Automated Quantification of 13C Labeled Peptides

A MASTER'S THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Joshua Elliot Goldford

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTERS OF SCIENCE

Igor G. Libourel, Adviser

August 2013

Acknowledgements

This thesis is the amalgamation of input from a wide range of incredibly intuitive and helpful scientists, engineers, mathematicians and everything in between. I'd like to thank Dominic Mandy for all the discussions that helped me understand mass spectrometry from a statistical point of view. The perspective gained from those discussions directly influenced the algorithms described in this thesis. I would have not made progress so quickly if it weren't for the incorporation of visualization tools into my development workflow, and I have Eli Krumholz to thank for that. Other members of the Libourel Lab and members of my graduate cohort, including David Burdge, Nagendra Palini, Hong Yang, Eric Lenneman and Audrey Harris provided helpful advice during practice talks and lab meetings. I'd like to that Doug Allen and James Gearse for generating Orbitrap data used for all analysis in this thesis. I'd especially like to thank my advisor for all the time spent discussing my project and providing me with timely advice and suggestions that have helped me grow tremendously as a scientist.

Dedication

This thesis is dedicated to my mother.

Abstract

Metabolic flux analysis (MFA) is a technique used to elucidate intracellular reaction rates (fluxes) in a metabolic network. Intracellular fluxes are determined by providing substrate enriched with stable, heavy isotopic label and subsequently measuring the incorporation of label into metabolic end products. This results in metabolic end products consisting of isomers of discrete mass states, termed *isotopomers*. The resulting isotopomer distributions (MIDs) for each metabolic end product are then used to infer fluxes.

Typically, metabolic end products used for MFA are derivatized protein-bound amino acids. Protein is extracted from the sample and hydrolyzed into constitutive amino acids, resulting in a amino acid pools derived from all cellular protein. Each amino acid pool contains amino acids potentially synthesized from different subcellular compartments, subspecies within a culture, or from different time points within the cell cycle. Thus, fluxes inferred from hydrolyzed total protein lack spatial and temporal resolution. However, if amino acid MIDs were to be measured directly from individual proteins, one could derive the fluxes at the time and place for which that particular protein was synthesized. Therefore, obtaining amino acid MIDs from individual proteins could enable spatial and temporal resolution for metabolic flux analysis. One solution would be to purify individual protein and hydrolyze and measure amino acid MIDs. This approach would require a significant amount of protein, is manually intensive and expensive. A much more viable solution utilizes high-throughput and high-resolution mass spectrometry to quantify and identify peptide MIDs, which can be used to infer constitutive amino acid MIDs. However, there is no well-defined, automated framework for the extraction and quantification of peptide

MIDs from raw mass spectra.

In the first chapter, the conceptual framework and vocabulary need for mass spectrometry and peptide-based MFA are provided, with a statistical emphasis. Chapter 2 provides a review of proteomics instrumentation for peptide based MFA followed by the algorithmic considerations and potential software solutions available for the extraction of peptide MIDs.

Chapter 3 will describe the methods developed for the automated extraction and quantification of isotopically enriched peptides, including parameter optimization of existing methods and description of novel clustering and quantification methods. Chapter 4 describes the validation of the methods using three different sets of labeled peptide MIDs. Chapter 5 provides a brief discussion of method and software improvements for both identification and quantification followed by a brief discussion of future work.

Contents

List of Tables **viii**

List of Figures **ix**

1 Preliminaries **1**

1.1 Isotopomers and Mass Distributions 1

1.2 Mass Spectrometry 4

1.3 Simulating MIDs 6

1.3.1 Mixing 6

1.3.2 Condensation reactions 8

1.3.3 Cleavage reactions 9

1.4 Metabolic Labeling 11

1.4.1 Uniformly Labeled Substrate 12

1.4.2 Positionally Labeled Substrate 13

1.5 Flux Analysis 15

1.5.1 ^{13}C Metabolic Flux Analysis 15

1.5.2 Peptide-based Metabolic Flux Analysis 16

1.6 Scope of this thesis 17

2 Proteomics for Peptide-based MFA **19**

2.1 Instrumentation: Orbitrap 20

2.1.1 Mass Accuracy, Resolution and Sample Rate 21
2.1.2 Dynamic Range and RIA Quantification 21
2.2 Isotopic Labeling for Proteomics 22
2.3 Data Analysis: Approaches and Tools 24
2.3.1 LC-MS Preprocessing 25
2.3.2 Software 26

3 Methods: Software Design and Development 28
3.1 Experimental Design 29
3.1.1 Workflow 29
3.1.2 Training Set 29
3.2 Peptide Identification 31
3.3 Reduction of Labeled Data to Mass Traces 35
3.4 Feature Identification 40
3.4.1 Filtering Raw Data 41
3.4.2 Clustering Mass Traces 42
3.5 Feature Quantification 50
3.5.1 Estimating MIDs From Experimental Data 51
3.5.2 Estimation of Experimental Error 53

4 Software Validation and Implementation 56
4.1 Unlabeled *E. coli* 56
4.1.1 Analysis of Feature Identification and Extraction 57
4.1.2 Analysis of Quantification 59
4.2 7% Uniform Labeled *E. coli* 62
4.2.1 methods 62
4.3 Labeled Soy Samples 66
4.4 Summary 68

5 Discussion and Future Work **69**

5.1 Software Improvements 70

5.1.1 Identification 70

5.1.2 m/z Calibration 71

5.1.3 Feature Annotation 71

5.2 Quantitative Capabilities 74

5.2.1 RIA Estimation 74

5.2.2 Variance Estimation 75

5.3 Future Work 75

References **77**

A Software Design **84**

A.1 File Preparation 84

B Statistical Interpretation of Mass Isotopomers **86**

B.1 Quantification of Isotopomer Ion Counts 87

B.2 Isotopomer Ion Counts as Random Multinomial Samples 88

C Raw Data **90**

D Validating RIA LC Bias **94**

List of Tables

3.1 Training set 31
3.2 Mass trace extraction parameter list 37

List of Figures

1.1 Isotopomers and Isotopologues 2
1.2 Mass Isotopomer Distribution (MID) 3
1.3 MID in Time 5
1.4 Example: Mixing of Isotopologue Populations 7
1.5 Example: Discrete Convolution 10
1.6 Uniform Labeling Within a Metabolic Network 12
1.7 Example: Relationship between Uniform labeling and Isotopomer Distributions 13
1.8 Positional Labeling Within a Metabolic Network 14
1.9 Example: Relationship between Flux and Isotopomer Distributions . 15
1.10 Peptide-based Metabolic Flux Analysis 17

2.1 Peptide MIDs: MFA vs. Quantitative Proteomics 24

3.1 Peptide MID Quantification Workflow 30
3.2 Comparing m/z recalibration for multiple samples 35
3.3 Mass trace similarity 39
3.4 Retention time length and ion abundance 40
3.5 Similarity Matrices 44
3.6 Clustering Similarity Matrices 45
3.7 Cluster analysis stopping criteria optimization 48

3.8 Feature Identification from Cluster Analysis 50
3.9 Training Data Example 52
3.10 Deriving Estimated Sample Variances 55

4.1 RIA and isotopic erichment for unlabeled *E. coli* 62
4.2 Cluster result example for labeled *E. coli* 63
4.3 Cluster results calculated MIDs for labeled *E. coli* 64
4.4 RIA and Isotopomer Enrichment Analysis of 7% *E. coli* 66
4.5 Simulated vs. measured RIA in *Soybean* 68

5.1 Using Average Carbon Label to Simulate PMDs 72
5.2 Experimental Design Example: Spiking with Known PMDs 73
5.3 Interactive Data Viewer 76

A.1 File formats 85

C.1 Feature Matrix: training data 92
C.2 Candidate Feature Sets: training data 93

D.1 Histogram of Weighted RIA Slope Differences 96

Chapter 1

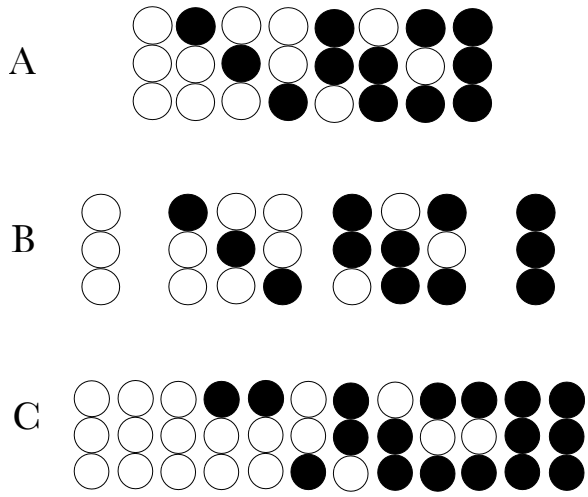
Preliminaries

The resurgence of central carbon metabolism as an active area of research has been motivated by a variety of scientific goals within the 21^{st}-century, ranging from the efficient production of biofuels to the characterization of cancer physiology (Clomburg and Gonzalez, 2010) (Keibler et al., 2012) . More fundamentally, an overarching goal in systems biology is to characterize the underlying metabolic response to genetic and environmental perturbations. Studying cellular metabolism from a systems level allows researchers to understand the underlying physiology for a given phenotype. Although the biochemical characteristics of metabolic enzymes are well established *in vitro*, the development of tools and techniques aimed at characterizing cellular metabolism *in vivo* are ongoing, and primarily rely on measuring the outputs of a metabolic network after the incorporation of isotopically enriched (labeled) substate (Wiechert, 2001). In this chapter, a description of the vocabulary and concepts associated with the use of stable isotope labeling to study metabolism is provided.

1.1 Isotopomers and Mass Distributions

For any given molecule, the collection of isotopic isomers are defined as **positional isotopomers**. Each positional isotopomer can be partitioned by mass state into

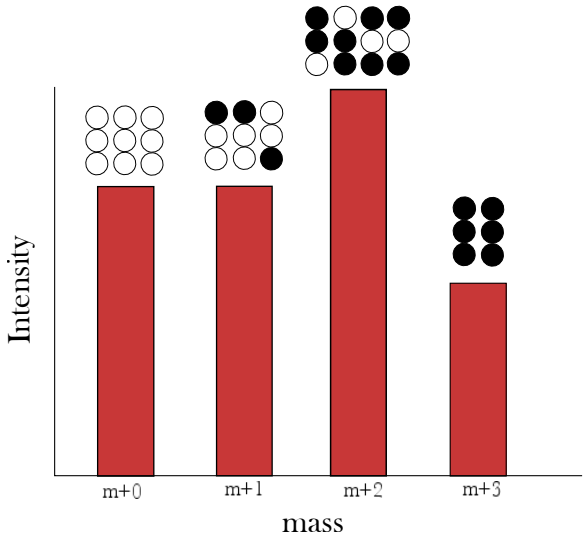
Figure 1.1: 3 atom molecule with 2 discrete mass states (black and white); A. Positional Isotopomers, B. Mass Isotopomers, C. Isotopologues



groups called **mass isotopomers**. For example, suppose we have a molecule with 3 atoms, where each atom existed in 2 distinct mass states. This molecule can exist in $2^3 = 8$ different atomic positional configurations. (see Figure 1.1). However, when grouping the set of positional isotopomers by mass, only four unique mass states are observed. Furthermore, we can define the mass state of a particular molecule to be a discrete random variable, X sampled from a discrete sample space of possible mass states, χ. It is conventional to define each mass isotopomer as m+i, where i are the number of heavy labeled atoms. Moreover, we will define $\chi = \{0, 1 \dots m\}$, where m are the number of atoms within a particular molecule.

For any population of molecules, mass isotopomers are present at specific frequencies. The entire group of isotopic isomers within a *sample* of this population are defined as **isotopologues**, or $\mathcal{I}$. For a given group of isotopologues, frequencies of mass isotopomers can be quantified using mass spectrometry. We can express mass isotopomer frequencies as a distribution, defined as a **mass isotopomer distribution (MID)**, which can be expressed in *absolute* or *relative* abundances (see Figure 1.2). Furthermore, we can express absolute or relative MIDs as vectors, I

Figure 1.2: Collection of isotopologues at specific frequencies: In mass spectrometry data, mass to charge (m/z) ratios paired with intensity measurements resolve the frequencies of mass isotopomers within a group of isotopologues



and p. The i^{th} element of I and p correspond to the absolute number or the relative frequency of sampled isotopologues for mass isotopomer m+i, respectively. The i^{th} element in the relative MID is defined as the **relative isotopic abundance** (RIA) for isotopomer m+i, where

$$p_i = \frac{I_i}{\mathcal{I}} \tag{1.1}$$

From a statistical point of view, we can define the probability of sampling mass isotopomer m+i as $\Pr\{X = i\}$, which can be estimated from the relative frequency of m+i. For example, in Figure 1.2 there are 12 molecules that make up the group of isotopologues, three of which contain no heavy label (i.e. mass isotopomer m+0). The probability of sampling m+0 can be estimated from the observed sampling frequency of m+0, where $\Pr\{X = i\} \approx p_0 = \frac{3}{12} = 0.25$. The sampling frequency approximates the probability when the sample size, $\mathcal{I}$, increases.

By establishing the connection between probability theory and MIDs, it is possible

to model a MID as a **multinomial distribution** where the observation of each molecule is an independent sample resulting in a success for one of many mass states, with each mass state having a fixed success probability (Casella and Berger, 2001). For a sample size of $\mathcal{I}$ isotopologues, the expected value for mass isotopomer i can be calculated, such that:

$$E(X = i) = \mathcal{I} \Pr\{X = i\} \approx \mathcal{I} p_i \tag{1.2}$$

Additionally, a sampling variance and standard deviation can be estimated using the following equation:

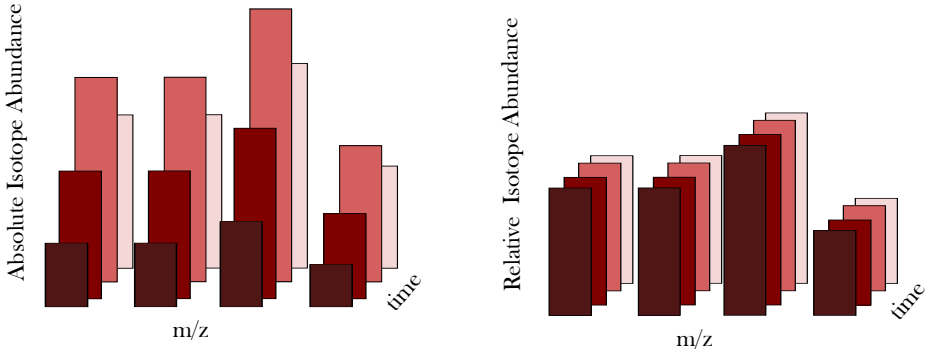
$$\mathrm{Var}\,(X = i) \approx \mathcal{I} p_i (1 - p_i) \quad \text{and} \quad \sigma(X = i) \approx \sqrt{\mathcal{I} p_i (1 - p_i)} \tag{1.3}$$

Equations (1.2) and (1.3) can be used to simulate isotopomer distributions, errors and confidence intervals for any isotopologue sample size.

1.2 Mass Spectrometry

The analytical method that resolves MIDs is high resolution mass spectrometry (MS), where signals proportional to the isotopologue abundance are measured as a function of mass over charge (m/z). Populations of ionized isotopologues are sampled simultaneously, and current is measured over a period of time. The transient signal is then converted to the *frequency domain* using the Fourier transform, where individual frequency measurements correspond with m/z, resulting in isotopically resolved mass spectra from a single sample. Moreover, liquid chromatography - mass spectrometry (LC-MS) is a powerful technique that couples the physical separation of molecules in a temporal domain with highly resolved mass measurements (Scigelova and Makarov, 2006).

Figure 1.3: MID as a function of time: In mass spectrometry data, MIDs are observed in multiple sampling events providing; Each isotopomer within a group of isotopologues exhibit different *absolute* intensities in time, however maintain the same *relative* intensity



In LC-MS, isotopologues elute off the column continuously and samples are injected into the mass analyzer at discrete time points, called **scan events**. A **mass trace** is defined as a single m/z measurement observed within contiguous scan events. Assuming a mass trace contains all measurements for a particular mass isotopomer, tracking the abundance of a mass trace across multiple scan events is called an **extracted ion chromatogram** (EIC). Figure 1.3 is an example of time-dependent mass spectra obtained from an LC-MS experiment. In this example, different scan events are denoted by different colors, while the collection of intensity measurements for one mass isotopomer would be the EIC (Yates et al., 2009).

From a statistical point of view, LC allows for multiple samples to be drawn from the same population of isotopologues, which enables more accurate quantification of MIDs. In Figure 1.3 it is shown that although the total number of sampled mass isotopomers vary through time, the RIA for each isotopomer remains the same. The temporal invariance of RIAs illustrate that differences in absolute isotopomer abundances though time is a function of isotopologue sample size, which is heavily exploited in the automated extraction and quantification of MIDs in Chapter 3.

1.3 Simulating MIDs

Techniques that use stable isotopes to generate information about the utilization of a metabolic network require the ability to formulate models of MIDs. Thus, in addition to rigorous quantification and characterization, the basic principles that are used to simulate MIDs warrant discussion (Hellerstein and Neese, 1999).

Three basic scenarios arise when attempting to simulate MIDs: (1) **mixing** of two isotopologue populations for the same molecule, (2) **condensation reactions** and (3) **cleavage reactions**. Predicted MIDs from mixing and condensation can be solved analytically, while cleavage reactions require an algorithmic approach. Characteristics and differences between each scenario are highlighted in the following subsections.

