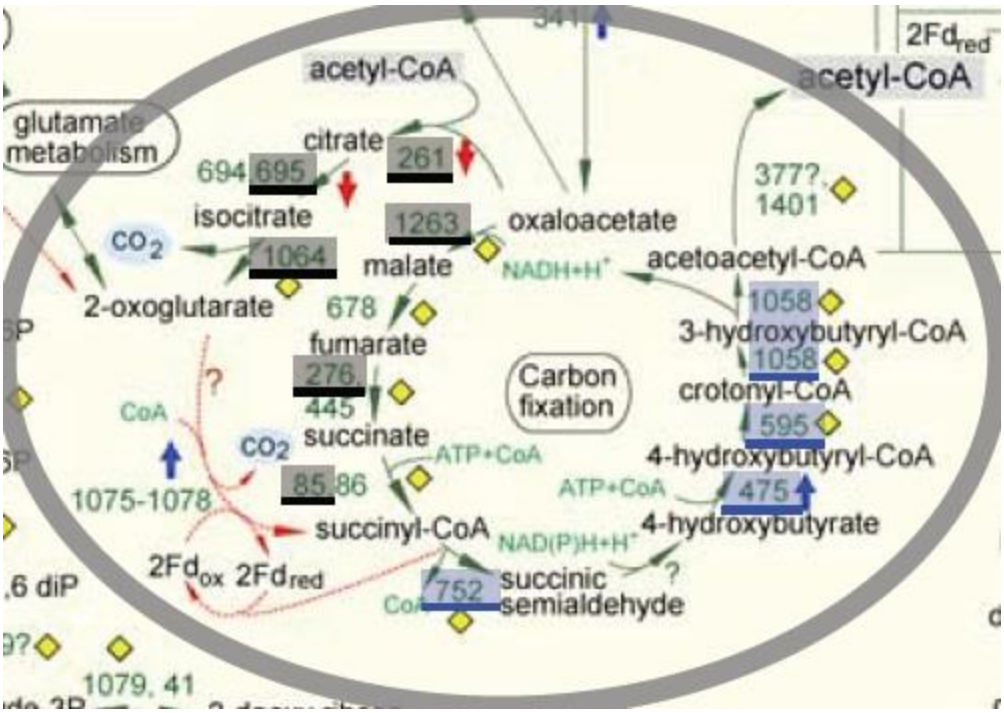


Figure 4.2 Excerpt from the updated reconstruction of I. hos metabolism showing the augmented reverse TCA cycle for carbon fixation [16]. The numbers that are highlighted and underlined in black identify the I. hos proteins for the standard TCA cycle that were successfully profiled. The numbers with blue emphasis identify the successfully profiled proteins of the I. hos exclusive augmentation, and they demonstrate the full inclusion of this critical addition by MRM pathway profiling.

statement points to the proteins of both organisms being well profiled across *ab initio* and empirical MRMs together, and is demonstrated in Table 4.8 by comparing the numbers of either method to the combined numbers. It may seem like an obvious outcome that the methods would perform better together than separately, but what is less pronounced is that a similar performance could be achieved by just one method, the *ab initio* method.

Considering each method's numbers for only the shared ECNs, as seen in Table 4.9, empirical has virtually exhausted its protein options, and is much closer to its limit of possible precursors than *ab initio*, hence there is little room for improvement for the former. Conversely, Table 4.9 demonstrates a wealth of options available to *ab initio*, and aids in method improvement by eliminating protein selection from consideration. Recalling from Table 4.5, the methods differ less in fragment selection, which focuses improvement measures on increasing peptide precursor selection. Looking at the detected precursor numbers from Table 4.9 reveals that 13 detected precursors were selected by the empirical method. One of these was the aforementioned +3 selection, which was excluded from *ab initio* availability. But as was also mentioned, each remaining precursor was +2 and came from a separate peptide, none of which were excluded by *ab initio*. Thus the 12 peptides that were selected only by empirical were still available to *ab initio* selection, 5 for *E. coli* and 7 for *R. pal*. However, increasing the *ab initio* selection number by just one peptide can bring a significant increase to transition list size, as this addition is multiplied across 7 proteins for each organism and further multiplied by 4 transitions for each peptide. Thus a compromise must be met between decreasing duty cycle and improving *ab initio* MRM quality. Table 4.10 displays the balance between assimilating empirical selections and overloading transition list. From this table, it can be determined that selecting the top 5 or top 6 peptides for *ab initio* would result in an appreciable increase of MRM quality, without sacrificing significant duty cycle. Thus the *ab initio* method can single-handedly provide a more complete TCA profile.