1.3.1 Mixing

Figure 1.4 illustrates the mixing of isotopologue distributions, which is also called *tracer* or *self-diffusion*. In this example, a two atom molecule initially exhists in two, separate isotopologue populations, denoted $\mathbf{A}_1$ and $\mathbf{A}_2$, with MIDs A_1 and A_2, respectively. Suppose a third population , $\mathbf{A}_{1+2}$ (with a MID, A_{1+2}) is comprised of 20% of isotopologues originated from $\mathbf{A}_1$ and 80% originating from $\mathbf{A}_2$. The RIA for the m+1 mass isotopomer of $\mathbf{A}_{1+2}$ would be determined by adding the m+1 RIAs for $\mathbf{A}_1$ and $\mathbf{A}_2$ scaled by the relative proportions of the isotopologues, where

$$a_{1,1+2} = 0.20 \times 0.50 + 0.80 \times 0 = 0.10$$

More generally, if we define the **population ratio parameter**, r, the i^{th} element in A_{1+2} can be uniquely determined, such that

$$a_{i,1+2} = r \times a_{i,1} + (1 - r) \times a_{i,2} \tag{1.4}$$

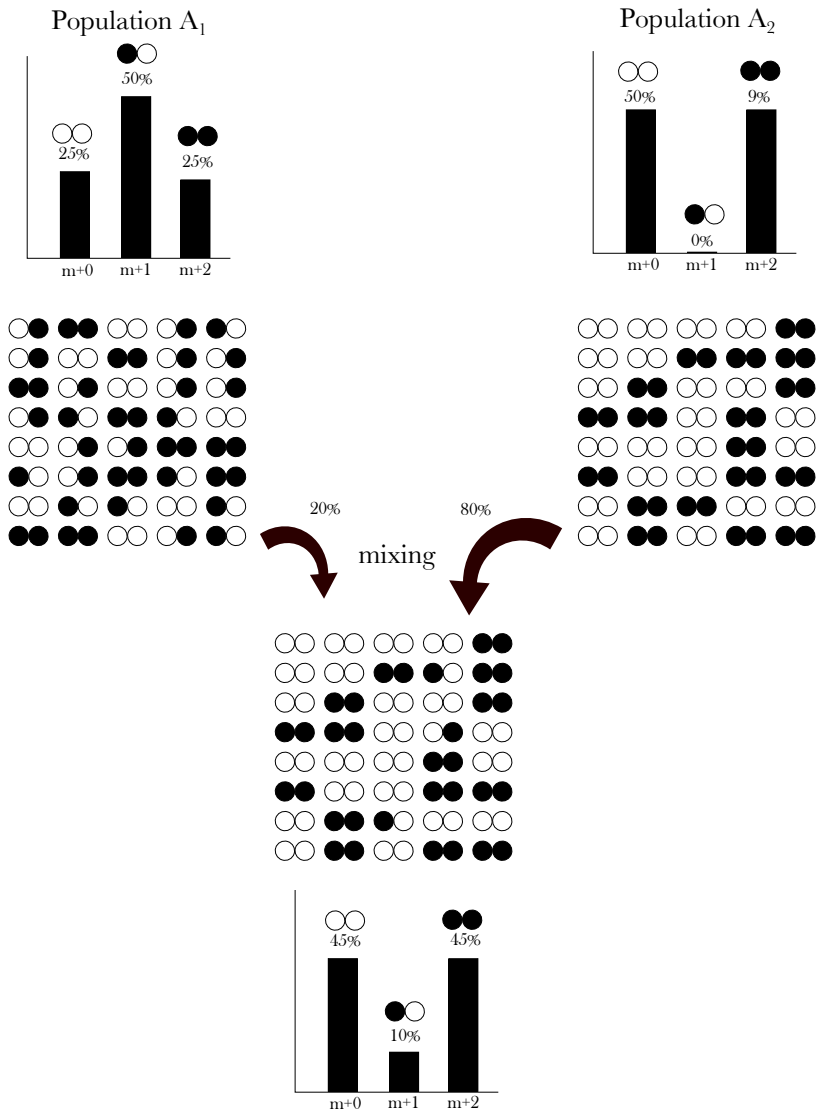


Figure 1.4: Mixing of Isotopologue Populations: Two isotopologue populations consisting of the same atomic composition are mixed with different frequencies - 20% of population $\mathbf{A}_1$ is mixed with 80% of $\mathbf{A}_2$

1.3.2 Condensation reactions

If MIDs for substrates involved in an addition reaction are known, then product MIDs can be calculated using an operation called **discrete convolution**. Suppose molecules **A** and **B** are substrates in a condensation reaction with MIDs A and B, respectively. The product, **C** has a MID, C that can be determined using the following relationship:

$$\mathbf{A} + \mathbf{B} \rightarrow \mathbf{C} \Longrightarrow A * B = C \tag{1.5}$$

where the $*$ operation is the convolution operator. An intuitive description of convolution is illustrated in Figure 1.5, where substrates in the condensation reaction consist of a single atom existing in one of two mass states; 70% of the population exists in the light mass state (m+0) while the remaining 30% is in the heavy mass state (m+1). Two independent, random samples of individual atoms are drawn from the precursor pool in a sampling event, polymerized and enter into a product pool. If the sampling and condensation occur enough times and without bias for a particular mass state, the RIAs of the product can be calculated based on the sampling probability of substrates. The entire space of sampling events can be generated and RIAs can be estimated for all possible product mass isotopomers. For example, synthesizing an m+1 consisting of one heavy and one light atom can occur in two different ways - a heavy atom can be sampled first, followed by a light or visa-versa. The probability of sampling an m+1 product, p_1 could be determined as follows:

$$p_1 = (0.7)(0.3) + (0.3)(0.7) = 0.42$$

Discrete convolution can be generalized for arbitrary substrate MIDs A and B with elements a_i and b_j, respectively, such that:

$$\begin{aligned} c_0 &= a_0 b_0 \\ c_1 &= a_1 b_0 + a_0 b_1 \\ c_2 &= a_2 b_0 + a_1 b_1 + a_0 b_2 \\ &\vdots \\ c_n &= a_n b_0 + a_{n-1} b_1 + \ldots + a_0 b_n \end{aligned} \tag{1.6}$$

where

$$\begin{aligned} i > |A| &\Rightarrow a_i = 0 \\ j > |B| &\Rightarrow b_j = 0 \end{aligned}$$

where c_k is the k^{th} element in the product MID, C. Additionally, MIDs for molecules consisting of elements with a fixed probability can also be calculated using a binomial distribution model, however suffer when the MID is simulated from substrates with different elemental or molecular distributions. A discussion of labeling scenarios is left for Section 1.4, within the context of isotopic propagation through a metabolic network.

1.3.3 Cleavage reactions

The third type of reaction is a cleavage reaction, which is of the form:

$$\mathbf{C} \rightarrow \mathbf{A} + \mathbf{B} \tag{1.7}$$

Where the substrate MID, C, is split into product distributions A and B. Determining product distributions A and B is an *inverse problem* which requires an algorithm-based approach to iteratively fit for optimal product MIDs - a processes called **deconvolution** (Jansson, 1996). The algorithm first guesses product MIDs, A and B,

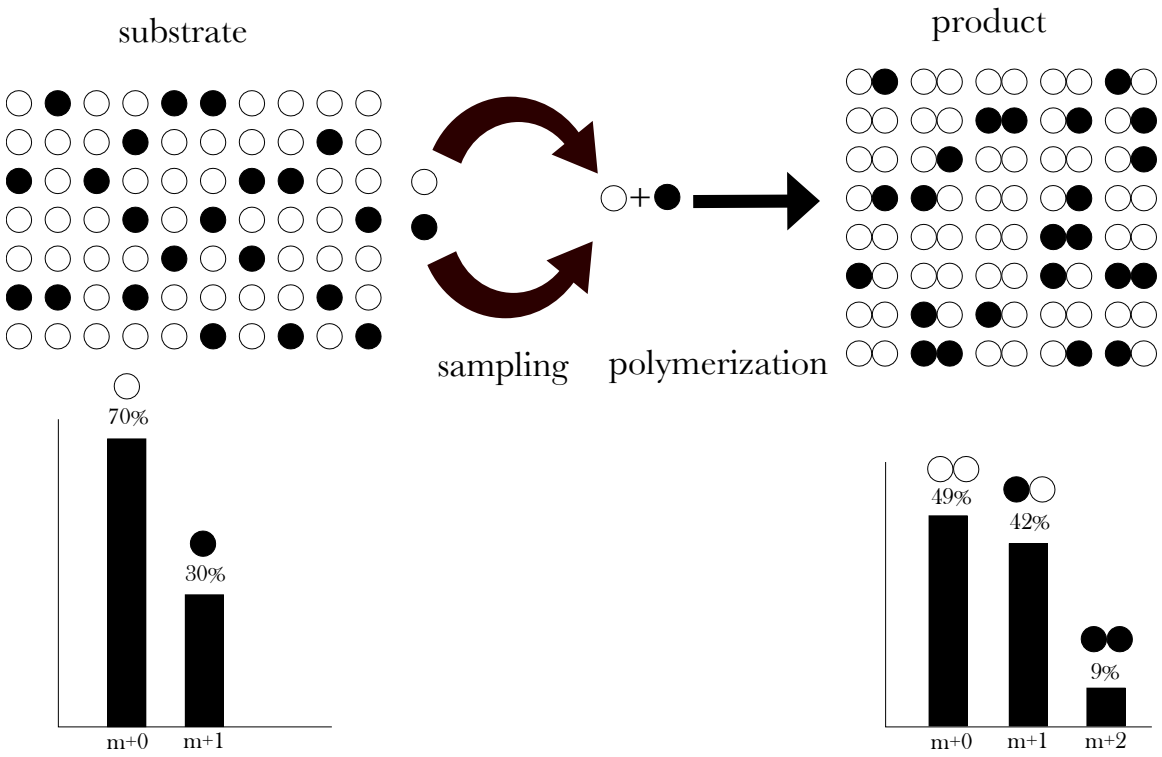


Figure 1.5: Discrete Convolution of Mass Isotopomer Distributions: Sample of the substrate population are taken independently, condensed and added to the product pool. The relative frequency for each product mass isotopomer can be calculated using discrete convolution

followed by simulation of substrate C via convolution. An objection function, f is then used to map the individual differences between measured and simulated RIAs of C, called *residuals* into a scalar value representing the quality of fit, called a *residuum*. An optimization algorithm (e.g. interior points or sqp) uses a gradient-based search to either simulate a new set of product MIDs that provide a better fit or terminate if optimality criteria are satisfied. More compactly, for cleavage reactions, the following relationship holds:

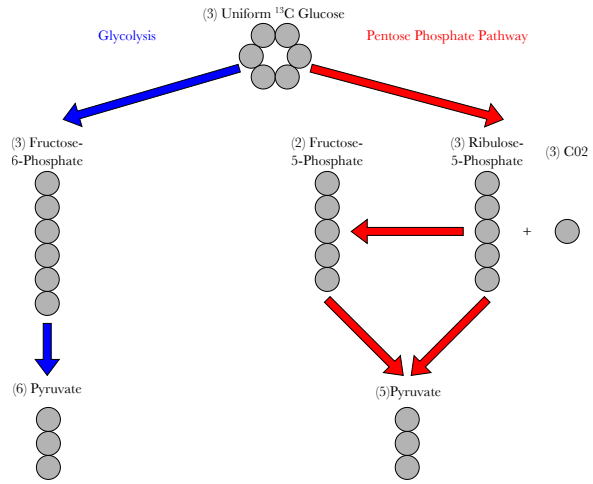
$$\mathbf{C} \rightarrow \mathbf{A} + \mathbf{B} \Longrightarrow \underset{A,B}{\text{minimize}} \;\; f(C - A * B) \tag{1.8}$$

There are situations when A and B can be calculated analytically, however in practice solutions are confounded when error is introduced in the known distribution C. Thus, fitting remains the most practical approach when faced with a cleavage problem.

1.4 Metabolic Labeling

The ultimate motivation for quantifying MIDs is to gain insight on the functional utilization of a biochemical networks by quantifying the molecular traffic, or *fluxes*, through reactions. Examples of substrate labeling conditions have been provided thus far have only served to illustrate the formulation of MIDs in single reactions, rather than within the context of fluxes through a metabolic network. To this end, metabolic end product MIDs are characterized based on two substrate labeling examples given different flux scenarios.

Figure 1.6: Example Metabolic Network: Uniformly labeled glucose enters in the metabolic network and is partitioned between glycolysis and the pentose phosphate pathway (PPP). The pyruvate product distribution contains the same elemental labeling distributions as the glucose substate, leading to a product MID invariant to flux distributions



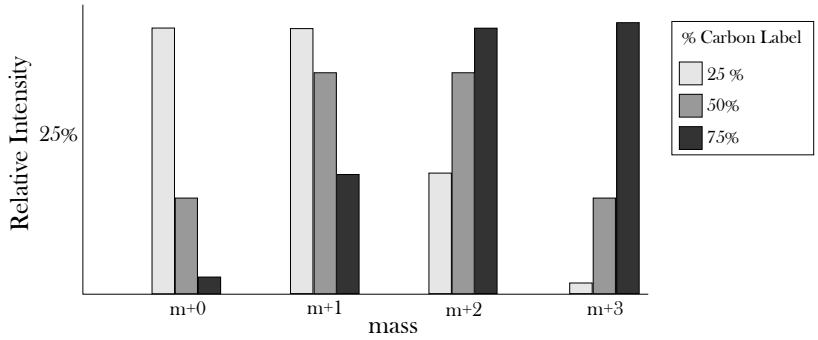
1.4.1 Uniformly Labeled Substrate

Figure 1.6 is an example of a simplified metabolic network for central carbon metabolism. Three molecules of glucose can either flow through glycoloysis or the pentose phosphate pathway (PPP) to produce pyruvate. The rate of flow is defined as **flux**, v, where v_{glyc} and v_{ppp} represent the molecular flow through glycolysis and PPP, respectively.

If the substrate is **uniformly labeled**, elements of the same type within the molecule have equivalent MIDs. Figure 1.6 illustrates this scenario by encoding each atoms' fractional abundance of heavy label in greyscale, where black means 100% of that atom exists in the labeled state and white means 100% of that atom exists in the unlabeled state.

Regardless of the flux distribution (e.g. $\frac{v_{\text{glyc}}}{v_{\text{ppp}}}$) the metabolic end product can be expressed as a convolution of elemental MIDs. For example, in Figure 1.6 pyruvate is generated from both glycolysis and PPP at different rates (e.g. $v_{\text{glyc}} \neq v_{\text{ppp}}$), but

Figure 1.7: Pyruvate MID as a function of different atomic RIAs: The MID of a uniformly labeled MID only dependent on the constitutive elemental MIDs.



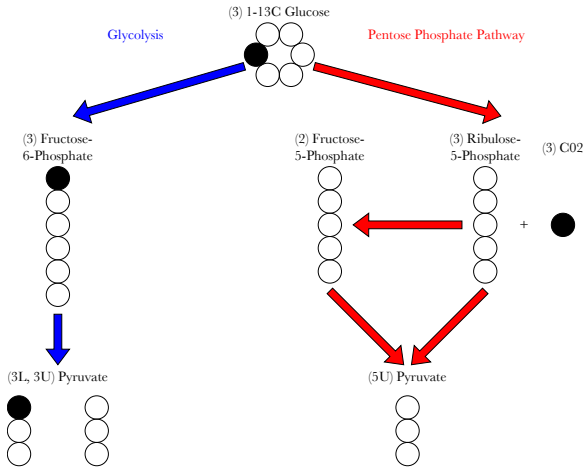
contain the same elemental label distributions. Thus, molecular MIDs at any point in the metabolic network can be expressed as a function of the percentage of elemental label enrichment (see Figure 1.7).

1.4.2 Positionally Labeled Substrate

Figure 1.8 illustrates a simple labeling experiment done to determine the flux through glycolysis and the pentose phosphate pathway (PPP) using ^{13}C labeled glucose at the one position as labeled substrate. When three molecules of glucose pass through glycolysis, six molecules of pyruvate are generated. If glucose is labeled at the one position, two mass isotopomers (m+0 and m+1) of pyruvate are observed in equal amounts. When three molecules of glucose pass through the pentose phosphate pathway, five molecules of pyruvate and three molecules of carbon dioxide are generated. However, when glucose is labeled in the one position, all label is incorporated into carbon dioxide, leaving only the m+0 mass isotopomer of pyruvate.

Figure 1.9 illustrates the relationship between the pyruvate MID and different flux distribution scenarios. In the first case, all flux is directed through glycolysis, which leads to an equal abundances of the m+0 and m+1 isotopomers, such that $p = [0.5, 0.5, 0, 0]$. In the second case all flux is directed through PPP, resulting in

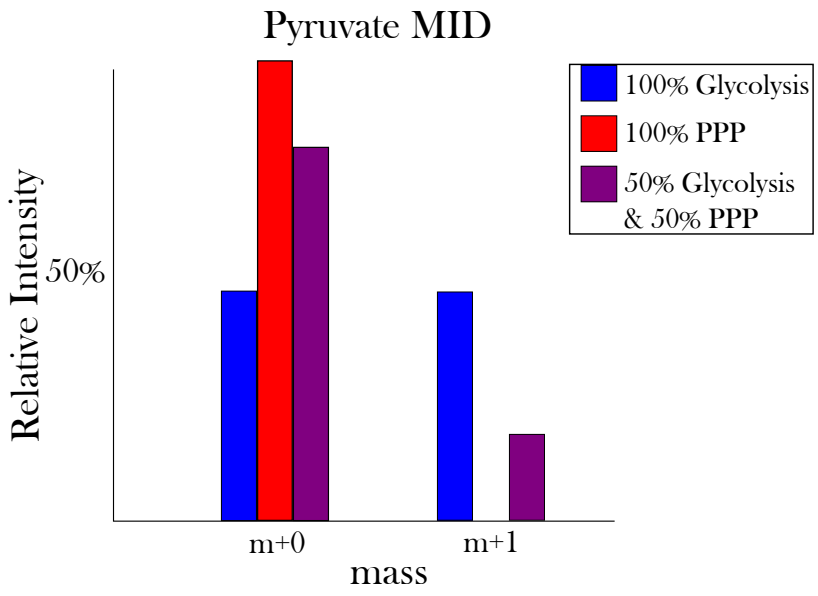
Figure 1.8: Positional Labeling Within a Metabolic Network



100% of the isotopologues occupying the m+0 state, or $p = [1, 0, 0, 0]$. In the third case, flux is split equally between glycolysis and PPP, resulting in a distribution that exists in between the first two cases. The distribution can be quantified via a simple example. Suppose 6 molecules of glucose were distributed equally through glycosis and PPP, resulting in 3 labeled and 8 unlabeled pyruvate molecules, such that $I = [8, 3, 0, 0]$ and $p = [\frac{8}{11}, \frac{3}{11}, 0, 0] = [0.73, 0.27, 0, 0]$.

This simple example illustrates that the metabolic end product MID is uniquely dependent on the distribution of fluxes, meaning that MIDs cannot be simulated unless the flux distribution is known *a priori*.

Figure 1.9: Pyruvate MID as a function of different flux distribution scenarios



1.5 Flux Analysis

1.5.1 ^{13}C Metabolic Flux Analysis

Techniques utilizing stable isotopic tracers to elucidate metabolic activiy has matured over the last decade, most notabily of which being ^{13}C-based metabolic flux analysis (^{13}C MFA) (Wiechert, 2001). The gaol of ^{13}C MFA is to quantify steady state intracellular reaction rates (fluxes) within a metabolic network, and has become a routine procedure within the last decade for experimental microbial metabolism (Zamboni et al., 2009). In this procedure, metabolic fluxes are inferred from isotopic distributions of proteinogenic amino acids (AAMDs) synthesized from ^{13}C enriched substrate. Key technological advances have made performing ^{13}C-MFA for central carbon networks feasible (Wiechert and de Graaf (1997), Wiechert et al. (1997), Wiechert et al. (1999), Möllney et al. (1999)) and computationally efficient using the EMU model (Antoniewicz et al., 2007), where open source computational tools have been become readily available (Quek et al. (2009),Weitzel et al. (2013)).

Conceptually, ^{13}C-MFA is performed using the approach presented in Section 1.3.3, however fluxes are model parameters instead of MIDs. Fluxes are used to simulate AAMDs (AAMD_{sim}) and compared to measured AAMDs ($\text{AAMD}_{\text{meas}}$) using a least squares comparison.

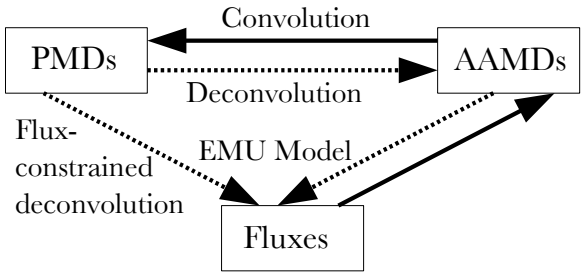
1.5.2 Peptide-based Metabolic Flux Analysis

As reconstructed metabolic networks expand beyond prokaryotic central metabolism and take into account compartmentalization in eukaryotic organisms (Krumholz et al., 2012), new experimental techniques aimed at validating these models are required. Currently ^{13}C-MFA has limitations; derived flux maps of metabolism are spatially and temporally invariant. Measured distributions are derived from all cellular protein, and thus contain amino acids synthesized at different times, sub-cellular compartments or within different sub-species. Because protein translation can be a temporally and spatially regulated process, constitutive AAMDs from specific proteins can report on metabolism at the time and location of translation.

Experimental techniques have been developed recently to determine spatially-specific flux maps. For example Allen et al. (2007) used labeling patterns in fatty acid groups, cell wall components, protein glycans, and starch synthesized in different compartments of the plant cell to derive compartment-specific flux maps. More recently, Rühl et al. (2011) measured AAMDs from an engineered GFP reporter to quantify fluxes from a single species within a microbial consortia.

We recently reported a new proteomics-based approach using peptide MIDs (PMDs) to quantify fluxes (Mandy et al., 2013). In this technique, a labeling experiment is performed, followed by protein extraction, gel purification and enzymatic digestion, resulting in a sample of isotopically enriched peptides. Peptides are separated further using HPLC followed by quantification of MIDs using high resolution MS. Peptide MIDs are then integrated into the fitting framework described for classical ^{13}C-MFA

Figure 1.10: Peptide-based MFA approaches: solid arrows represent forward problems while dotted lines represent inverse problems. Case(1): PMDs are freely deconvolved to AAMDs and fluxes are fitted to deconvoved AAMDs. Case(2): Fluxes are fitted directly to PMDs, resulting in flux-constrained AAMD solutions.



using the EMU model implemented in OpenFLUX^{TM}. Figure 1.10 shows the two approaches used for peptide based MFA. In the first case, PMDs are first deconvolved into constitutive AAMDs, followed by fitting fluxes to deconvolved AAMDs. In the second case, fluxes are directly fitted to PMDs. While the first case allows AAMDs to assume any distribution that best fits the PMDs, the second approach constrains AAMDs with the metabolic network.

The key advantage of peptide-based MFA over other methods is that it utilizes high-throughput proteomic data. Like other high-throughput techniques, rapid technical advances are making proteomics experiments more tractable, enabling researches to produce highly complex proteomic datasets cheaply and quickly. However as datasets become richer, new algorithmic approaches are required to extract meaningful biological information.

1.6 Scope of this thesis

The scope of this thesis project was to utilize the basic principles of probability theory, high resolution MS, and isotopic label propagation through metabolic networks to develop an algorithmic framework and software implementation for the automated

extraction and quantification of PMDs for peptide based metabolic flux analysis. In the next chapter, a brief discussion of current the instrumentation, algorithmic and software requirements needed for peptide based MFA are given. The limitations of algorithms and software designed for the proteomics and metabolomics community is of central focus and provide motivation for methods developed in Chapter 3.

Chapter 2

Proteomics for Peptide-based MFA

Proteomics is a scientific discipline that has seen rapid development over the last 15 years (Mallick and Kuster, 2010). In addition to advances in technology aimed at high-throughput peptide sequencing in complex biological samples, a variety of techniques have been developed to obtain quantitative information using stable isotopic labeling both *in vitro* and *in vivo* (Bantscheff et al. (2007), Bantscheff et al. (2012)). Comprehensive overviews of key technologies and techniques that capture the evolution of the field are vast and are beyond the scope of this thesis. Instead, it is imperative that we describe the instrument and software specifications required for the experimental design of peptide-based MFA, while simultaneously highlighting differences between current techniques and algorithms used for the high-throughput quantitative proteomics experiments.

This chapter first provides an overview of the instrumentation required for peptide-based MFA with a description of the data generated after an experiment is performed. Next, different techniques used in quantitative proteomics are described, highlighting the differences between the data produced using these techniques and data generated from a flux analysis experiment. Next a review of algorithms and software is provided, with an emphasis on frameworks available for the automated extraction of isotopically enriched proteomics datasets.

2.1 Instrumentation: Orbitrap

Peptide-based MFA requires instrumentation that is capable of quantifying RIAs for an unknown number of mass isotopomers. For isotopologues of peptides taken from samples with no isotopic enrichment, isotopic clusters are observed from the naturally occurring heavy isotopes of C, N, S, H and O. However, the number of observable mass isotopomers for a peptide synthesized from natural substrate are significantly less than when peptides are synthesized from enriched substrate. This places greater demand on the instruments ability to provide *accurate*, *precise*, and *isotopically precise* mass measurements. For clarity, *accurate* is defined as the instruments ability to report the true mass of a molecular species while *precise* means the reproducibility of mass measurements from different biological, technical or chromatographic samples. *Isotopically precise* means that the mass differences between each isotopomer and the theoretical mass should be consistent. For example, if the measured m+0 isotopomer is 5 ppm from the theoretical mass, then all other mass isotopomers should be 5 ppm away from their theoretical masses.

The instruments most suitable for these requirements are bench-top mass spectrometers coupled to an **Orbitrap** mass analyzer (Hu et al., 2005). The Orbitrap has been usd to identify an unprecedented number of peptides and proteins from highly complex samples. For example, Nagaraj et al. (2012) were able to identify 4000 proteins from 30,000 sequenced peptides from a single orbitrap experiment, resulting in nearly complete coverage of the yeast proteome. Such high identification rates are a function of the key attributes of the Orbitrap: **mass accuracy, resolution** and **dynamic range** (Makarov et al., 2006).

2.1.1 Mass Accuracy, Resolution and Sample Rate

Since the invention of the Orbitrap in 2005, a number of studies have verified high mass accuracy, ranging from 2-5 ppm (Makarov et al. (2006), Perry et al. (2008)). However, sub-ppm mass accuracy has been demonstrated when using real-time, "lock" mass calibration (Olsen et al., 2005). Additional computational techniques have also increased the mass capability of the instruments, enabling measurements to reach the ppb range by matching mass measurements from identical peptides in different charge states (Cox and Mann, 2009).

While the linear trap quadrapole (LTQ) coupled to an Orbitrap can achieve a resolution as high as 150,000, they are typically run between 30,000 because of slower scanning rates, resulting in a smaller number of samples. Typically, the resolution for an LTQ Orbitrap is set at 30,000, a mass accuracy of 5ppm is achievable with an average sample time of 1 second (Yates et al., 2006). Newer instruments like the Q-Exactive Orbitrap have been shown to perform better than the LTQ-Orbitrap by simultaneously achieving faster scanning rates and mass resolution. For example, for the ultra high resolution Orbitrap Elite^{TM}, the spectra for a complete isotopically resolved protein in the +47 charge state was recently demonstrated at a 2.5 second sample time with a resolution over 100,000 (Michalski et al., 2012).

2.1.2 Dynamic Range and RIA Quantification

Compared to characterizing the mass accuracy and resolution of Orbitrap instruments, the quality of RIA quantification has received little attention. In particular, studies have not rigorously assessed the quality of RIAs from any statistical point of view by accounting for sampling variance of isotopologue populations. Initial studies on the dynamic range give some idea on the instruments capacity. For example, a signal dynamic range was shown to be approximately 5000 for early Orbitrap models

(Makarov et al., 2006).

RIA quantification errors were reported to fall within 4-5% of theoretical measurements (Okawa et al., 2013). However, identified peaks were normalized prior to error calculation, thus no consideration was made for sampling variance. Furthermore, there was no discussion on correcting for potential RIA bias. It has been shown that isotope discrimination exists for certain enzymes, most notably being RuBisCO (Wong and Sackett, 1975), and has recently been shown to directly effect labeling patterns (Wasylenko and Stephanopoulos, 2013). RIA accuracy was also shown to decrease for increasing resolution, which is likely caused by decreased sampling frequency (Okawa et al., 2013). However, the tradeoffs between resolution and scan scanning rates on the quantification of RIAs have yet to be rigorously investigated.

Recently, Kaufmann and Walker (2012) have shown that accuracy of RIA quantification could be effected by a number of physical and mathematical phenomena, such as ion suppression, ion coalescence, and artifacts from the Fourier transform (FT). Regardless, these effects are less pronounced in Orbitrap mass spectrometry relative to other forms of FT-MS due to high resolving power and large space filling capacity of the trapping device.

Although the quantitative capacity for measuring RIAs have received little attention, the high mass accuracy, dynamic range, and protein and peptide identification rates makes the technology suitable for the quantification of PMDs for peptide-based metabolic flux analysis.

2.2 Isotopic Labeling for Proteomics

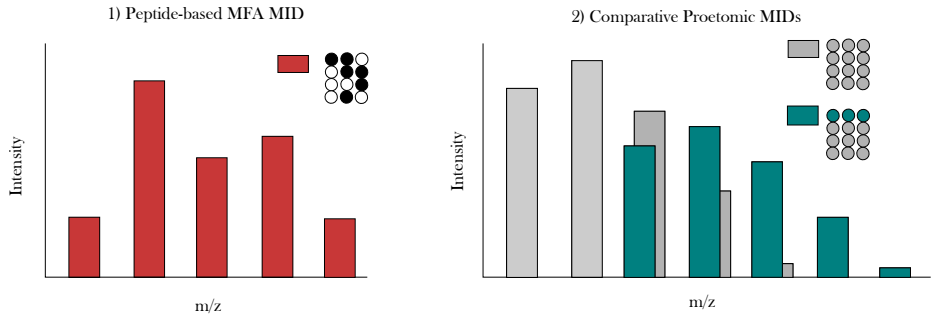
The use of stable isotopes in proteomics is well-established, with techniques ranging from *in vitro* labeling of residues (ICAT), termini (^{18}O), and functional groups (mTRAQ) as well as metabolic, *in vivo* labeling using elemental isotopic enrichment

with ^{15}N or ^{13}C, or incorporation of heavy variants of amino acids (e.g. SILAC) (Becker, 2008).

Most quantitative proteomic experiments that utilize the incorporation of stable isotopes into proteins *in vivo* are used for comparative or differential proteomics. In these experiments, one sample is grown with natural substrate while the other is exposed to isotopically labeled substrate. For example, in a SILAC experiment, the isotopically labeled substrate is typically a ^{13}C, ^{15}N or ^{18}O varient of an amino acid supplied to the cell culture media (Geiger et al., 2011). The two samples are combined and relative protein quantities are determined by the relative abundances of light and heavy isotopic clusters. The key difference between data produced from this type of experiment and the data produced from a metabolic labeling experiment is that comparative proteomic experiments produce two flux-independent isotopic distributions, while peptide-based flux analysis produces one flux-dependent isotopic cluster. Figure 2.1 illustrates the differences between this type of data. In the first case, a flux-dependent MID is observed for a 4 carbon species. As described in Section 1.5, case (1) cannot be accurately simulated unless a flux distribution is known. However, in case (2), both MIDs are flux independent because each isotopologue population is uniformly labeled. In case (2), the "labeling" illustrated is *in vivo*, but the principles can be extended to *in vitro* labeling as well.

In most quantitative proteomics experiments, the objective is to measure the *ratio* between heavy and light isotopologue populations. This is performed by taking the ratio between either the **monoisotopic peaks** (i.e. m+0 peak) or the **base-peaks** (i.e. most abundant peak) from each population. Extracting MIDs and accurately quantifying all isotopomer abundances for a group of isotopologues is often not necessary. Thus, software designed for quantitative proteomics to date is not sufficient for automated extraction of flux-dependent PMDs.

Figure 2.1: Example of difference between PMDs generated from a flux-dependent labeling experiment and a comparative quantitative proteomics experiments



2.3 Data Analysis: Approaches and Tools

Several software frameworks consisting of suites of algorithms for the analysis of mass spectrometry been developed over the last decade in response to the development of new mass spectrometry technology. Some software provide fully automated workflows for the identification and quantification of peptides using established techniques, such as MaxQuant, while others provide interfaces for the user defined workflows, such as XCMS/mzMatch (Smith et al., 2006), OpenMS (Reinert and Kohlbacher, 2010), Trans Proteome Pipeline (TPP) (Kohlbacher et al., 2007), and ProteoWizard (Kessner et al., 2008). Although these software suites are constantly growing to accompany current technology and techniques, established workflows for the automated extraction of non-natural peptide distributions have not been demonstrated within any of these frameworks. Although this specific goal has not been achieved, modular algorithms and paradigms for pre-processing LC-MS data have been developed in effort to allow for flexible software development for new techniques. As Listgarten and Emili (2005) state, most low level pre-processing algorithms have been "parenthetically performed within the larger goal of sequence-based identification", thus not rigorously tested or developed within the proteomics community. However, significant work as been done within the metabolomics community to identify flux-dependent isotopically

enriched metabolites.

In this section, a brief discussion of algorithms and software available for preprocessing LC-MS data are described, with an emphasis on the limitations of these approaches when attempting to extract and quantify PMDs from MFA experiments.

2.3.1 LC-MS Preprocessing

LC-MS preprocessing algorithms first attempt to distinguish noise from **peaks**, or pairs of m/z and intensity measurements that represent a single measurement within an extracted ion chromatogram (EIC). EIC's are then grouped into the full set of isotopologue measurements for a single molecule, or **features**. Algorithms used for identifying peaks and features have limited use for peptide based MFA PMDs. For example, the first step in most algorithms is to remove low-frequency noise from data sets by using one of many filtering approaches, including *Savitzky-Golay* and *lowess* smoothing algorithms. However, filtering techniques often have difficulty distinguishing between low abundant peaks and noise.

Filtering and smoothing is typically followed by peak identification. Zhang et al. (2009) recently reviewed approaches for LC-MS peak picking. Currently three approaches are taken to distinguish peaks from noise: (1) the shape of the extracted ion chromatogram, (2) the isotope pattern and (3) a combination of (1) and (2). To date, the most successful peak picking algorithms are based the isotope pattern of peptides rather than the shape of the ion chromatogram because of the discrete, sampling nature of peaks throughout time. Common isotope-matching peak picking algorithms such as *Superhirn* (Mueller et al., 2007) and *VIPER* (Monroe et al., 2007) cannot be used for the detection flux-dependent PMDs: the isotopic distribution is simulated using the uniformly-labeled "averagine model," and fitted to the measurement.

Approaches that solely rely on the shape of the ion chromatogram are not capable of extracting low abundant peptides. Furthermore, the poor performance of

chromatographic peak detection methods can be attributed to the assumption that elution profiles are Gaussian shaped, which is not necessarily true and depends on the saturation state of the stationary phase on the column (Rouessac and Rouessac, 2007). Some improvements have been developed within the metabolomics community recently to account for this by applying a continuous wavelet transoformation in the chromatographic domain (Tautenhahn et al., 2008), however have not been applied to high resolution proteomic datasets.

2.3.2 Software

Most of the software development developed for the automated extraction and visualization of isotope-labeled mass spectrometry data have come from within the metabolomics community. Currently, there are four software tools available for the extraction and quantification of labeled metabolites: MAVEN, MetExtract and mzMatch-Iso and the OpenMS tool *"FeatureFinderMetabo"*. MAVEN, developed in 2010, is a fully integrated framework for the analysis and visualization of mass spectrometry data, which uses a neural network to detect and group peaks based on a set of metrics developed from manually extracted data (Melamud et al., 2010). However, the metrics used in the algorithm are based on manually annotated peaks from small metabolite MIDs rather than peptide MIDs.

Bueschl et al. (2012) developed MetExtract for a specific subset of isotopically enriched LC-MS datasets, requiring the presence of isotopologues in the fully unlabeled and full labeled states, which is not typically observed for peptide-based MFA. Chokkathukalam et al. (2013) recently developed an R-based tool for the annotation and quantification of isotope labeled mass spectrometry data, mzMatch-ISO, however require multiple biological samples for optimal performance. OpenMS algorithm "FeatureFindeterMetabo" is a flexible algorithm that attempts to identify metabolites with arbitrary isotopic enrichment, however was not capable of identify-

ing isotopically enriched PMDs. After parameters had been optimized, the algorithm annotated single PMDs as multiple molecular species resulting in subsets of identified PMDs. For example, mass isotopomers m+0 and m+1 would be grouped as a single peptide species, while m+2 and m+3 would be classified as another.

All four software packages are limited in scope to metabolomics experiments and are not currently applicable for the detection of larger metabolites such as peptides. The following chapter describes a new approach to automatically extract and quantify PMDs that circumvents the algorithmic and software limitations of these software packages by compressing the elution characteristics of ion chromatograms and the mass accuracy and precision capabilities of the Orbitrap into a single data type.

Chapter 3

Methods: Software Design and Development

In this chapter, the methods developed for the automated extraction of peptide MIDs from labeled data sets are described. In Section 3.1, a complete workflow is provided, as well as a description of the training data set used for method development. Section 3.2 provides a summary for the working protocol used to identify peptide species from unlabeled data sets, specifically focusing on additional data provided via algorithms used in MaxQuant for m/z recalibration. Section 3.3 describes the working protocol developed for the reduction of data using the open source software suite OpenMS, including methods developed to optimize software parameters. Section 3.4 describes the development of in-house MATLAB modules that filters and groups mass traces into peptide spectral features, including a discussion of parameter optimization and tradeoffs associated with parameter choices. Section 3.5 describes novel algorithms developed for quantification and error estimation of peptide MIDs for high resolution, high-throughput mass spectrometry.

3.1 Experimental Design

3.1.1 Workflow

The quantification of flux-dependent isotopically enriched peptide MIDs requires an **unlabeled** sample as well as the **labeled** sample (see Figure 3.1). Both data sets are generated from samples subjected to identical protein preparation, LC-MS protocols and culturing conditions, differing only in isotopic enrichment of the substrate in the culture media. The unlabeled sample is used to determine the set of proteins and peptides one could observe under the specific culturing conditions using standard identification software suites. The number of peptides observed within an unlabeled sample is maximized in a two-step search protocol described in Section 3.2 using MaxQuant/Andromeda (Cox and Mann (2008), Cox et al. (2011)). Furthermore, each peptide is annotated with an observed m/z, corrected for instrument m/z bias or inaccuracy. The labeled sample is processed with the open source software OpenMS to reduce the raw data into a groups of m/z measurements observed in time. The peptide database and mass traces are imported into MATLAB for the identification of peptide features using a hierarchical clustering algorithm. This is based on the construction of a **similarity matrix** for groups of mass traces that integrates isotopic accuracy information with the covariance of ion chromatograms. MIDs are then quantified using a constrained non-linear optimization technique that finds the optimal MID and standard errors for a peptide.

3.1.2 Training Set

In order to develop a suitable workflow for unpredictable peptide MIDs from labeling experiments, we used a training set consisting of 17 peptides identified from labeled and unlabeled Soybean, *Glycine max* cv. Jack. Embryo culture conditions, protein

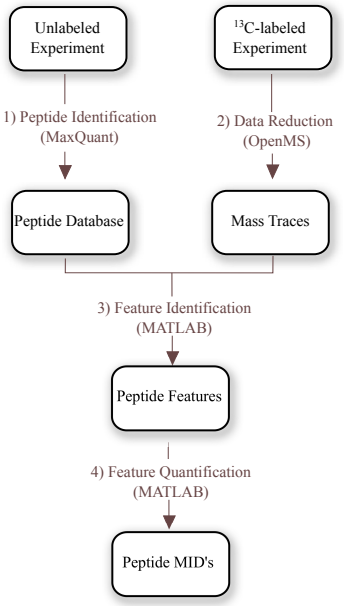


Figure 3.1: Peptide MID Quantification Workflow

preparation and LC-MS protocol are described in (Mandy et al., 2013) (submitted, in review). All samples were run on a LTQ-Orbitrap Velos mass spectrometer in positive ion mode within an m/z range of 50 - 2000 Da, with a resolution of 60,000 at 400 m/z and with an automatic gain control set to 1,000,000 charges. Peptides of the proteins glycinin and conglycinin were identified from unlabeled samples using Mascot Distiller v2.4 and searched using Mascot Daemon (Matrix science, London, U.K.). Identification of peptides was based upon NCBI library and a trypsin cleavage pattern, 0.80 Da and 15 ppm fragment and parent tolerances, respectively. Protein modifications inlcuded were carbamidomethylation of cysteine, oxidation of methionine and deamination of glutamine and asparagine. m/z measurements for the m+0 isotopomer were recorded and used as a reference value to search for that peptide within labeled data. Table 3.1 contains a list of all manually identified peptides, charge states, theoretical and reference m/z measurements, mass error, full-width at half-height (FWHH) spanning retention times, number of mass isotopomers and the

	charge	theoretical m/z	reference m/z	mass error (ppm)	FWHH(s)	isotopomers	sample rate
LLK	2	187.14	187.15	6.55	19.14	10	1
DYR	2	227.11	227.11	6.61	12.12	12	0.98
VLFGR	2	296.18	296.19	2.54	9.85	12	0.97
NKNPFHFNSK	4	308.91	308.91	3.33	11.52	18	1
VLFSR	2	311.19	311.19	3.14	13.96	13	0.97
NPFHFNSK	3	330.83	330.83	2.75	9.03	13	0.97
SRDPIYSNK	3	360.52	360.52	6.18	12.48	29	0.98
NFLAGSK	2	368.70	368.70	2.52	16.41	9	0.98
FEEINK	2	390.20	390.20	3.72	7.42	24	0.96
FQTLFK	2	392.22	392.22	3.19	4.93	19	0.99
NKNPFHFNSK	3	411.54	411.55	3.39	12.34	20	0.97
DIENLIK	2	422.74	422.74	3.14	9.02	28	0.98
LQSGDALR	2	430.24	430.24	3.61	9.91	25	0.95
NKNPFLFGSNR	3	431.89	431.90	3.43	17.23	23	0.70
FEEINKVLFGR	3	451.25	451.25	3.13	7.38	36	0.98
LQESVIVEISKK	3	458.27	458.28	4.68	14.77	47	0.95
FEEINKVLFSR	3	461.25	461.26	3.46	5.74	34	0.92

Table 3.1: Training set

sample rate for all expected measurements.