Table 4.8 Direct comparison by shared protein of number detected at each level

		Detected Protein	Precursor	Fragment
e.coli	Ab Initio	5	11	42
	Empirical	4	10	35
	Combined	5	17	67
r.pal	Ab Initio	5	9	30
	Empirical	5	12	44
	Combined	5	16	59
Both	Ab Initio	10	20	72
	Empirical	9	22	79
	Combined	10	33	126

Table 4.9 Direct comparison by shared ECNs of numbers available, selected, and detected at each level

		Protein			Pecursor			Fragment	
		Possible	Selected	Detected	Possible	Selected	Detected	Selected	Detected
e.coli	Ab Initio	19	5	5	222	15	11	60	42
	Empirical	7	5	4	54	14	10	56	35
	Shared	7	5	4	29	4	4	10	10
r.pal	Ab Initio	6	5	5	84	15	9	84	30
	Empirical	5	5	5	23	13	12	52	44
	Shared	5	5	5	20	5	5	16	15
Both	Ab Initio	25	10	10	306	30	20	120	72
	Empirical	12	10	9	77	27	22	108	79
	Shared	12	10	9	49	9	9	26	25

Table 4.10 Comparison of top n selections showing the number of empirical selections assimilated, number of total peptides, and number of total transitions. Dark blue values are most desirable, light blue values are more desirable, purple values are desirable, light red values are less desirable, and dark red values are least desirable.

	Pick	Peptides		Transitions
		Emprical	Total	
e.coli	top 3	0	21	84
	top 4	0	28	112
	top 5	2	35	140
	top 6	2	42	168
	top 9	3	63	252
	top 12	5	77	308
r.pal	top 3	0	21	84
	top 4	1	28	112
	top 5	3	35	140
	top 6	5	42	168
	top 9	6	63	252
	top 12	7	83	332

4.6 Summary

The chapter presented the project of applying MRMs designed for complex samples to profile the proteins of TCA cycles for *E. coli*, *R. pal*, and *I. hos*. As global MS/MS provided insufficient empirical data to design MRMs for all three organisms, this project introduces an *ab initio* approach to MRM design, which presented the opportunity to compare two competing methods. This project also took heed of the previous chapter's conclusion on peptide length by applying it to the *ab initio* method for designing MRM, as well as applying the conclusion on peptide abundance by changing its definition for the empirical method in this chapter. An explanation of how MRM profiles a metabolic pathway provided context for the common precautions and criteria observed by both *ab initio* and empirical, then a fundamental contrast set the stage for their three comprehensive comparisons. The two methods were first evaluated on what each provided in terms of proteins, peptides, and precursors. As the empirical method's deficiency was the origin for this comparison study, it is no surprise that *ab initio* had the clear advantage in numbers possible at each level; *ab initio* was also shown to provide both a better opportunity for profiling *R. pal*'s cycle and the only opportunity for *I. hos*. The methods were then assessed for diversity in selections of precursors and fragments, and it was demonstrated that they differed more in their ranking of precursors than that of fragments. The final comparison was based on the detection results of each methods design, which revealed that MRM profiling of metabolic pathways in a complex sample is achievable, and the methods are essentially even in selecting viable peptides and transitions for shared proteins. But, *ab initio*'s strength in availability was proven to be of vital importance as it enabled this method to solely profile the TCA cycle of *I. hos*, and to assimilate some precursor selections that were originally unique to empirical, thereby granting *ab initio* the ability to increase its MRM performance.