The remaining sections of the chapter aim to comprehensively describe the methods and development of the automated workflow via discussion of three key parts:

- Description of the procedure from a pragmatic and algorithmic motivation
- Discuss key parameters required each method or algorithm
- Discuss characteristics from the training set used to develop the method

3.2 Peptide Identification

The workflow requires an unlabeled sample to first identify peptides sequences and charge states. A **targeted search** is performed using these peptides species to find peptides within labeled samples based on observed m/z measurements. An **un-targeted search** would use either MS2 data from labeled samples, or attempt to identify peptides solely based on m/z. Currently, there are no proven software solutions for un-targeted identification of isotopically enriched peptides. The XCMS extension, mzMatch-ISO is currently the only software package capable of un-targeted

identification of labeled metabolites. However, we were unable to accurately identify and quantify any of the peptides within the training set when using mzMatch-ISO.

Procedure

The computational proteomics software suite, MaxQuant v 1.3.0.5 (Max Planck Institute of Biochemistry, Martinsried, Germany) was used to identify peptides within an unlabeled sample. The unlabeled sample was first loaded into the MaxQuant GUI hosted by the Minnesota Supercomputing Institute (MSI). The reference proteome for *Glycine hispidia* was used for all searches and was obtained from UniProt (UniProt Consortium). The database was loaded into Andromeda and configured according to the supplied protocol. Within the group specific parameters, the variable modifications included were carbamidomethylation of cysteine, oxidation of methionine and deamination of glutamine and asparagine. The digestion enzyme trypsin was specified with a multiplicity of 1, and a mass accuracy for the first and second peptide searches was set to 20 and 6ppm, respectively. A maximum of 5 modifications per peptide were allowed, and peptides contained a maximum of 4 missed cleavage sites. Peptides were allowed to carry +1 to +7 charge units. MS/MS fragmentation parameters were left to default, while identification and quantification parameters were set manually. The minimum length of a peptide used for identification was 3 amino acids. Peptide and protein false discovery rates were set to 0.01. One positively identified peptide was required for the protein identification. However, this peptide was not required to be unique. This was done to ensure maximal coverage and identification for m/z recalibration. All other parameters were set to default. MaxQuant produces a directory containing multiple output files. The directory contains information integrated by the MaxQuant parser MATLAB module to create a protein data structure described in Appendix A. The protein data-structure was then used to generate a new FASTA file, but only with the subset of proteins previously

identified. The MaxQuant identification protocol is then performed using the new FASTA database, resulting in a new protein data structure with a higher number of unique peptides per identified protein. This new protein data structure is used as an input into the feature identification MATLAB module described in Section 3.4.

MaxQuant identifies more peptides than Sequest or Mascot

Peptide identification rates using MaxQuant, Sequest and Mascot were compared by identifying peptides from the unlabeled soy sample. All parameter choices for Sequest were identical to those used for MaxQuant searches. MaxQuant outperformed both Sequest and Mascot by identifiying significantly more peptides within the sample; the number of peptides identified increased by 75% when using MaxQuant compared to Sequest (MaxQuant/Andromeda: 230, Sequest: 131, Mascot: 29). Furthermore, the probability associated with the identification of peptides did not suffer: the mean posterior error probability (PEP) for the 230 peptides was 0.0113 ± 0.0174, indicating that 99% of the measurements have atleast a 95% probability of being correctly being annotated, compared to 131 for Sequest.

Two-step search results in a larger number of unique peptides

Using the full UniProt FASTA file, 154 peptides were identified using MaxQuant while after the FASTA file was reduced 230 peptides were identified. To analyze the quality of identifications, we looked at how the PEP for peptides identified in both the first and second stages changed by taking the ratio of the PEP of the first identification and the second. We identified 151 peptides identified in both stages and observed an average decrease in PEP by three orders of magnitude, indicating that by limiting the possible set of peptides in our search database, the probability that the identified peptide was correctly annotated increases.

Potential uses for *m/z* recalibration data

MaxQuant was chosen because of its ability to rescale the m/z domain by using identified charge state pairs within a sample. Precursor ions in the MS1 spectra are first matched to sequenced peptides from MS2 data. MS1 spectra for peptides from different charge states are then used to rescale the raw, MS1 spectra. The identification algorithm provides adjusted scores for identfieid peptides based on the rescaled MS1 spectra.

If m/z deviation is a machine-dependent phenomena, we would expect that m/z calibration could be described as a function of the m/z measurement. The calibration function could then be used to transform labeled datasets into datasets with more accurate masses. To investigate this, we looked at the difference between uncalibrated m/z measurements and calibrated m/z measurements in both the time-invariant m/z domain and the m/z-invariant time domain, respectively. MaxQuant was able to match 692 MS2 spectra to a set of redundant peptide species, with 434 being used for m/z recalibration. Figure 3.2 shows the relationship between uncalibrated m/z measurements and calibrated m/z measures as a function of m/z (left panel) and as a function of retention time (right panel). The figure indicates that m/z calibration is primarily a function of mass drift in time rather than mass bias within the m/z domain.

Currently, we do not attempt to use this information to transform data from labeled samples because the m/z recalibration function is sample dependent, rather than simply instrument dependent. If error associated with m/z measurements were sample independent, then the non-linear m/z recalibration function would be similar for multiple samples. Figure 3.2 shows m/z calibration data for two separate unlabeled samples run on the same instrument. This shows that the relationship between m/z or retention time and m/z calibration is not necessarily reproducible. Therefore, currently the main advantage of using MaxQuant is identification of more peptides

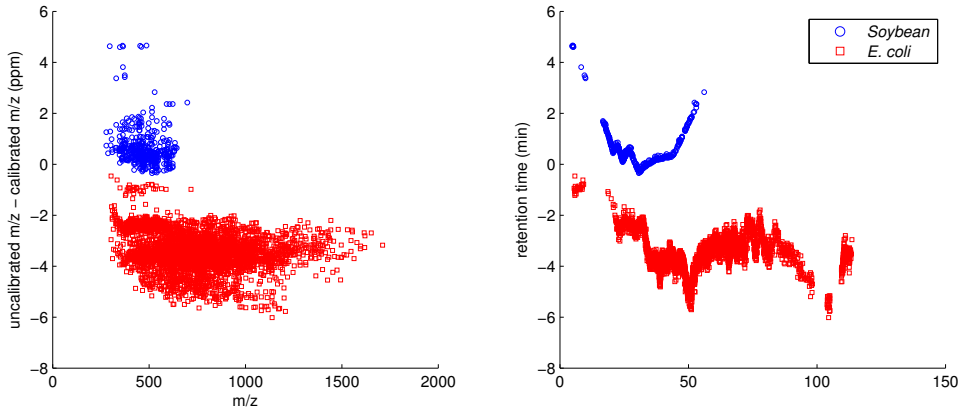


Figure 3.2: Comparing m/z recalibration for multiple samples

in the unlabeled sample.

3.3 Reduction of Labeled Data to Mass Traces

In this section, the reduction of raw data from labeled samples into **mass traces** is described using a workflow developed with The OpenMS Proteomic Pipeline (TOPP). The section first focuses on the workflow implemented followed by a discussion of key parameters within the workflow. The optimization of parameters based on the training set is then used to develop a scheme to determine optimal mass trace extraction parameters.

A **mass trace** is defined as a set of similar m/z measurements observed within spanning retention time at high frequency. The isotopologue abundance as a function of time for a mass trace is equivalent to an extracted ion chromatogram (EIC) if the mass trace contains all m/z measurements associated with an ion species. However, it is often the case that the true EIC is made up of several mass traces. It is the goal of this work to find the parameter set that maximizes the likelihood a mass trace contains all measurements within an ion chromatogram spanning FWHH retention

time.

Method

The raw data in mzXML format was then loaded into the TOPPAS environment and converted to mzML format. Peak reduction was performed using a S/N cutoff of 1, using the "PeakPickerHiRes" algorithm. The peaks were then reduced to mass traces using the "Mass Trace Extractor" algorithm. Table 3.2 lists the optimized values, default values and descriptions of all parameters used with the OpenMS mass trace extraction tool.

Parameter optimization

Parameter values were adjusted to maximize the similarity between any pair of mass traces within an isotopic cluster. For a set of mass traces within an isotopic cluster, the similarity can by captured by determining the degree of overlap of mass trace measurements in time. We can use the *Jaccard similarity* statistic, $J(A, B)$, to determine the degree of measurement overlap between mass trace A and mass trace B such that

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

where A and B are scan sets for mass trace A and B, respectively. Within the context of LCMS data, the jaccard similarity statistic measures the number of shared scans between two mass traces as a fraction of number of total scan events for both mass traces. If scan event sets for mass trace A and mass trace B were equal, then their jaccard similarity would be 1. If scan event sets for mass trace A and mass trace B were disjoint, then their jaccard similarity would be 0.

The jaccard similarity index for scan sets captures multiple characteristics of mass

Parameter	Current(Default)	Description
noise threshold	500(10)	Intensity threshold below which peaks are regarded as noise
chromatogram peak S/N	1(3)	Minimum S/N ratio a mass trace should have
chromatogram FWHM	10(5)	Expected chromatographic peak width (seconds)
mass error (ppm)	4(20)	Allowed mass deviation
reestimate mass trace SD	true(true)	Enables dynamic re-estimation of m/z variance during mass trace collection
minimum sample rate	0.5(0.5)	Minimum fraction of scans along the mass trace that must contain a peak
minimum trace length	0.8(5)	Minimum expected length of mass trace (in seconds)
width filtering	off(off)	Enable filtering of unlikely peak widths
min FWHH	0.15(3)	Minimum full-width-at-half-maximum of chromatographic peaks
max FWHH	65(60)	Maximum full-width-at-half-maximum of chromatographic peaks
S/N filtering	false (false)	Apply post-filtering by signal-to-noise ratio after smoothing
maximum trace length	850 (300)	Maximum length of a mass trace (in seconds)
enabled	false(true)	Enables/disables the chromatographic peak detection of mass traces

Table 3.2: Mass trace extraction parameter list

traces, including the tradeoffs between sample-rate and mass consistency. For example, if the sample-rate or the mass consistency parameter is too large, two distinct ion chromatograms with similar m/z values can be artificially connected into one mass trace. If we were to compare the scan event sets for this combined mass trace with a true mass trace from another isotopomer, we would see a decrease in similarity. Additionally, if the sample rate is too low then real mass traces will be truncated into multiple mass traces. If the mass consistency constraint is too low, gaps will be introduced within mass traces, which reduces the sample rate.

To explore these trade-offs, we built mass traces from extracted ion chromatograms at FWHH for all peptides in the training set at different mass consistency cutoffs. First, the S/N-weighted mean for each mass trace was calculated using measurements within FWHH. Measurements that fell within the mass consistency cutoff for each scan event were identified. Scan events that had a measurement were then grouped into sets based on a specified maximal gap-size, which indirectly represents the sample rate parameter. All scan event sets that contained the scan events at FWHH were then assigned as the true mass trace scan set.

The jaccard similarity statistic was then calculated for each mass trace combination within a peptide. The collection of jaccard statistics were summed over all combinations, and subsequently summed for all peptides. This produced one summed jaccard statistic, which was plotted for different mass consistency and gap-size cutoffs in Figure 3.3. It was observed that the maximum similarity between mass traces within the training set with a 4 ppm mass consistency parameter a gap-size of 0. However, at this mass consistency parameter cutoff and gap-size, 13% of the mass traces were not identified in our sample (47/372). This was because these mass traces contained very few measurements. Therefore, increasing the sample rate paramater results in more mass traces extracted, albeit at the expense of isotopic cluster similarity. We chose to use a sample-rate cutoff of 0.5 to ensure the collection of sparse

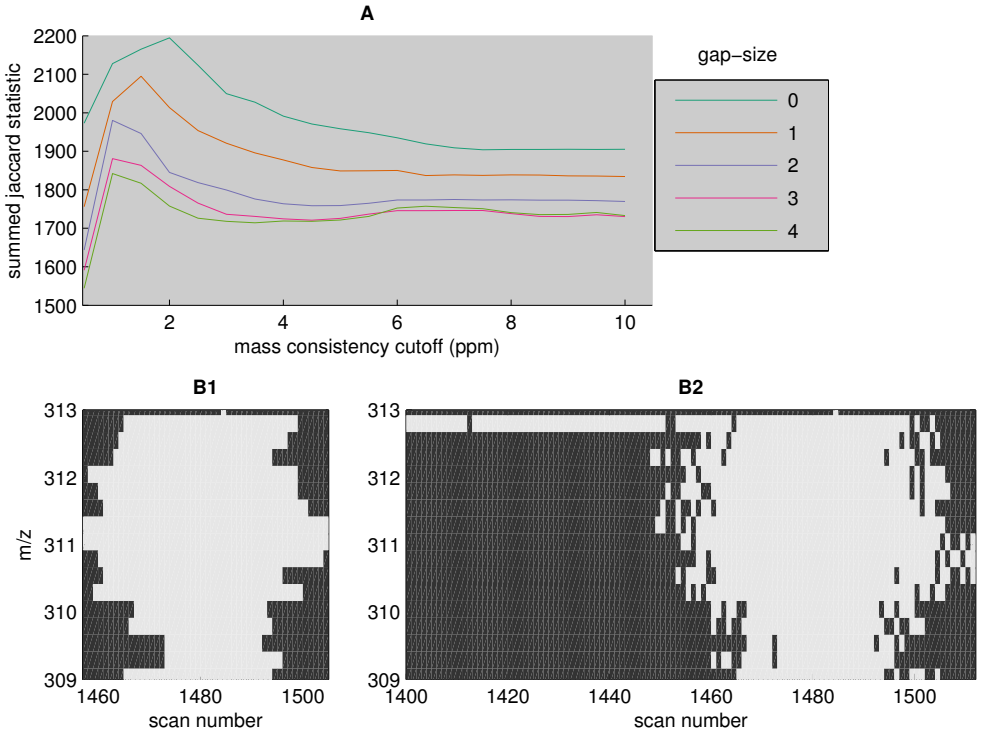


Figure 3.3: Relationship between m/z consistency and sample rate: Panel **A**: mass consistency cutoff vs summed jaccard similarity. Each line represents a set of mass traces constructed using a different maximum gapsize cutoff. Panel **B**: two constructed mass traces for NKNPFHFNSK using an m/z consistency cutoff and gapsize cutoff of (**B1**) 4 ppm and 0 and (**B2**) 4 ppm and 2, respectively

mass traces while retaining a significantly high similarity.

The remaining parameters were directly calculated using the mass traces generated using the optimal mass consistency parameter and sample gap-size. 372 mass traces were extracted, resulting in a total of 34,564 individual measurements. The range of mass trace lengths were 0.82 - 838.19 seconds, with a mean of 77.65 and median of 27.54 s. The minimum intensity value was 524, and the minimum S/N ratio was 1.06.

For 22 of the 372 extracted mass traces, we coud not calculate FWHH ion chromatogram estimates because measurement observance was either too sparse in fre-

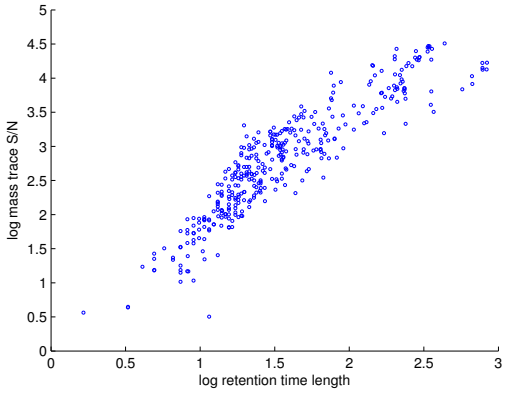


Figure 3.4: Relationship between m/z retention time length and ion abundance

quency or too low in abundance. Of the 350 mass traces, a mean FWHH of 9.70(SD: 9.21s) was observed with a full range of $0.15 - 63.5$s.

Lastly, it was observed that there is a strong relationship between mass trace retention time length and ion abundance (see Figure 3.4). The large range of full peak widths and peak widths at FWHH reflect the large range in abundance for each isotopomer. For example, in Figure 3.4 the largest mass trace is over 5 orders of magnitude larger than the smallest mass trace. To ensure the inclusion of mass traces with low abundance, we chose to disable peak width filtering within mass trace extraction. Limited peak-shape information is available in low abundant mass traces because the isotopologue abundance is dominated by sampling probability. Thus, it is difficult to include such contraints in the mass extraction algorithm without incorporating probabilistic information.

3.4 Feature Identification

Once mass traces were properly constructed and peptide information was obtained from MaxQuant, mass traces needed to be assigned to their appropriate mass isotopomer within an isotopic cluster. The collection of mass traces that belong to

one molecular species is defined as a **feature**. This section describes the algorithm developed for identification of features for isotopically enriched peptides. Unlike algorithms described in Section 2.3.1, our method requires minimal information of isotopic distributions.

Depending on the conditions of the labeling experiment, mass isotopomers typically observed under natural conditions might fall below the limit of detection. This might result in discontinuous mass states or light mass states that are not observable. Therefore, it was our goal to design an algorithm that is flexible enough to identify peptide features consisting of random MIDs. The workflow and algorithms described in this section are encapsulated in the feature identification MATLAB module. Section 3.4.1 describes the reduction of potential mass traces that could belong to a particular peptide species using reference m/z and retention time values, while Section 3.4.2 describes the algorithm designed to group mass traces into peptide features.

3.4.1 Filtering Raw Data

Before mass traces were clustered for a particular peptide species, the full mass trace set was reduced to a set of mass traces that fell within a **mass accuracy cutoff** (MAC) and a **retention time window cutoff** (RTW). For a given peptide species, m/z values were simulated for all possible ^{13}C mass isotopomers using the reference m/z provided by MaxQuant. Mass traces previously characterized were used to calculate an appropriate MAC paramater and RTW parameter. The mass accuracy parameter was determined by comparing the S/N-weighted average m/z for a mass trace to the simulated m/z. For the 325 mass traces previously discussed, a mass difference of as much as 60 ppm was observed. However, the majority of mass traces fell below 12 ppm. The maximum difference between observed retention time in the unlabeled sample and retention time in the labeled sample was 1.75 minutes. Default values for the filtering algorithm is currently set at MAC of 20 ppm and RTW of ± 2

minutes.

3.4.2 Clustering Mass Traces

Mass traces were clustered based on the assumptions that mass traces belonging to the same isotopic cluster should contain a degree of similarity in time and mass. This section describes the procedure developed to integrate both types of data to form **similarity vectors** for each mass trace.

Signal Information

A **signal matrix**, X, was constructed for the set of filtered mass traces, $\mathcal{M}$ consisting of M mass traces. Suppose that $\mathcal{S}_i$ represents the scan event set for the i^{th} filtered mass trace in $\mathcal{M}$,

$$\{s_{i,1}, \ldots, s_{i,j}, \ldots, s_{i,N_i}\} \in \mathcal{S}_i$$

The union of all scan event sets were then taken, such that $\mathcal{S}_t = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_M$. Thus, $\mathcal{S}_t$ contains all scan events observed within the collection of mass traces, where

$$\{s_1, \ldots, s_{i,k}, \ldots, s_{i,N}\} \in \mathcal{S}_t$$

with N representing the total number of scans observed in $\mathcal{S}_t$.

The S/N ratio $x_{i,k}$ was the calculated for each mass trace i and scan event $s_{i,j}$. An M by N matrix, X, was constructed with elements $x_{i,k}$. Note that if mass trace i contained no observable signal at scan event k, $x_{i,k}$ was set to zero. Rows and columns in X represent mass traces and independent scan events, respectively.

The sample covariance matrix, $C = \mathrm{Cov}\,(X^T)$, was calculated to exploit the observation that mass traces that belong within the same isotopic cluster share similar elution profiles. In other words, peaks drawn from the same isotopologue population should have peak shapes that correlate with one another in time. This is expected

when the RIA for any mass trace is independent of when it comes off the chromatography column (see Appendix D). C contains elements $c_{i,j}$ that reflect the covariance between mass trace i and mass trace j. The covariance was then transformed into the correlation matrix, R, by element-wise division by the dyadic of mass trace standard deviations.

m/z Information

A m/z weighting matrix, W (with elements $w_{i,j}$), was constructed to capture the **isotopic precision** between mass traces that belong within the same isotopic cluster. Although the mass accuracy of a single ion species measured in different samples can vary up to 60 ppm, the difference between measured m/z values and theoretical m/z values of ions coming from the same group of isotopologues should be consistent. This is because for all pairs of adjacent ^{13}C isotopomers (e.g. m+0 and m+1), their mass is distinctly separated by 1.003355 amu. Thus, we can simulate theoretical m/z values for all mass isotopomers based on a reference m/z from any isotopomer observed in a separate sample, and subsequently measure the difference from the observed mass trace to the closest simulated m/z value.

First, the S/N-weighted m/z value for each mass trace was calculated for each mass trace. For each m/z value, m_i, the **isotopic accuracy**, d_i, was calculated by taking the signed difference from the closest simulated isotopomer, $m_{k,\text{sim}}$ such that

$$d_i = \frac{m_i - m_{k,\text{sim}}}{m_{k,\text{sim}}} \times 10^6$$

The **isotopic precision**, $\Delta_{i,j}$, between mass trace i and j was calculated by taking the absolute difference between respective isotopic accuracies, where $\Delta_{i,j} = |d_i - d_j|$. The isotopic precision was then mapped to values between 0 and 1, using the following gaussian transformation:

$$w_{i,j} = e^{\frac{-(\Delta_{i,j})^2}{2\sigma^2}} \tag{3.2}$$

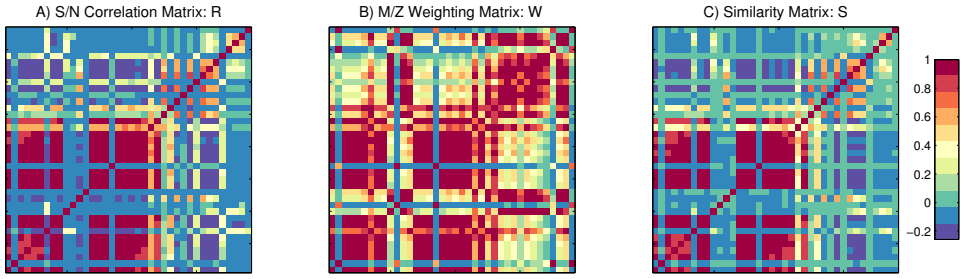


Figure 3.5: Comparison between different similarity matricies: A) S/N correlation matrix, B) m/z weighting matrix and C) similarity matrix

where σ is a function of the difference in ppm one would expect at half similarity, such that $\sigma = {}^{\Delta_{i,j,1/2}}/\sqrt{2ln2}$. This functional mapping is called **isotopic similarity**. Thus, $\Delta_{i,j,1/2}$ can be parameterized based on observed isotopic similarity between mass traces belonging to the same peptide.

Similarity Matrix Construction

A similarity matrix, Q, was constructed to integrate both the signal information and m/z information into a single rank-deficient matrix to be used for clustering. The goal of this transformation was to preserve both both high positive and negative correlation between mass traces while reducing correlation between mass traces with high signal correlation, but low isotopic similarity. This was achieved by constructing elements of Q, $q_{i,j}$, using the following formula:

$$q_{i,j} = \begin{cases} r_{i,j} & r_{i,j} \leq 0 \\ r_{i,j} \times w_{i,j} & r_{i,j} > 0 \end{cases}$$

By requiring that the correlation is positive when multiplied by the weighting factor, we ensure that negative correlations between mass traces with low isotopic similarity are not lost. Figure 3.5 compares R, W and Q. Rows of Q are defined as **similarity vectors** for each mass trace.

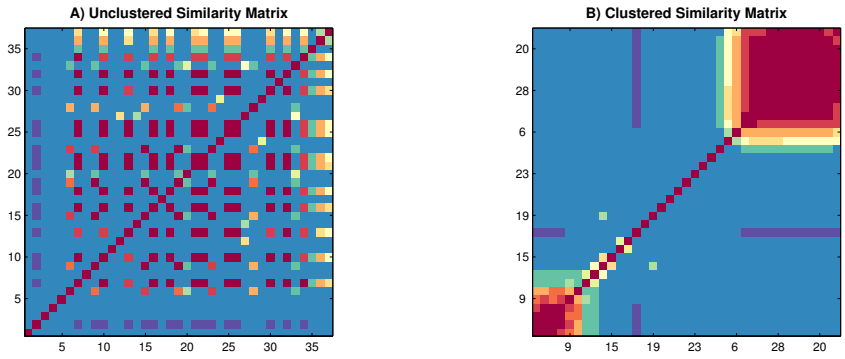


Figure 3.6: Comparison between unclustered and clustered similarity matricies: The columns and rows are rearranged via the optimal leaf ordering acccording to a clustered dendrogram

Mass Trace Clustering

To group mass traces with high similarity, an allogomerative hierarchal clustering algorithm was developed. Similarity vectors were clustered using Ward's minimum variance method with a Euclidean distance metric. This was the appropriate method and distance metric for cluster analysis because vectors associated with mass traces that belong within the same isotopic cluster should have minimal variance. At each stage in the method, a mass trace is selected from the remaining set of mass traces such that the inter-cluster variance between information vectors is minimal. This ensures that each mass trace appended to the growing cluster belongs within the cluster more than the remaining mass traces.

The **Linkage** utility in MATLAB was used to perform clustering on the similarity matrix, which produces a matrix encoding the information needed to construct the dendrogram. An algorithm was developed to identify each node in the dendrogram as a potential feature. Features were built from the bottom of the dendrogram by first identifying the node with least variance. Parent nodes were identified as potential features iteratively until a group contained two of the same type mass isotopomer. For example, if a child node contained potential m+0 and m+1 mass isotopomers and its

parent node contained another potential m+0, then the child node would be a feature. The feature is removed from the dendrogram and the algorithm is repeated on the remaining node set. This was performed until the dendrogram had been completely segregated into features. Figure 3.8 shows an example of an identified feature that represents VLFSR in the training set. Note that mass isotopomers are represented exactly once per feature and multiple features can be extracted from $\mathcal{M}$. The mass trace set that corresponds to a true feature is denoted $\mathcal{F}$.

There are scenarios when the inclusion of isotopomers of the same type is not a sufficient stopping criteria for feature identification. For example, suppose $\mathcal{M}$ consisted of three mass traces, two of which are highly correlated and are potential m+0 and m+1 isotopomers, and a third less correlated mass trace that is potentially an m+100 isotopomer. The m+0 and m+1 mass traces would be grouped first, followed by the m+100. However, it is unlikely that these mass isotopomers belong to the same peptide because of the large number of non-observed mass isotopomers within the feature. To address this issue, another parameter wused as a stopping criteria for the algorithm is the **dendrogram height**, H, which is a function of the variance within a mass trace cluster. To determine how this information could be used as stopping criteria, the relationship between the clustering of mass traces within a true feature and the dendrogram height was established and is shown in Figure 3.7.

First, true mass traces that belonged to the feature were identified and tracked through the clustering algorithm, where the x-axis in Figure 3.7 represents the algorithm stage. Clustering began on the node containing the two most correlated mass traces. In subsequent stages of the algorithm, the cluster node was assigned to the parent node for the previous cluster. This was performed until there existed no other parent nodes. At each stage in the algorithm, three metrics were calculated: 1) the absolute dendrogram height (green line; right axis), 2) the percent of true mass traces within the cluster (blue line; left axis) and 3) the percent of unique isotopomers within

the cluster (red line; left axis).

It can be seen that in all cases, the stage of the algorithm that results in the incorporation of isotopomers of the same type (represented by a drop in the red line) results in a sharp increase in H, indicating that the *change* in dendrogram height, or ΔH, at each stage in the algorithm might serve as a sufficient additional cutoff. The optimal ΔH could be inferred visually from Figure 3.7 by finding a maximum ΔH that occurs prior the incorporation of all true isotopes (represented as the sharpest increase in the green line prior to the blue line reaching 100%), which was roughly 2.5.

After candidate features are identified using this algorithm, we can impose additional constraints on candidate features based on observations from the training set. For example, it is rare that there is a high degree of **isotopomer discontinuity** within a feature. Therefore, the software segregates identified features based on a maximum isotopomer discontinuity prior to annotation. As default, the software requires a feature to consist of contiguous isotopomer species.

Feature Annotation

After the feature finding algorithm segregates the hierarchical cluster into potential features, the feature that most likely belongs to the molecule of interest needs to be identified. Currently, more information is required to determine which feature likely belongs to the peptide of interest, thus the software requires that a unique feature be identified per peptide.

In order to maximize recovery, two approaches were developed: the first allows incorporation of a parameter associated with the probability of observing a user-defined mass isotopomer, while the second calculates an expected isotopomer from the initial set of identified features. While the first approach relies on prior knowledge of the experiment, the second is an unsupervised approach to characterize isotopic

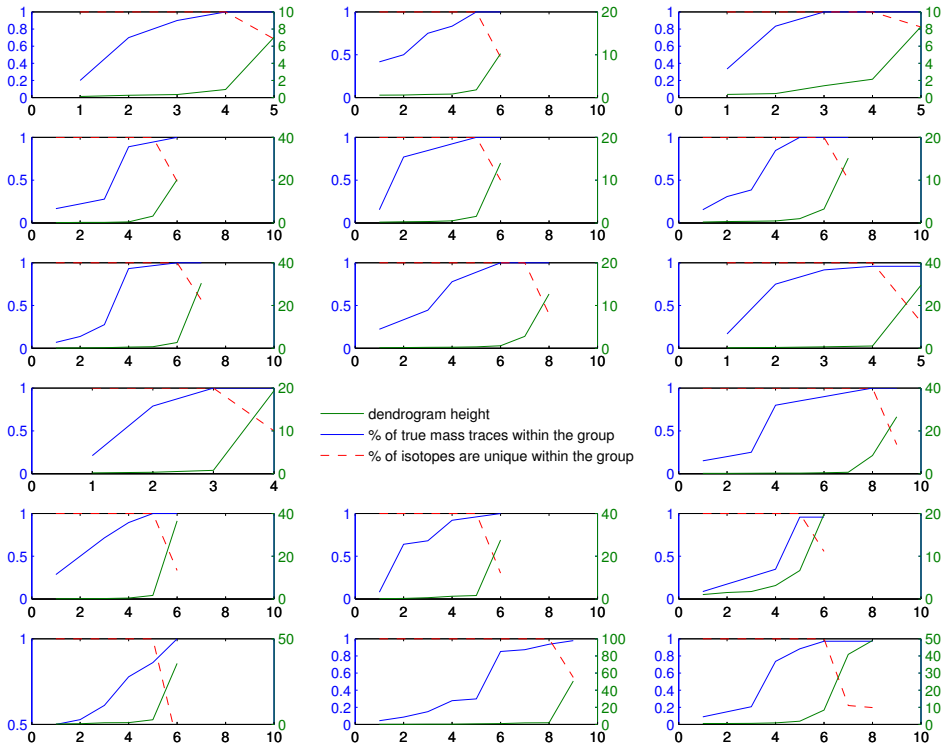


Figure 3.7: Analysis of potential stopping criteria for feature identification algorithm:

distributions within the sample.

In most experiments, a simple rule is sufficient for distinguishing the correct feature from other candidate features. For example, in most labeling experiments there exists a population of isotopologues generated from unlabeled substrate prior to growth with labeled media. This means there will be a significant number of isotopologues in lower mass states. We would then require that features must include low mass isotopomers, such as a m+0 or m+1. If prior knowledge is known for which isotopomers should be above the limit of detection, then a score can be assigned to clusters based on the isotopomer index, ι_i, for the i^{th} mass trace in the feature set $\mathcal{F}$. A probability-weighting function, $w(\iota_i)$ can be generated based on the experimental design and an annotation score can formulated, such that

$$Z_{\text{annotation}} = \sum_{i \in \mathcal{F}} w(\iota_i)$$

Careful experimental design aimed at maximizing the annotation score of true features is discussed in Chapter 5.

In the second approach, calculated RIAs for peptides with unique features are used to calculate an the average ^{13}C enrichment for the k^{th} peptide, defined as the **average carbon label**, l_k. The sample mean for all l_k is used to simulate MIDs for all peptides in the original database assuming uniformly labeled substrate. The maximum peak is determined and serves as an estimate for the **base peak** for that peptide. Candidate features are then re-filtered prior to annotation by removing all candidate features that do not include the base peak estimate. The power of this approach is demonstrated in Section 4.2 and Section 4.3.

Feature Reduction

After the feature has been annotated, the FWHH spanning retention time is calculated by summing all mass trace signals observed per scan event to generate a total

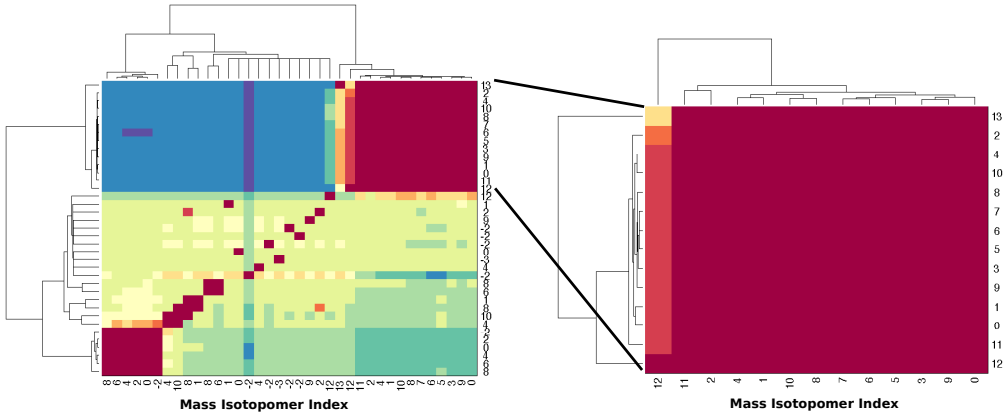


Figure 3.8: Feature Identification from Cluster Analysis

ion chromatogram (TIC). The maximum of the TIC is determined, and the scan event set spanning FWHH, $\mathcal{S}$, is calculated. The feature is then reduced into a **feature matrix**, F, consisting of elements $f_{i,j}$ which are are the number of isotopologues for mass trace $i \in \mathcal{F}$ and the j^{th} scan event in $\mathcal{S}$, using the following forumula:

$$f_{i,j} = \frac{S_{i,j}}{N_{i,j}} \frac{K\sqrt{\frac{R}{R_0}}}{z_i} \tag{3.3}$$

where $\frac{S_{i,j}}{N_{i,j}}$ is the signal-to-noise ratio, z_i is the charge state, R is the resolution setting and K and R_0 are instrument-specific parameters. Figure 3.9 provides an example of a feature matrix for the peptide NKNPFLFGSNR from the training set.

3.5 Feature Quantification

Once the feature matrix is determined, individual measurements are refiltered based on their isotopic accuracy. This is performed because of the variety of conditions that

could account for inaccuracies in m/z measurements, most importantly the presence of isobaric ions that cause shifts in the m/z domain (Kaufmann and Walker, 2012). Whatever the reason for m/z inaccuracy, requiring high isotopic precision minimizes the possibility that a contaminated peak bias the quantification of the MID. Figure 3.9 provides an example of a feature matrix after m/z measurements are filtered. In this figure, the grid is a 4 dimensional representation of the feature, where each rectangle in the grid represents a measurement for each mass isotopomer (y-axis) within a specific scan event (x-axis). Filled rectangles represent peaks observed at that particular mass and scan event combination, while empty rectangles indicate no peaks were observed. The size and color density of each rectangle correspond to the number of isotopologues and isotopic accuracy, respectively.

For this feature, 213 measurements were observed out of 299 expected measurements (71%) using an isotopic accuracy cutoff of 5 ppm. However, there does not exist a scan event that contains all observable mass isotopomers. This is problematic because it means the MID will be calculated from a set of incomplete sampling events.

Instead of throwing away scan events that do not contain all expected isotopomers, we chose to use a fitting approach with constrained nonlinear optimization MATLAB utility *fmincon*. The advantage of this method is that makes use of all measurements in the data set and can potentially resolve the relative abundances for isotopomers rarely observed.

3.5.1 Estimating MIDs From Experimental Data

In order to find the most representative MIDs, we fitted relative isotopic abundances for each observed isotopomer, p_i and the total number of ions samples in scan event s_j, η_j to observed ion count estimates, $f_{i,j}$. In the procedure, variables p_i and η_j were used simulate measured ion counts from each scan event. The fitting algorithm

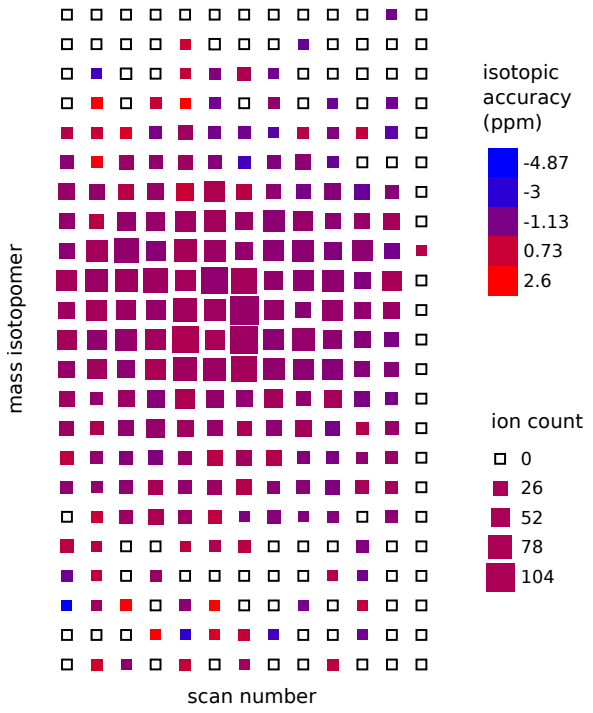


Figure 3.9: Example peptide from training set:

developed was:

$$\begin{aligned} \text{minimize} \quad & \sum_{i\in\mathcal{F}}\sum_{j\in\mathcal{S}} L_{i,j}|p_i\eta_j - f_{i,j}| \qquad \text{subject to} \\ & \sum_{i\in\mathcal{F}} p_i = 1 \\ & \eta_j \geq \sum_{i\in\mathcal{F}} f_{i,j} \quad \forall j \end{aligned} \tag{3.4}$$

where

$$L_{i,j} = \begin{cases} 0 & \text{if } f_{i,j} \text{ is } 0 \\ 1 & \text{otherwise} \end{cases}$$

$L_{i,j}$ was simply used as to eliminate the contribution of non-observed mass isotopomers to the residuum. The objective function was chosen to be of the ℓ_1-norm type because we wanted all residuals to contribute equally to the residuum. While the first contraint is a result of (B.1), the second constraint requires that the predicted

sample size per scan, η_j, be bounded below by the observed sample size. The fitting provided optimized mass isotope distribution probability estimates, p_i^*, and sample sizes per scan event, η_j^*. From these two optimized parameters, we can calculate expected values at each scan event such that

$$E(X_{i,j}) = \eta_j^* p_i^* \tag{3.5}$$

3.5.2 Estimation of Experimental Error

In addition to finding optimal RIAs to construct the MID, error estimates for RIAs were determined. If peptide MIDs are to be used for metabolic flux analysis (or any technique requiring fitting a set of parameters to MIDs), then error estimates are required for the calculation of parameter confidence intervals. This sections describes the methods developed to estimate errors for RIAs using a multinomial sampling model.

Each scan is a discrete sampling event of the total isotopologue population, where the sample size is directly related to the set of isotopologues eluting off the liquid chromatography column at that specific retention time. Although there was no evidence for systematic bias for later scan events enriched with heavy isotopes (see Appendix D), the absence of measurements described in Section 3.5 as well as intrinsic sampling variance described in Appendix B provided reason for developing a more sophisticated approach to estimate the error for each RIA.

We determined an estimate of variance for a particular isotopomer by fitting the multinomial standard deviation to an estimation of measured standard deviation using a multinomial sampling model. It was hypothesized that the relative frequency of each isotopomer from scan to scan varies as a function of multinomial sampling variance, m/z inaccuracy, and unknown experimental variance. The goal was to derive a relationship between the multinomial variance and the measured variance,

where we could use the theoretical error to predict experimental variance.