Chapter 5: Discussion

The objective of this thesis was to expand the scope of the MRM approach in proteomics to include the reproducible analysis of complex samples without the assistance of extensive sample preparation. This goal was achieved by executing two projects which dually demonstrated an increased understanding of MRM capability, and provided depth for this understanding through the success of secondary objectives. The first project affirmed MRM reproducibility in a complex sample through the demonstration of robust peptide measurements, as evidenced by characterization of the influence of predicted and unpredicted factors on MRM quality. The second project illustrated the significance of complex sample MRM design by successfully profiling the same metabolic pathway in three microbes from the same complicated sample. This project further displayed the value of MRMs in a complicated mixture through a comparison of two evenly matched approaches for MRM design.

By meeting the objectives of both projects, the work presented in this thesis successfully challenged conventional perceptions of MRM limitations. Compared to the traditional MRM samples of blood, plasma, or urine, the 4iso sample presented a vast labyrinth with the complete proteomes of *E. coli, R. pal, I. hos, and N. equi*. Next, the TCA cycle was an ideal pathway choice because of its large number of enzymatic reactions and its presence in three of the organisms, but mostly because of its vital role in cell life. Finally, the conventional complex sample strategies of fractionation, 2D-LC, and 24 hour run-times were eliminated by the employment of 60 minute 1D-LC-MS/MS.

In the first thesis project (chapter 3), the generation of reproducible spectra for selected peptides from the 4iso sample validated both the MRM of complex samples, and the workflow developed to design and evaluate those MRMs. From global MS/MS results for a pool of over three thousand peptides, this workflow sufficiently defined regions of differing chromatographic congestion by identifying at least four hundred candidates that were consistently detected, and could be grouped into 5 minute bins. This workflow also provided an appropriate range in peptide abundance by defining three levels therein and identifying one peptide from each level for each bin. The explanation for removing weak peptides also provided characterization of peptide length

as an unpredicted third influential factor, while congestion and abundance were well characterized on the results of the 3 replicate MRMs of the remaining peptides. These characterizations defend the conclusions that congestion and abundance have little if any influence on MRM quality in a complex sample, while peptide length is a more likely source of loss.

The second related thesis project (chapter 4) built on the conclusions from the chapter 3 work to accomplish the goal of MRM characterization for the TCA cycles in a complex sample, while also providing a venue for the comparison of the *ab initio* and empirical methods for MRM design. Assessing each method on their availability numbers proved *ab initio*'s advantage in making more proteins accessible for attempted profiling, while examining diversity in selection revealed that the methods differed most in how each scored precursors. Although the detection numbers revealed that *ab initio* and empirical were relatively equal in performance on shared proteins, they also clearly demonstrated that *ab initio*'s strength in availability provided more total profiles and greater opportunity for improvement.

Through the demonstration of both MRMs capabilities in complex samples, and *ab initio*'s comparable performance in MRM design, this thesis has authenticated a method for metabolic profiling that requires minimal sample preparation and preliminary measurements. The reduced cost in time and money by applying such a method carries the potential to greatly increase the wide-spread application of proteomic research, as well as the number of complete profiles generated in one study. The predictive software and bioinformatics programs used in this study are freely available, and the QQQ platform on which the MRMs were performed is among the most attainable instruments. Even if a researcher cannot afford such an instrument and must borrow time on one from another lab, the minimal sample preparation means that most of the design can be carried out in an office, allowing for a more efficient use of instrument time. Utilizing *ab initio* MRM design to profile metabolic pathways in complex samples allows researchers greater accessibility for contributing to the ever-growing proteomics field and its parent field of systems biology.

Many improvements to method design, separation, and instrumentation are on the

horizon, as suggested by this and related studies. For example, further research into the effect of peptide length on MRM quality could produce filters that reduce the availability of less meaningful peptides. Also, as explicitly noted, selecting the top 5 or top 6 rated precursors provides a confident expectation of improving MRM quality. Improvements to 1D-LC could further reduce sample complication, or alternately allow even more complex samples to be attempted. Specifically, improvements to retention time prediction software could recover some the peptides lost to ambiguity, by providing more a confident chromatographic identity. As for instrumentation, the aforementioned Q-Exactive is becoming increasingly popular in proteomic MRM, and represents a substantial leap forward in acquiring MRM data at superior mass resolution and accuracies, thereby reducing ambiguity and interference overlaps. Performing improved *ab initio* based MRMs of complex samples on such an instrument while utilizing the latest updates to 1D-LC will provide the ultra-fast and highly accurate metabolic pathway profiles necessary to feed the rapidly increasing demand for comprehensive protein level views of cellular genetic expression.

In total, the MRM-MS approach has seen remarkable advancements and implementation in recent proteome research applications. While the MRM-MS method has been widely used in other scientific arenas, the application to systems biology is a recent venue. The work demonstrated in this thesis assists in defining the experimental landscape for this methodology by systematically examining how to best define and optimize approaches for high-throughput measurements of moderately complex microbial systems.

LIST OF REFERENCES

1. Sabido, E., N. Selevsek, and R. Aebersold, *Mass spectrometry-based proteomics for systems biology.* Current Opinion in Biotechnology, 2012. **23**(4): p. 591-597.
2. Sauer, U., M. Heinemann, and N. Zamboni, *Genetics - Getting closer to the whole picture.* Science, 2007. **316**(5824): p. 550-551.
3. Baitaluk, M., *System biology of gene regulation.* Methods in molecular biology (Clifton, N.J.), 2009. **569**: p. 55-87.
4. Romualdi, C. and G. Lanfranchi, *Statistical Tools for Gene Expression Analysis and Systems Biology and Related Web Resources*, in *Bioinformatics for Systems Biology*, S. Krawetz, Editor. 2009, Humana Press. p. 181-205.
5. Snoep, J. and H. Westerhoff, *From isolation to integration, a systems biology approach for building the Silicon Cell*, in *Systems Biology*, L. Alberghina and H.V. Westerhoff, Editors. 2005, Springer Berlin Heidelberg. p. 13-30.
6. Gall, Z., et al., *Liquid chromatography-mass spectrometric determination of rufinamide in low volume plasma samples.* Journal of chromatography. B, Analytical technologies in the biomedical and life sciences, 2013. **940**: p. 42-6.
7. Science, M.; Available from: http://www.matrixscience.com/help/fragmentation_help.html.
8. Zhao, Y. and A.R. Brasier, *Applications of selected reaction monitoring (SRM)-mass spectrometry (MS) for quantitative measurement of signaling pathways.* Methods, 2013. **61**(3): p. 313-322.
9. Picotti, P. and R. Aebersold, *Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions.* Nature Methods, 2012. **9**(6): p. 555-566.
10. Van Oudenhove, L. and B. Devreese, *A review on recent developments in mass spectrometry instrumentation and quantitative tools advancing bacterial proteomics.* Applied Microbiology and Biotechnology, 2013. **97**(11): p. 4749-4762.
11. Hewel, J.A., et al., *Targeted protein identification, quantification and reporting for high-resolution nanoflow targeted peptide monitoring.* Journal of Proteomics, 2013. **81**: p. 159-172.
12. Abbatiello, S.E., et al., *Design, implementation and multisite evaluation of a system suitability protocol for the quantitative assessment of instrument performance in liquid chromatography-multiple reaction monitoring-MS (LC-MRM-MS).* Molecular & cellular proteomics : MCP, 2013. **12**(9): p. 2623-39.
13. Li, G., et al., *Quantitative determination of arenobufagin in rat plasma by ultra fast liquid chromatography-tandem mass spectrometry and its application in a pharmacokinetic study.* Journal of chromatography. B, Analytical technologies in the biomedical and life sciences, 2013. **939**: p. 86-91.
14. Vierikova, M., E. Hrnciarikova, and J. Lehotay, *DETERMINATION OF NATAMYCIN CONTENT IN CHEESE USING ULTRA PERFORMANCE LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY.* Journal of Liquid Chromatography & Related Technologies, 2013. **36**(20): p. 2933-2943.

15. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-968.
16. Giannone, R.J., et al., *Proteomic Characterization of Cellular and Molecular Processes that Enable the Nanoarchaeum equitans-Ignicoccus hospitalis Relationship.* Plos One, 2011. **6**(8).
17. Dionex. *Proteomics*. Available from: http://www.dionex.com/en-us/markets/life-science/protein-sciences/proteomics/lp-80196.html.
18. Neta, P., et al., *Collisional Energy Dependence of Peptide Ion Fragmentation.* Journal of the American Society for Mass Spectrometry, 2009. **20**(3): p. 469-476.
19. Scientific, T.F., *LTQ Series Hardware Manual*. B ed. 2011.
20. Molina, H., et al., *Comprehensive comparison of collision induced dissociation and electron transfer dissociation.* Analytical Chemistry, 2008. **80**(13): p. 4825-4835.
21. Finnigan, T., *TSQ Quantum Hardware Manual*. Vol. C. 2002.
22. Tabb, D.L., C.G. Fernando, and M.C. Chambers, *MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis.* Journal of Proteome Research, 2007. **6**(2): p. 654-661.
23. Holman, J.D., Z.-Q. Ma, and D.L. Tabb, *Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker.* Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], 2012. **Chapter 13**: p. Unit13.17-Unit13.17.
24. Monroe, M.E., et al., *MASIC: A software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features.* Computational Biology and Chemistry, 2008. **32**(3): p. 215-217.
25. Mallick, P., et al., *eComputational prediction of proteotypic peptides for quantitative proteomics.* Nature Biotechnology, 2007. **25**(1): p. 125-131.
26. Frank, A. and P. Pevzner, *PepNovo: De novo peptide sequencing via probabilistic network modeling.* Analytical Chemistry, 2005. **77**(4): p. 964-973.
27. Stevenson, S.E., N.L. Houston, and J.J. Thelen, *Evolution of seed allergen quantification - From antibodies to mass spectrometry.* Regulatory Toxicology and Pharmacology, 2010. **58**(3): p. S36-S41.
28. Kanehisa, M.G., S; Kawashima, S; Okuno, Y; Hattori, M. *Kyoto Encyclopedia of Genes and Genomes*. Available from: http://www.genome.jp/kegg-bin/show_pathway?org_name=map&mapno=00020&mapscale=&show_description=hide.
29. Evans, M.C.W., B.B. Buchanan, and D.I. Arnon, *A NEW FERREDOXIN-DEPENDENT CARBON REDUCTION CYCLE IN A PHOTOSYNTHETIC BACTERIUM.* Proceedings of the National Academy of Sciences of the United States of America, 1966. **55**(4): p. 928-&.
30. Horton, R.M., Laurence; Ochs, Raymond; Rawn, David; Gray, Scrimgeour, *Principles of Biochemistry*. Third Edition ed. 2002, Upper Saddle River, NJ: Pearson Education.

31. Allen, L.C. and A.M. Karo, *BASIS FUNCTIONS FOR ABINITIO CALCULATIONS.* Reviews of Modern Physics, 1960. **32**(2): p. 275-285.

VITA

Adam Martin was born in Greeneville Tennessee. After spending 3 years training as an EMT and working in various medical professions, he enrolled at University of Tennessee Chattanooga where he majored in chemistry. After two semesters of undergraduate research he began investigating graduate school as a follow up to his undergraduate studies. After receiving his B.S. in 2009, he enrolled in the graduate chemistry program at the University of Tennessee, where studied mass spectrometry under Dr. Compton and Dr. Hettich at UT and Oak Ridge National Laboratory respectively.