The sampling variance was first determined using the following equation:

$$\mathrm{Var}\left(X_i\right) = \sum_{j \in \mathcal{S}} L_{i,j}(f_{i,j} - E(X_{i,j}))^2 = \sum_{j \in \mathcal{S}} L_{i,j}(f_{i,j} - \eta_j^* p_i^*)^2 \tag{3.6}$$

From the experimentally determined variance, we then calculated the sample standard deviation $s(X_i) = \sqrt{\mathrm{Var}\left(X_i\right)}$. Because the sampling variance was calculated using events were no observations occurred, we investigated whether we could use the calculated multinomial sampling variance as a predictor for measured variance. Multinomial sampling variance, $\mathrm{mnVar}(X_i)$ was calculated using the fitted RIA and sample size, such that

$$\mathrm{mnVar}(X_i) = \left(\sum_{j \in \mathcal{S}} \eta_j^*\right) p_i^*(1 - p_i^*) \tag{3.7}$$

where the multinomial standard deviation, $\mathrm{mnSD}(X_{i,j})$ was calculated from multinomial standard deviation, such that

$$\mathrm{mnSD}(X_i) = \sqrt{\mathrm{mnVar}(X_i)}$$

For each peptide isotopomer in the training set, $s(X_i)$ and $\mathrm{mnSD}(X_i)$ were calculated and plotted in Figure 3.10. The figures shows a linear relationship between the standard deviation, with 91% of the variance in $s(X_i)$ explained by $\mathrm{mnSD}(X_i)$. Furthermore, the slope slightly greater than one indicates that observed variance is dominated by sampling variance rather than instrument noise. To calculate standard deviation estimate for a given measurement, we multiplied the calculated multinomial

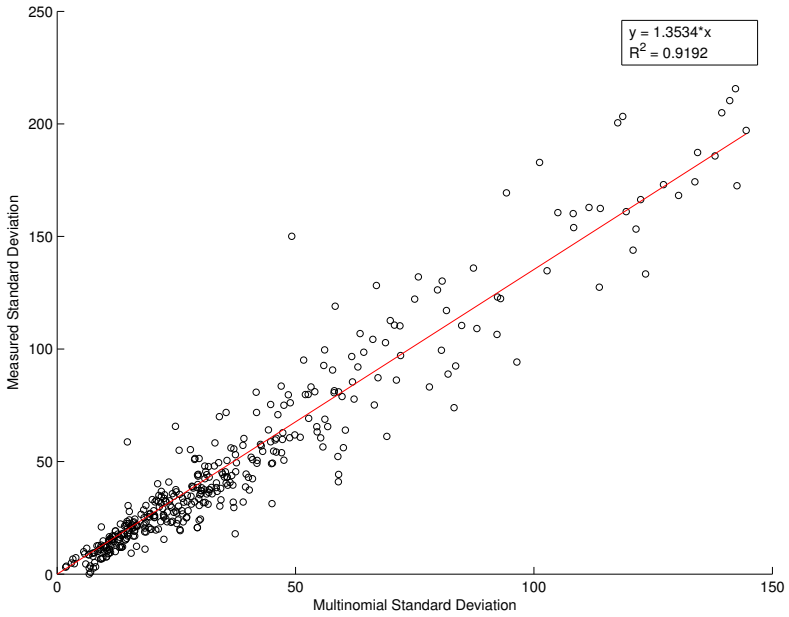


Figure 3.10: Using multinomial variance to predicted measurement variance

standard deviation by the slope of the regression plotted in Figure 3.10, β, such that:

$$s(X_i)_{\text{estimate}} = \beta \text{mnSD}(X_i) = \beta \sqrt{\left(\sum_{j \in \mathcal{S}} \eta_j^*\right) p_i^*(1 - p_i^*)} \tag{3.8}$$

where the estimated measurement variance would be:

$$\text{Var}\,(X_i)_{\text{estimate}} = \beta^2 \text{mnVar}(X_i) = \beta^2 \left(\sum_{j \in \mathcal{S}} \eta_j^*\right) p_i^*(1 - p_i^*) \tag{3.9}$$

Chapter 4

Software Validation and Implementation

In this chapter, validation experiments performed using datasets from different labeling experiments are discussed with an emphasis on the numerical and visual validation of the of the software. For each validation data set, parameter specifications to increase peptide recovery rates are discussed. Quantitative methods described in Section 3.5 are used to extract relative isotope abundances (RIA) for peptides and compared to simulated RIAs.

4.1 Unlabeled *E. coli*

Feature extraction and MID quantification was carried out using methods described in Chapter 3 for *E. coli* grown in glucose minimal media with no isotopic enrichment. *E. coli* strain G1655CGSC was cultured, protein preparation and MS data were collected according to Allen et al. (2013). Cells were grown with M9 media with 0.4% unlabeled glucose, total cellular protein was extracted and separated using SDS-Page. A protein band was extracted at 43KD and an in-gel digestion was performed; samples were reduced, alkylated and digested with trypsin. Orbitrap mass spectra were obtained and peptides were identified according to the procedure and parameter set listed in

Section 3.2. 37 proteins were identified in the sample with 10 or more peptides. Mass traces were constructed using the protocol discussed in Section 3.3 and clustered according to Section 3.4.2.

4.1.1 Analysis of Feature Identification and Extraction

To determine if the output of MaxQuant provided correct isotope assignments for precursor ions, the isotope index for the precursor ion was determined. m/z values were simulated for all ^{13}C isotopomers based on the modified sequence provided by MaxQuant. The precursor m/z value was mapped to the closest simulated m/z value, subsequently providing a matching isotope index. The matched isotope index was then compared to the isotope index provided by MaxQuant. Surprisingly, only 54% (349/648) of the peptides were assigned as the correct isotope. However, two distinct cases were observed for peptides assigned the incorrect isotope index: (1) observed monoisotopic masses provided by MaxQuant and masses calculated in-house were nearly identical, however the peptide had been assigned as an m+1 isotopomer instead of the correct m+0 isotopomer and (2) the masses provided by MaxQuant were roughly 57-58 amu higher than masses calculated in-house. Although there was no clear explanation for (1), the discrepancy in (2) was likely caused by an unrecognized carboxymethyl modification. This hyphothesis was corroborated with the observation that each peptide in case 2 contained atleast one cysteine. Roughly 2/3 of the isotopomer miss-assignment was explained by case (1) while approximately 1/3 was explained by case (2), indicating that 10% of peptide sequences identified by MaxQuant contain unrecognized modifications. Once precursor ions were assigned to their correct isotope index, peptides with undetected modifications were removed from the validation database.

Data Reduction

Once precursor ions were assigned to their correct isotope index, peptides with undetected modifications were removed from the database resulting in a set of 558 peptides for the 37 proteins originally identified. The protocol described in Section 3.3 was used to generate mass traces for the same unlabeled sample used for identification. Mass traces containing the precursor ion for each peptide were then identified. All precursor ions existed within a mass trace when using optimized parameters described in Section 3.3.

Feature Finding

Features were determined using the protocol described in Section 3.4.2. Mass traces were filtered by mass and retention time, and clustered using an isotopic precision of 4 ppm and a dendrogram height of 1.5. 10 potential features were identified on average per peptide. To assess the difference between feature clusters containing the true peptide and other features, we identified clusters that contained the precursor ion mass trace. Approximately 30% (168/558) of identified features containing the precursor ion mass trace consisted of a set of discontinuous isotopomers, with 15% (89/558) containing isotopomer gaps of at-least 15. If a quality score based on isotopomer continuity were employed, we would potentially lose up to 30% of the peptide features. Thus, clustered features were broken up prior to feature identification to prevent the existence of high isotopoic discontinuity as explained in Section 3.4.2. If no gaps were allowed, a small percentage (5%) of peptide feature sets contained only the precursor ion mass trace. Thus, RIAs were not calculated for theses "orphaned" features and removed from the peptide set for further analysis. Features were refiltered with a mass accuracy cutoff of 2 ppm and feature matrices were extracted for each peptide. Peptide MIDs were calculated using the procedure described in Section 3.5.

A small percentage of peptide features (13/558) contained a significant peak at the m−1 position, which suggests the peptide MID is contaminated with of another molecular species at that specific m/z and retention time. These peptides were removed, leaving 484 peptides that were used for validation of the quantification software.

4.1.2 Analysis of Quantification

Unlabeled peptide MIDs from *E. coli* were quantified in effort to validate the quantification module as well as determine the quantitative characteristics of Orbitrap data. Under natural conditions, it is assumed that the mean ^{13}C isotopic enrichment is 1.078%. However, it has been shown that isotope bias can occur in enzymes within central carbon metabolism, most notibly RuBisCO (Wong and Sackett, 1975). Therefore, we first attempted to calculate the mean ^{13}C isotopic enrichment given the measured distributions. First, the isotopic enrichments from non-carbon elements were removed from the measured distributions using deconvolution, resulting in carbon-only MIDs, $\mathbf{x}_{\text{carbon}}$ with elements $x_{i,\text{carbon}}$ representing the RIA of carbon mass isotopomer ι_i. The **average carbon label**, l_k, per peptide k was calculated using the following formula:

$$l_k = \frac{\sum_{i \in \mathcal{F}} \iota_{i,k} \times x_{k,i,\text{carbon}}}{C_k} \tag{4.1}$$

where C_k are the total number of carbons in the peptide k. The weighted mean of l_k for all peptides was determined by weighing each peptide by the total number

of observed ions, such that:

$$\bar{l} = \frac{\left(\sum_{j \in \mathcal{S}_k} \eta^*_{j,k}\right) l_k}{\sum_{k \in \mathcal{P}, j \in \mathcal{S}_k} \eta^*_{j,k}} \tag{4.2}$$

and $\bar{l}$ was compared to the theoretical ^{13}C enrichment of 1.078%. We found that the mean of the average carbon label for all identified peptides($n = 484$) was $1.03 \pm 0.09\%$, which was significantly smaller than the expected value of 1.078% ($p < 5\%$). The average carbon label deviation could be explained by an isotopomer bias at low mass states, which was further investigated by comparing measured RIAs to simulated RIAs for each observed isotopomer. The theoretical RIA was simulated by first generating a carbon-only MID using the measured fractional enrichment for ^{13}C. The carbon-only MID was then convolved with naturally occurring elemental mass distributions for each remaining element within the peptide. The simulated MID and the measured MID were both renormalized to eliminate the effects of missing isotopomer values. This was required because isotopomers that were not observed have to be removed from the simulated MID, resulting in simulated distributions that no longer sum to 1. After normalization, a nearly one-to-one relationship was observed between simulated RIA and measured RIA (see Figure 4.1, panel (A)). A chi-squared statistic was calculated to test the goodness of fit between the measured RIA and the theoretical RIA. For each isotopomer, ι_i of peptide k, the measured RIA, $p^*_{i,k}$, and the theoretical RIA, $p_{i,k,\text{pred}}$ were used to generate a reduced chi-squared test statistic, χ^2_{red}, for a total number of data points M and parameters ρ, such that

$$\chi^2_{\text{red}} = \frac{1}{\nu} \sum_{i,k} \frac{N_k^2 (p^*_{i,k} - p_{i,k,\text{pred}})^2}{\text{Var}\left(p_{i,k,\text{pred}}\right)_{\text{estimate}}} \tag{4.3}$$

where $N_k = \sum_{j \in \mathcal{S}_k} \eta^*_{j,k}$ and $\nu = M - \rho - 1$. The parameter set for this model is simply the average carbon label use to generate our estimated RIAs, meaning that $\rho = 1$ (Taylor, 1996). Equation (4.3) can be reduced using equation (3.8) to the following equation:

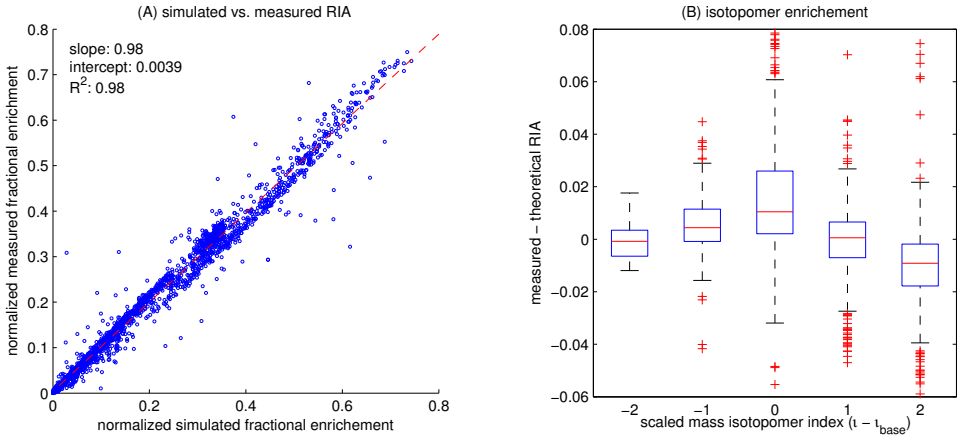
$$\chi^2_{\text{red}} = \frac{1}{\nu} \sum_{i,k} \frac{N_k}{\beta^2} \frac{(p^*_{i,k} - p_{i,k,\text{pred}})^2}{p_{i,k,\text{pred}}(1 - p_{i,k,\text{pred}})} \tag{4.4}$$

χ^2_{red} is measure of goodness of fit between observed RIAs and simulated RIAs, weighted by the estimated variance. Using (4.4), a χ^2_{red} of 10.49 was calculated, resulting a in p-value close to 0. These results indicate the theoretical RIA did not fit the data well based on using the multinomial sampling model.

To investigate the possibility of isotopic bias, the difference between the measured RIA and predicted RIA was plotted as a function of isotopomer mass state relative to the base peak, denoted as $\delta_{\text{RIA}}(\iota_i)$. Figure 4.1, panel (B) shows the distribution of $\delta_{\text{RIA}}(\iota_i)$ for rescaled mass isotopomers. If there were sampling bias for more abundant isotopomers, one would expect that the mean $\delta_{\text{RIA}}(\iota_i)$ (referred to herein as $\bar{\delta}_{\text{RIA}}(\iota_i)$) would be significantly higher for isotopomers closest to the base peak, which is observed in Figure 4.1, panel (B).

In summary, using a direct approach for the identification of peptides within an unlabeled sample, we were able to identify precursor ion mass traces for all peptides within the validation database. Due the high-rate of missannotation, our inability to identify more than one isotopomer species per peptide and the existence of potential overlapping MIDs, MIDs could be calculated for 75% (484/648) of peptides. However, this recovery rate will improve when the software takes into consideration the effects of modifications as well as causes for missannotation.

Figure 4.1: RIA and isotopic enrichment analysis for unlabeled *E. coli*: (A) simulated RIAs using the calcuated average carbon label of 1.02% are plotted against measured RIAs, (B) The difference between simulated and measured RIA as a function of mass isotopomer state scaled by the base peak (most abundant peak). Oversampling is observed for the most abundant peak.



4.2 7% Uniform Labeled *E. coli*

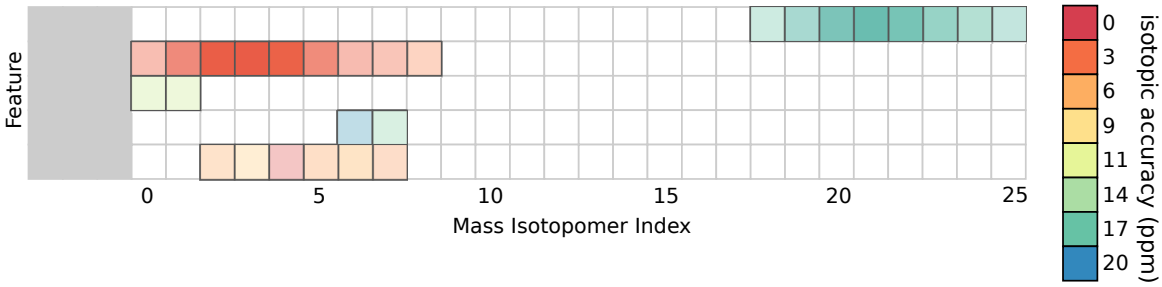
Software was validated with data containing well-defined isotopic enrichment; *E. coli* was cultured under the same conditions described in Section 4.1 except in the presence of 7% [U-$^{13}C_6$]-glucose described by Allen et al. (2013).

4.2.1 methods

The 484 peptides used for the quantification validation of unlabeled *E. coli* in Section 4.1.2 were used for software validation on labeled data. This was done to ensure that peptide MIDs for the validation set could be quantified from unlabeled material and were not falsely identified.

Peptide features were generated using the parameter and protocol described in Chapter 3 with a mass accuracy cutoff (MAC) of 20 ppm and a retention time window (RTW) of ±2 minutes. The clustering algorithm used a dendrogram height cutoff of 2.5, and features were broken up into features containing contiguous isotopomer sets

Figure 4.2: Cluster result example for labeled *E. coli*: Each block represents a mass trace and rows and columns represent candidate features mass isotopomers, respectively. The color gradient is a function of isotopic accuracy and opacity is a function of the abundance of isotopologues.



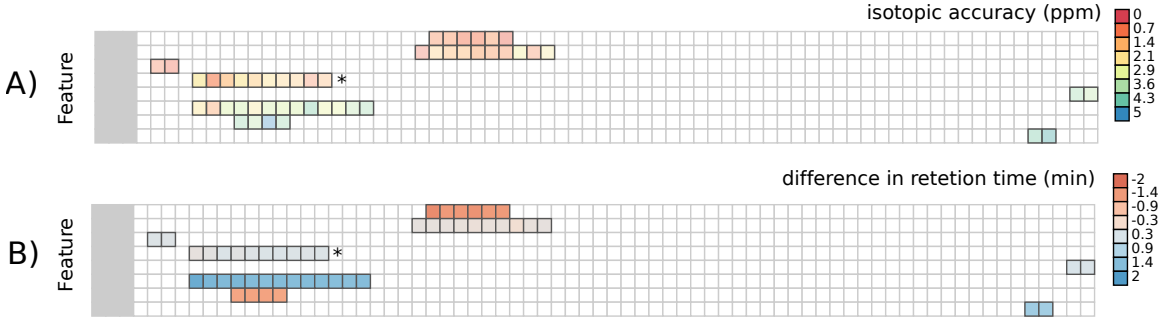
prior to annotation. Figure 4.2 is an example of the output provided by the clustering algorithm for labeled samples, where each block represents a mass trace within a potential feature. The row corresponds to a candidate feature for a peptide, while the columns represent the mass isotopomer index for that mass trace. The color gradient is a function of the isotopic accuracy while the opacity is a function of isotopologue abundance. In this example, the true peptide feature is observed in the second row. Without additional information, for labeled populations it would be impossible to distinguish the true feature from other candidates. For example, Figure 4.2 shows five candidate features, all of which are viable options if abundance and labeling are unknown. On average, there were 8 candiate features for each peptide, with only 10 % of peptides (49/484) containing one unique feature.

Re-parameterizing MTC and RTW

To reduce the number of candidate features per peptide, visualizations analogous to Figure 4.2 were generated for each peptide. It was observed that true features exhibit higher mass accuracy than other candiate features. For example, filtering with a MAC of 5 ppm would be sufficient to remove all other candidate clusters in Figure 4.2.

Visualizations similar to Figure 4.2 were generated to encode both isotopic ac-

Figure 4.3: Cluster result example for labeled *E. coli*: similar to Figure 4.2 except color gradient in the figure below represents the difference in retention time between the reference precursor peak and the retention time of the mass trace. Features denoted with an asterisk (*) represented the true feature for this peptide. In this example, although the m/z values are very similar between two candidate features, the true feature is much closter in retention time to precursor.



curacy and retention time to find other distinguishing characteristics between true features and other candidates. Figure 4.3 (A) and (B) show an example of the clustering results, where the color gradient corresponds to mass accuracy and retention in (A) and (B), respectively. While Figure 4.2 shows that features can be unique determined when reducing the MAC, very similar candidate features exist in both mass and isotopomer sets, but differ dramatically in retention time. To eliminate candidate mass traces, the RTW was lowered to ± 1 minutes.

The mean number of features per peptide was 8.3 at a mass accuracy cutoff of 20 ppm and retention time window of 4 minutes. However, once the mass accuracy cutoff was lowered to 5 ppm and the retention time window was reduced to 2 minutes the average number of features were reduced to 1.8, which allowed for the unique identification of many more peptides.

Maximizing Recovery

After re-filtering mass traces with a MAC of 5 ppm and RTW of ± 1 minute, candidate features were determined by the clustering procedure described previously. Nearly all

of the peptides contained at-least one feature while 53% (258/484) peptides contained one unique feature per peptide. MIDs for the 258 unique peptide features were quantified and average peptide carbon label was calculated using (4.1). A population mean for all average peptide carbon label was calculated and used to simulate MIDs. Unlike the methods described in Section 4.1, the mean average carbon label was used to simulate MIDs for all peptides in the training set, and the base peak was determined for each peptide. Candiate features were removed if they did not contain the base peak, which resulted in 94% (454/484) of peptides containing one unique feature while only 3% (13/484) of peptides had more than one feature containing the base peak.

Quantification

RIAs were quantified and the average carbon label per peptide was calculated. For the labeled dataset, 7 % (1525/21,476) of expected peaks were missing, and 40% of scan events contained at least one missing peak. If the fitting procedure were not employed, 18% (85/484) of peptides would not be quantifiable.

RIAs were simulated using a mean average carbon label of 6.68% and compared to measured values. A nearly one-to-one relationship was observed with high and significant correlation between simulate and measured RIAs (Slope: 0.99 and R^2 : 0.98). However, multinomial sampling variance could not fully explain the observed variance ($\chi^2_{\text{red}} = 8.31$), much like results obtained from the unlabeled sample.

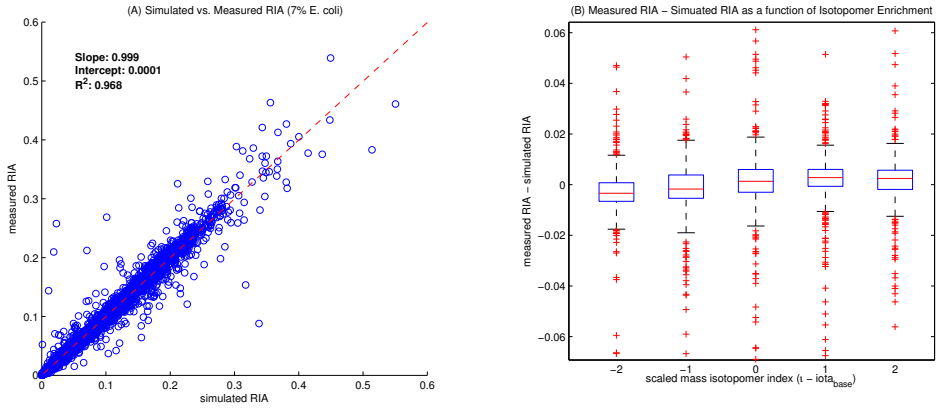


Figure 4.4: RIA and isotopomer erichment analysis for uniformly labeled *E. coli*: (A) simulated RIAs using the calcuated average carbon label of 6.68% are plotted against measured RIAs, (B) The difference between simulated and measured RIA as a function of mass isotopomer state scaled by the base peak. Oversampling is not as prenounced for the heavily labeled data set.

4.3 Labeled Soy Samples

The final software validation was performed data from the peptide-based metabolic flux analysis of soy seeds. The data were generated under the same experimental conditions as listed in Section 3.1, except chymotrypsin was used for protein digestion instead of trypsin. MaxQuant was able to identify 19 proteins from 193 peptide spectra.

Recovery Results

Mass traces were filtered using a 5 ppm MAC with a RTW of ±1 minute. A dendrogram height cutoff set to 2.5. Peptides were kept if there was one unique feature remaining. 34% (66/193 , 966 mass isotopomers) of peptides were quantified, with 11/19 proteins containing at least one measurable peptide. MIDs were quantified and the average carbon label was determined to be 23.3%. The mean carbon label was then used to simulate base peaks for all 193 peptides in the original database and

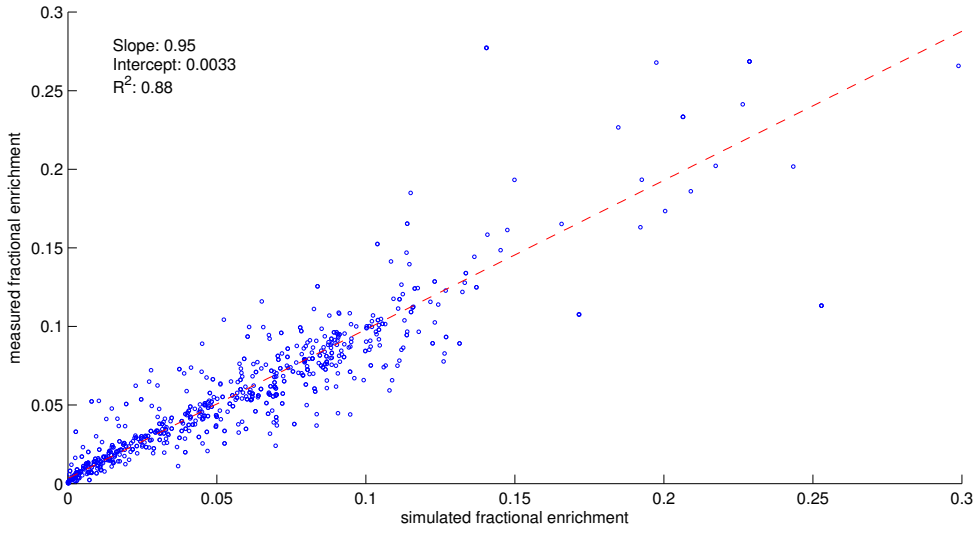
second search was performed. Candidate features were further reduced by requiring the existence of the simulated base peak. The number of quantifiable peptides increased by 42% (29/193) resulting in a final recovery rate of 49%. 74% of proteins had at least one measurable peptide (14/19) while a total of 1426 mass isotopomer RIAs were quantified.

Quantification

To test the quantitative capabilities for flux-dependent labeling samples, the extracted RIAs were compared to simulated RIAs using deconvolution derived amino acid MIDs generated by Mandy et al. (2013). Amino acid MIDs (AAMDs) were determined for peptides in the training set originating from soy storage proteins glycinin and beta-conglycinin. AAMDs were calculated using flux-constrained deconvolution described in Section 1.3.3.

51 peptides were extracted from glycinin and beta-conglycinin from the chymotrypsin digestion data set, and simulated RIAs were generated by first simulating the labeled and unlabeled isotopologue populations, then mixing the populations using the optimized ratio. The labeled population was derived by convolving optimized AAMDs from flux-constrained deconvolution, while the unlabeled population was generated by convolving natural elemental distributions. The labeled population was then mixed with 2.23% unlabeled, original biomass. Simulated and measured RIAs were normalized and weighted linear regression was performed, where each RIA was weighted by the total number of ions. Figure 4.5 show a nearly one to one relationship between predicted and measured RIA's (Slope: 0.95, R^2: 0.88).

Figure 4.5: Simulated vs. measured RIA in Soybean: RIAs were simulated by convolving AAMDs generated by flux-constrained deconvolution of the training set followed by mixing in 2.23% of unlabeled isotopologues.



4.4 Summary

The extraction of unlabeled, uniformly labeled and flux dependent labeled data sets were demonstrated high accuracy and recovery. Visualizations were used to optimize extractions for labeled datasets: a MAC of 5 ppm and a RTW of ± 1 minute was observed to sufficiently reduce the number of candidate features. This resulted in approximately 50% of peptides with one unique feature, allowing for the calculation of RIAs and average carbon label. Using the unsupervised approach discussed in Section 4.2 to estimate base-peaks for the peptide set, peptide feature recovery increased by 50% for both labeled data sets, resulting in recovery rates of 94% and 49% for uniformly and flux-dependent labeled samples, respectively. The high occurrence of missing data was demonstrated for labeled data sets, highlighting the need for fitting RIAs to measured data. The quantification method produced estimated RIAs that showed extremely high similarity and correlation with simulated values.

Chapter 5

Discussion and Future Work

In Chapter 4, the quantification capabilities were investigated using unlabeled data prior to automated extraction. Measured RIAs were shown to be very similar to simulated RIAs. However, the multinomial sampling variance did not fully explain the experimental variance based on the reduced chi-squared test, which could be attributed to instrument or signal-dependent variance, or the observed sampling bias for highly abundant peaks.

The use of automated extraction of PMDs from uniformly labeled and flux-dependent labeled experiments were demonstrated with 94% and 49% recovery rates, respectively. Furthermore, the quantification of uniformly labeled data showed a slightly better fit to simulated RIAs and reduced observed sampling bias. In all cases, nearly one-to-one relationships were observed between simulated and measured RIAs with strong correlation.

In this section software and experimental approaches are discussed for improving identification rates and feature annotation, followed by the experimental approaches for improving RIA quantification and error estimation. The final section touches on the software and interactive visualization tools that will provided to biological community.

5.1 Software Improvements

5.1.1 Identification

Roughly 75% of peptides identified from MaxQuant were quantifiable using our software implementation. The largest reason for why peptides were not quantified were missannotations within the peptide database. Missannotations can be improved by developing a workflow that integrates other peptide identification search software to corroborate the existence of the peptide. Furthermore, incorporating the possibility of modifications results in a larger false discovery rate for tandem MS-MS (fragmentation) based identification approaches (Mallick and Kuster, 2010). Additionally, the percentage of precursor ions sequenced in any proteomic experiment represents only a subset of peptides within the sample. As instruments provide improved mass resolution, accuracy and signal dynamic range for LC-MS, non-fragmentation based identification methods could be used to corroborate sequenced based methods and reduce the rate of false positives.

Another factor for why peptides were not quantified was that isotopomers other than precursor mass isotopomer was not observed. This was most likely attributed to the low abundance of the precursor, but has not been investigated.

Lastly, peptides were not quantified when an intense and highly correlated peak was observed in the m-1 state. Although the loss of discrete molecular groups from the peptides are commonly observed (e.g. neutral loss), intense peaks in the the m-1 state are often not observed. Thus, the observation of intense peaks in the m-1 state are potentially caused by overlapping MIDs from other molecular species. However, better characterization of the probability associated with neutral loss in peptides will allow for improved detection of potential contaminations.

Currently, the software does not take into consideration the score provided by the identification software. We investigated the use of the score by comparing the mean

scores of quantifiable peptides for the unlabeled *E. coli* with the mean scores of peptides that were not quantified. However, only a slightly higher score for quantifiable peptides was observed.

5.1.2 *m/z* Calibration

The *m/z* domain was shown to exhibit a significant amount of drift throughout a single LC-MS experiment. To correct for this, a simple "lock" mass calibration used in traditional proteomic experiments should be employed for labeled data. By using a standard that is injected into the mass analyzer, the software could rescale each scan in the *m/z* domain. Using a calibrant will allow for smaller isotopic precision cutoff, subsequently increasing isotopic similarity between mass isotopomers.

It was also observed that the isotopic precision between mass traces could be dependent on the absolute difference in *m/z*. This effect is most likely attributed to a potential **space charging** effect, where intensely charged groups of isotopomers repel one another, resulting a larger measured difference in *m/z*. Approaches for correcting for this effect have been proposed, but are not implemented within the software (Gorshkov et al., 2010).

5.1.3 Feature Annotation

In order to distinguish between candidate features for a peptide, additional information for which isotopomers are more likely to be observed is required. From Section 4.2 and Section 4.3, requiring the feature contain the most likely peak based on a simulated distribution increased recovery rates by 50%. The approach that worked best was to simulate a base peak using the mean label enrichment within subset of the data, followed by requiring feature candidates to contain the simulated base peak. This approach works best for larger peptides due to the *central limit theroem*. As pep-

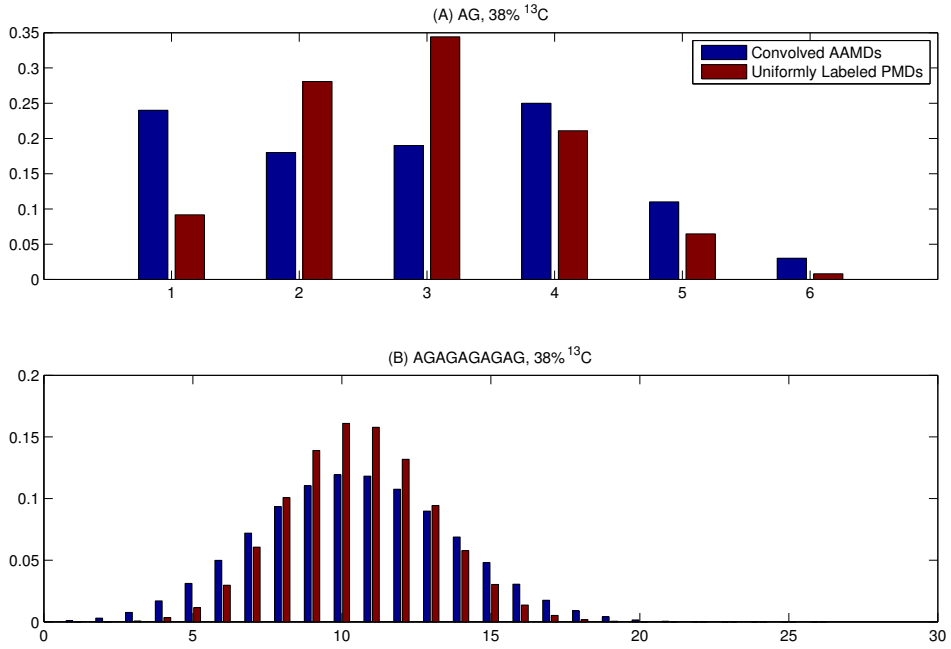


Figure 5.1: Using Average Carbon Label to Simulate PMDs: MIDs were simulated by convolving either non-uniformly labeled AAMDs (blue bars) or uniformly labeled elements (red bars). All distributions contain the same average carbon label of 38%. The simulated base peaks are equivalent for the longer peptide (B), but not for the short peptide (A)

tide length increases, the PMDs are constructed from a larger number of AAMDs. As a consequence, the PMD approaches a unimodal, normal distribution with a base peak that approaches the base peak of a uniformly labeled peptide. This effect is illustrated in Figure 5.1 for peptides AG and AGAGAGAGAG, both with an average carbon label of 38%.

If the investigator was interested in obtaining smaller peptides, an experimental approach is suggested in Figure 5.2. The investigator runs two experiments, one designed to produced flux dependent PMDs and the other from unlabeled (uniform and naturally abundant) substrate. Material from both conditions are mixed using a fixed ratio. MS spectra are collected from the mixed sample as well as an unlabeled sample,

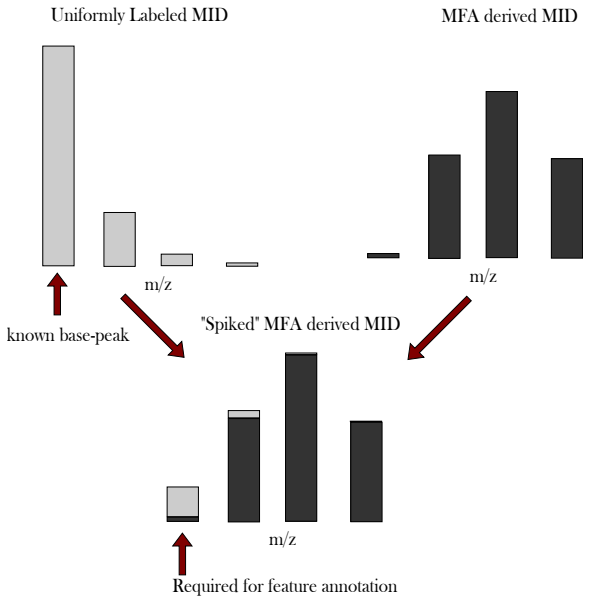


Figure 5.2: Experimental Design Example: Spiking with Known PMDs: Biomass is grown seperately using uniformly labeled substrate and substate that provides flux information. Both are mixed and the software uses the basepeak of the uniformly labaled PMD to identify the feature.

and a base peak is calculated from the average label within the the unlabeled sample. The software uses this base-peak as annotation criteria for the mixed sample. Using the fixed ratio between unlabeled and flux-labeled biomass, the unlabeled population is then subtracted from the labeled population (see Section 1.3.1).

Lastly, the feature annotation method makes no attempt to distinguish true from false annotations. Falsely annotated features could be observed using the reduced chi-squared residuum per for each peptide to remove PMDs with significantly different average carbon label. For example, the peptide with the largest chi-squared residuum had a 2.8% average carbon label in the 7% uniformly labeled data set. This suggests that false positives might be detectable using the outliers in the chi-squared residuum for uniformly labeled experiments or the objective function in PMD optimization scheme for flux-labeled experiments.

5.2 Quantitative Capabilities

In addition to the feature identification improvements previously discussed, the quantitative capabilities of our methods should be compared to current practices. Additionally, new statistical models can be developed that potentially could account for the additional variance described in Chapter 4

5.2.1 RIA Estimation

Section 3.5 describes a new method for quantifying MIDs using LC-MS data. The model assumes that missing peaks are primarily attributed to software limitations within the pipeline, ranging from collecting data from the mass analyzer to preprocessing of peaks using low isotopic precision cutoffs, rather than that the signal was below the limit of detection. This was motivated by the fact that for non-observed peaks, the number of isotopologues at the limit of detection was more than three standard deviations less than the expected number of isotopologues. However, an established statistical method for distinguishing between peaks below the limit of detection and missing peaks has not been demonstrated.

The rate of missing peaks is a function of the isotopic precision specification within the software. In Section 3.5, the feature matrix is filtered prior to quantification based on the isotopic precision for each m/z measurement. If this parameter is too small, completely valid peaks might be artificially removed from the feature matrix. The robustness of the quantification algorithm has not been throughly tested to demonstrate the performance as a function of isotopic precision.

Lastly, the quantification algorithm has not yet been compared to current methods of calculated RIAs. RIAs are typically quantified by determining the fractional abundance of each isotopomer over a retention time range. We have initially observed slightly better fits between simulated RIAs and measured RIAs using the fitting ap-

proach over current methods. However, this analysis could be inherently biased because the simulated model uses an average carbon label calculated from fitted RIAs. Unbiased tests are currently being devised to compare these quantification methods.

5.2.2 Variance Estimation

Currently the most sophisticated statistical model of mass high resolution LC-MS data is based on the multinomial distribution, which assumes that sampling occurs with replacement. In this model, the observed signal is assumed to come from a small fraction of the total population of isotopologues from the experiment. However, in reality LC-MS can be thought of as two sampling events: (1) the sample from the experiment to the LC column, and (2) the sampling from the LC to the mass analyzer. While (1) would still come from a multinomial distribution, (2) would come from a hypergeometric distribution. At each scan event, a sample of isotopologues are drawn from the finite number of isotopologues that exists on the column. Each subsequent scan event then samples from the remaining set of isotopologues, which is smaller than the previous scan event. Currently, the model assumes that each scan event samples from the same population of isotopologues. More accurate estimation of measurement variance could be derived using a model such as this, however has not been tested.

5.3 Future Work

In addition to software that was developed for the extraction and quantification of PMDs, a substantial work went into developing new visualization tools to explore mass spectrometry data and the output at each stage of our method. For example, Section 4.2 demonstrates the use of flexible visualization for software development and data analysis, aiding both the parameterization and troubleshooting process. In

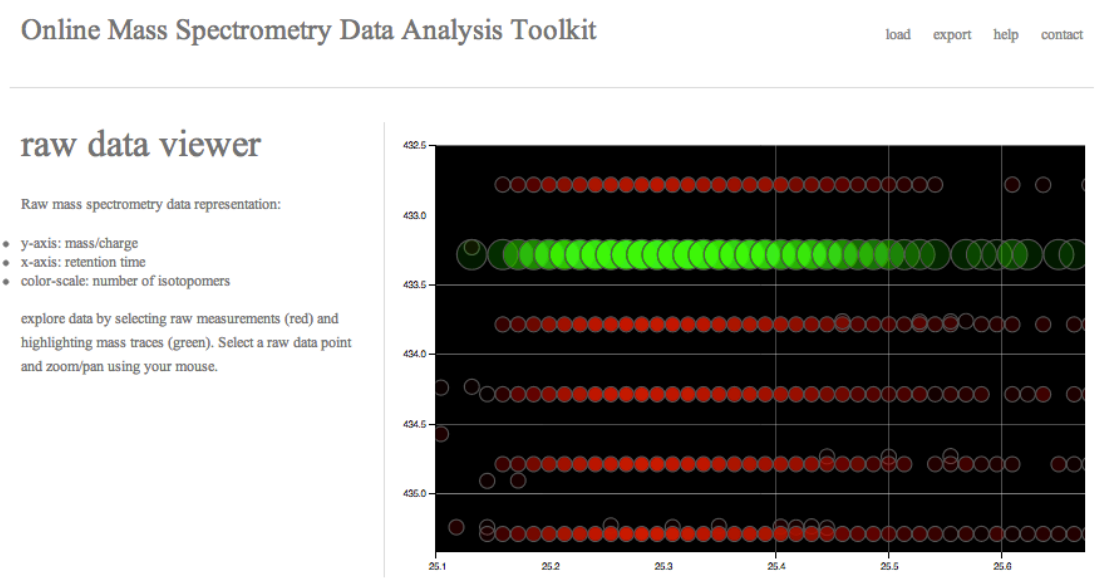


Figure 5.3: interactive mass spectromtery web-page screen shot: raw data reduced at any stage of the algorithm can be loaded and visualized interactively

addition to making software available, there will be an online visualization toolkit allowing the user to explore their data sets at each stage of the processes. Figure 5.3 is an example screenshot from a web-interface currently in development. This data set was generated after peaks had been grouped into mass traces, but before features had been identified. Having access to easy-to use and intuitive visualization tools such as this will give the user a better understanding of their data.

The workflow shown in Figure A.1 will initially be implemented as modular pieces of MATLAB scripts and functions. The code will be cleaned, commented and hosted on and online project hosting page, such as GitHub. After debugging and code optimization, the software will be compiled and hosted by an online bioinformatics platform such as Galaxy.

References

Allen, D. K., Goldford, J. E., Gierse, J. K., Mandy, D., Diepenbrock, C., and Libourel, I. (2013). Orbital trap mass spectrometry measurements faithfully represent m/z ion distributions of peptides from isotope labeled cultures.

Allen, D. K., Shachar-Hill, Y., and Ohlrogge, J. B. (2007). Compartment-specific labeling information in 13C metabolic flux analysis of plants. *Phytochemistry*, 68(16-18):2197–210.

Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic engineering*, 9(1):68–86.

Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry*, 404(4):939–65.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389(4):1017–31.

Becker, G. W. (2008). Stable isotopic labeling of proteins for quantitative proteomic applications. *Briefings in functional genomics & proteomics*, 7(5):371–82.

Bueschl, C., Kluger, B., Berthiller, F., Lirk, G., Winkler, S., Krska, R., and Schuhmacher, R. (2012). MetExtract: a new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. *Bioinformatics (Oxford, England)*, 28(5):736–8.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Cengage Learning.

Chokkathukalam, A., Jankevics, A., Creek, D. J., Achcar, F., Barrett, M. P., and Breitling, R. (2013). mzMatch-ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics (Oxford, England)*, 29(2):281–283.

Clomburg, J. M. and Gonzalez, R. (2010). Biofuel production in Escherichia coli: the role of metabolic engineering and synthetic biology. *Applied microbiology and biotechnology*, 86(2):419–34.

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–72.

Cox, J. and Mann, M. (2009). Computational principles of determining and improving mass precision and accuracy for proteome measurements in an Orbitrap. *Journal of the American Society for Mass Spectrometry*, 20(8):1477–85.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research*, 10(4):1794–805.

Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., Kruger, M., Ishihama, Y., and Mann, M. (2011). Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nature protocols*, 6(2):147–57.

Gorshkov, M. V., Good, D. M., Lyutvinskiy, Y., Yang, H., and Zubarev, R. A. (2010). Calibration function for the Orbitrap FTMS accounting for the space charge effect. *Journal of the American Society for Mass Spectrometry*, 21(11):1846–51.

Hellerstein, M. K. and Neese, R. A. (1999). Mass isotopomer distribution analysis at eight years: theoretical, analytic, and experimental considerations. *The American journal of physiology*, 276(6 Pt 1):E1146–70.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry : JMS*, 40(4):430–43.

Jansson, P. A. (1996). *Deconvolution of Images and Spectra, Second Edition*. Academic Press.

Kaufmann, a. and Walker, S. (2012). Accuracy of relative isotopic abundance and mass measurements in a single-stage orbitrap mass spectrometer. *Rapid communications in mass spectrometry : RCM*, 26(9):1081–90.

Keibler, M. A., Fendt, S.-M., and Stephanopoulos, G. (2012). Expanding the concepts and tools of metabolic engineering to elucidate cancer metabolism. *Biotechnology progress*, 28(6):1409–18.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008). ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics (Oxford, England)*, 24(21):2534–6.

Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007). TOPP–the OpenMS proteomics pipeline. *Bioinformatics (Oxford, England)*, 23(2):e191–7.

Krumholz, E. W., Yang, H., Weisenhorn, P., Henry, C. S., and Libourel, I. G. L. (2012). Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *Journal of experimental botany*, 63(6):2353–62.

Listgarten, J. and Emili, A. (2005). Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular & cellular proteomics : MCP*, 4(4):419–34.

Makarov, A., Denisov, E., Lange, O., and Horning, S. (2006). Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *Journal of the American Society for Mass Spectrometry*, 17(7):977–82.

Mallick, P. and Kuster, B. (2010). Proteomics: a pragmatic perspective. Supp Info. *Nature biotechnology*, 28(7):695–709.

Mandy, D., Goldford, J., Yang, H., Allen, D., and Libourel, I. (2013). Metabolic flux analysis using 13C peptide label measurements.

Melamud, E., Vastag, L., and Rabinowitz, J. D. (2010). Metabolomic analysis and visualization engine for LC-MS data. *Analytical chemistry*, 82(23):9818–26.

Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Müller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J.-J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012). Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Molecular & cellular proteomics : MCP*, 11(3):O111.013698.

Möllney, M., Wiechert, W., Kownatzki, D., and de Graaf, a. a. (1999). Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotechnology and bioengineering*, 66(2):86–103.

Monroe, M. E., Tolić, N., Jaitly, N., Shaw, J. L., Adkins, J. N., and Smith, R. D. (2007). VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics (Oxford, England)*, 23(15):2021–3.

Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., and Müller, M. (2007). SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*, 7(19):3470–80.

Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012). System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Molecular & cellular proteomics : MCP*, 11(3):M111.013722.

Okawa, S., Fischer, B., and Krijgsveld, J. (2013). Properties of isotope patterns and their utility for peptide identification in large-scale proteomic experiments. *Rapid communications in mass spectrometry : RCM*, 27(9):1067–75.

Olsen, J. V., de Godoy, L. M. F., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005). Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & cellular proteomics : MCP*, 4(12):2010–21.

Perry, R. H., Cooks, R. G., and Noll, R. J. (2008). Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass spectrometry reviews*, 27(6):661–99.

Quek, L.-E., Wittmann, C., Nielsen, L. K., and Krömer, J. O. (2009). OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis. *Microbial cell factories*, 8:25.

Reinert, K. and Kohlbacher, O. (2010). OpenMS and TOPP: open source software for LC-MS data analysis. *Methods in molecular biology (Clifton, N.J.)*, 604:201–11.

Rouessac, F. and Rouessac, A. (2007). *Chemical Analysis: Modern Instrumentation Methods and Techniques*. Wiley.

Rühl, M., Hardt, W.-D., and Sauer, U. (2011). Subpopulation-specific metabolic pathway usage in mixed cultures as revealed by reporter protein-based 13C analysis. *Applied and environmental microbiology*, 77(5):1816–21.

Scigelova, M. and Makarov, A. (2006). Orbitrap mass analyzer–overview and applications in proteomics. *Proteomics*, 6 Suppl 2:16–21.

Smith, C. a., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, 78(3):779–87.

Tautenhahn, R., Böttcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, 9(1):504.

Taylor, J. R. (1996). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books.

Wasylenko, T. M. and Stephanopoulos, G. (2013). Kinetic isotope effects significantly influence intracellular metabolite (13) C labeling patterns and flux determination. *Biotechnology journal*.

Weitzel, M., Nöh, K., Dalman, T., Niedenführ, S., Stute, B., and Wiechert, W. (2013). 13CFLUX2–high-performance software suite for (13)C-metabolic flux analysis. *Bioinformatics (Oxford, England)*, 29(1):143–5.

Wiechert, W. (2001). 13C metabolic flux analysis. *Metabolic engineering*, 3(3):195–206.

Wiechert, W. and de Graaf, a. a. (1997). Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. *Biotechnology and bioengineering*, 55(1):101–17.

Wiechert, W., Möllney, M., Isermann, N., Wurzel, M., and de Graaf, a. a. (1999). Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnology and bioengineering*, 66(2):69–85.

Wiechert, W., Siefke, C., de Graaf, a. a., and Marx, a. (1997). Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. *Biotechnology and bioengineering*, 55(1):118–35.

Wong, W. and Sackett, W. M. (1975). Isotope Fractionation in Photosynthetic Bacteria during Carbon Dioxide Assimilation. *Plant physiology*, 55(3):475–9.

Yates, J. R., Cociorva, D., Liao, L., and Zabrouskov, V. (2006). Performance of a linear ion trap-Orbitrap hybrid for peptide analysis. *Analytical chemistry*, 78(2):493–500.

Yates, J. R., Ruse, C. I., and Nakorchevsky, A. (2009). Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering*, 11:49–79.

Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). (13)C-based metabolic flux analysis. *Nature protocols*, 4(6):878–92.

Zhang, J., Gonzalez, E., Hestilow, T., Haskins, W., and Huang, Y. (2009). Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current genomics*, 10(6):388–401.

Appendix A

Software Design

A.1 File Preparation

Because many programs contain different data formate requirements, it is important to describe the correct formats for each data sample one would obtain in this workflow. Msconvert (ProteoWizard) was used to convert thermo .RAW files to mzXML files. Vendor supplied centroided peaks were used within the peak picking algorith. Additional parameters used with msconvert are provided in the software package. Figure A.1 shows the file flow at each stage of the procedure. The files ending in MAT are data structures within MATLAB.

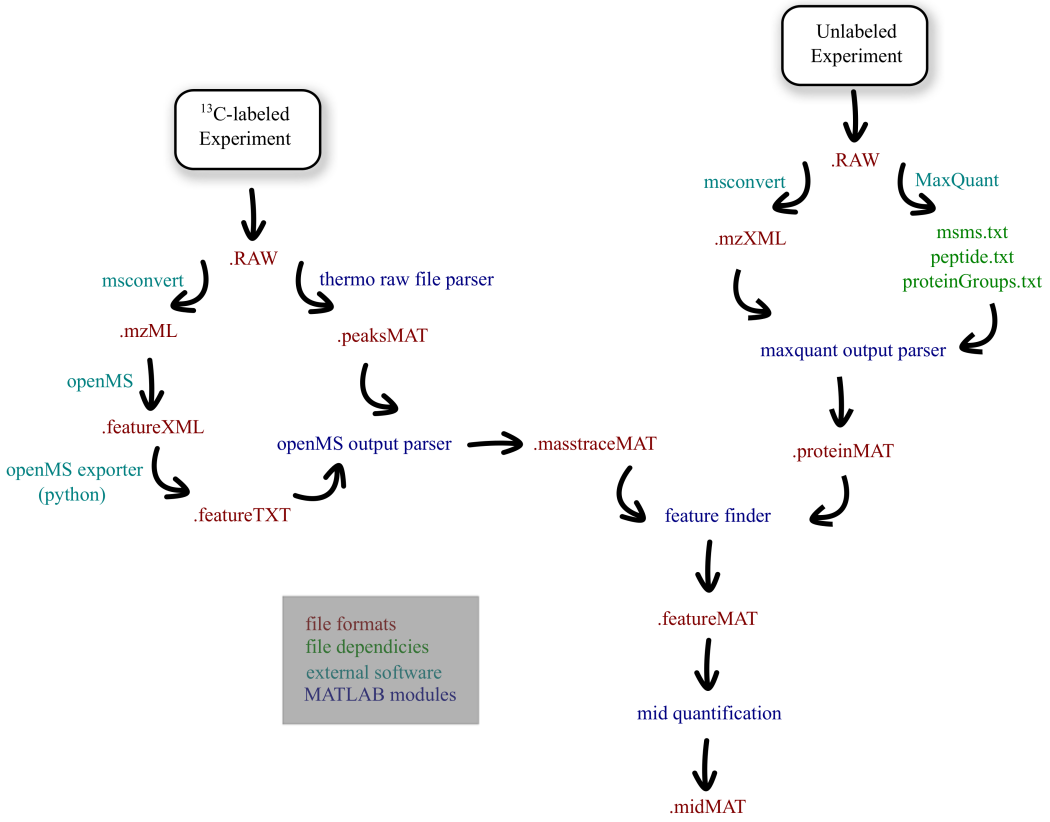


Figure A.1: Detailed software design flow and file formats:

Appendix B

Statistical Interpretation of Mass Isotopomers

For a given peptide, i, we can define every observed isotopomer as a random variable, X_i, drawn from a finite sample space, χ_i consisting of all possible mass isotopomers, N_i. The discrete probability of drawing mass isotopomer j is $\Pr\{X_i = j\}$, where $j \in \chi_i = \{0, 1, \ldots, N-1\}$. From this point further, we will use the notation $p_{i,j} \equiv \Pr\{X_i = j\}$ for clarity.

$$\sum_{j}^{Ni} p_{i,j} = 1 \tag{B.1}$$

B.1 Quantification of Isotopomer Ion Counts

For the i^{th} peptide in our measurement set, we can estimate the number of ions for isotopomer j, $I_{i,j}$, from the absolute peak intensity and noise, $S_{i,j}$, and $N_{i,j}$, respectively, using the following equation:

$$I_{i,j} = \frac{S_{i,j}}{N_{i,j}} \frac{K\sqrt{\frac{R}{R_0}}}{z_i} \tag{B.2}$$

where K is the noise band, R is the resolution setting and R_0 is the reference resolution setting for time-dependent acquisition.

We can extend our ion count estimate, $I_{i,j}$, to include an additional index, k, which corresponds to the retention time, or **scan event**, associated with that measurement. Therefore, the total number of ions for a particular isotopomer in a given peptide is given by

$$\sum_{k=s_1}^{s_M} I_{i,j,k} = I_{i,j} \quad \forall i, j \tag{B.3}$$

where $\{s_1 \ldots s_M\}$ is the set of scans that contain the j^{th} isotopomer of peptide i. Furthermore, it is convenent to define the set of all measuremed isotopologue for

peptide i by $\mathcal{I}_i = \sum_j I_{i,j}$.

We can estimate the probability of sampling isotopomer j for peptide i, $\Pr\{X_i = j\}$, by taking

$$\Pr\{X_i = j\} = \frac{I_{i,j}}{\sum_j I_{i,j}} = \frac{I_{i,j}}{\mathcal{I}_i} \tag{B.4}$$

where $\sum_j I_{i,j}$ is the total number of ions for all isotoplogues for peptide i. This procedure is analogous to summing all ions observed for a particular isotopomer and dividing by the total number of isotopologues extracted for that peptide within the scan event set $S_i = \{s_1 \dots s_M\}$. This enables us to generate an expected value, $E(X_i = j)$ of sampling the j^{th} isotopomer in peptide i for any sample size, η, such that

$$E(X_i = j, \eta) = \eta \Pr\{X_i = j\} = \eta \frac{I_{i,j}}{\sum_j I_{i,j}} \tag{B.5}$$

B.2 Isotopomer Ion Counts as Random Multinomial Samples

It was proposed that the observed sampled isotopologue population is equivalent to a *multinomial distribution*. The multinomial distribution is an extension of the binomial distribution, where independent trials (observation of a single mass isotopomer ion) leads to success in exactly one of many categories (i.e. sampling mass isotopomer m+0).

This is relevant because this represents the physical sampling of the total population of isotopologues within the cell. Assuming that the observed group of isotopologues are only a fraction of the total population within the cell, the multinomial distribution generalization allows us to calculate variances and covariances for mea-

sured sampling probabilities for each isotopomer.

Extending from Appendix B.1, if the sampling of isotopomer j has probability $p_{i,j}$, we define the expected value as:

$$E(X_i = j) = \mathcal{I}_i p_{i,j} \tag{B.6}$$

with variance and covariance defined as:

$$\mathrm{Var}\,(X_i = j) = \mathcal{I}_i p_{i,j}(1 - p_{i,j}) \tag{B.7}$$

$$\mathrm{Cov}\,(X_i = j, X_i = k) = -\mathcal{I}_i p_{i,j} p_{i,k}\ (j \neq k) \tag{B.8}$$

Appendix C

Raw Data

Feature matricies for the training data are shown in Figure C.1 while cluster results are shown in Figure C.2.

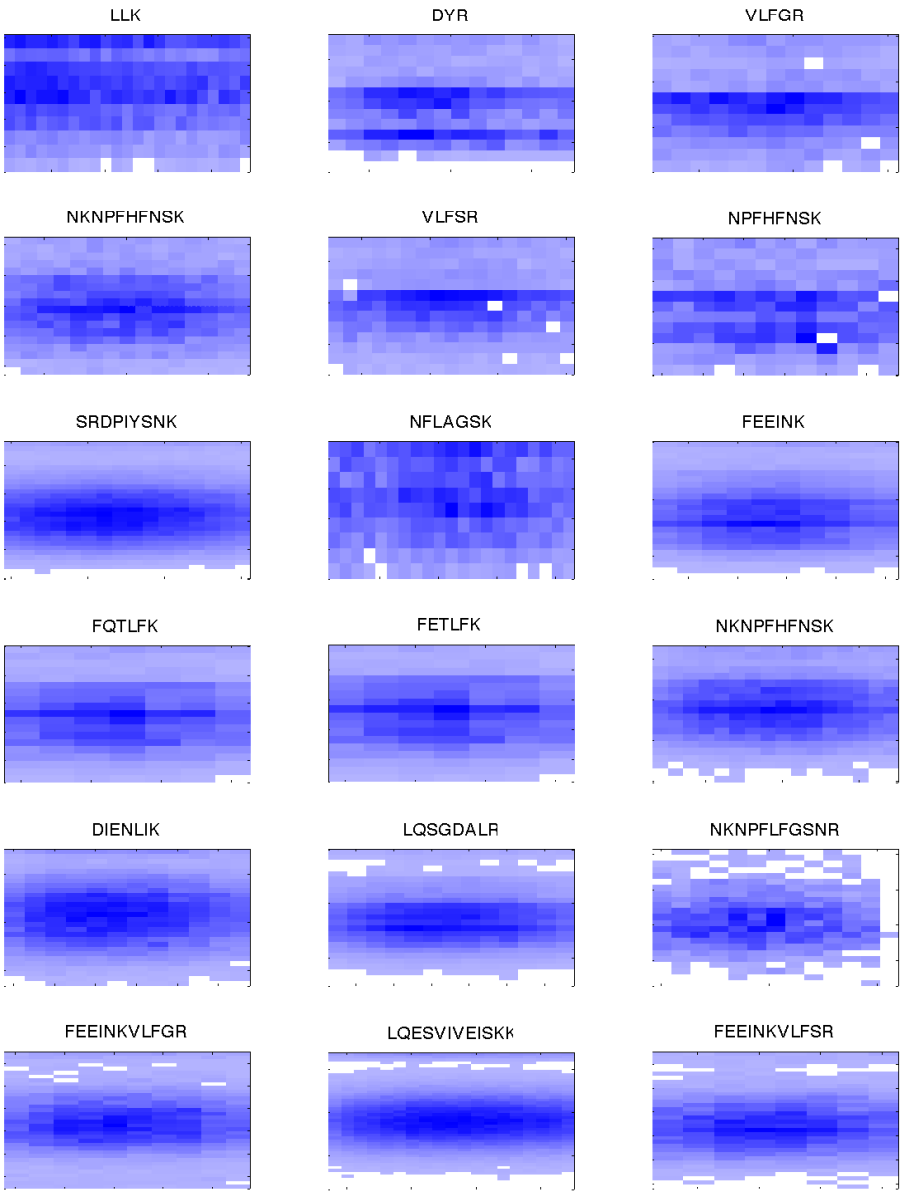


Figure C.1: Feature Matrix: training data

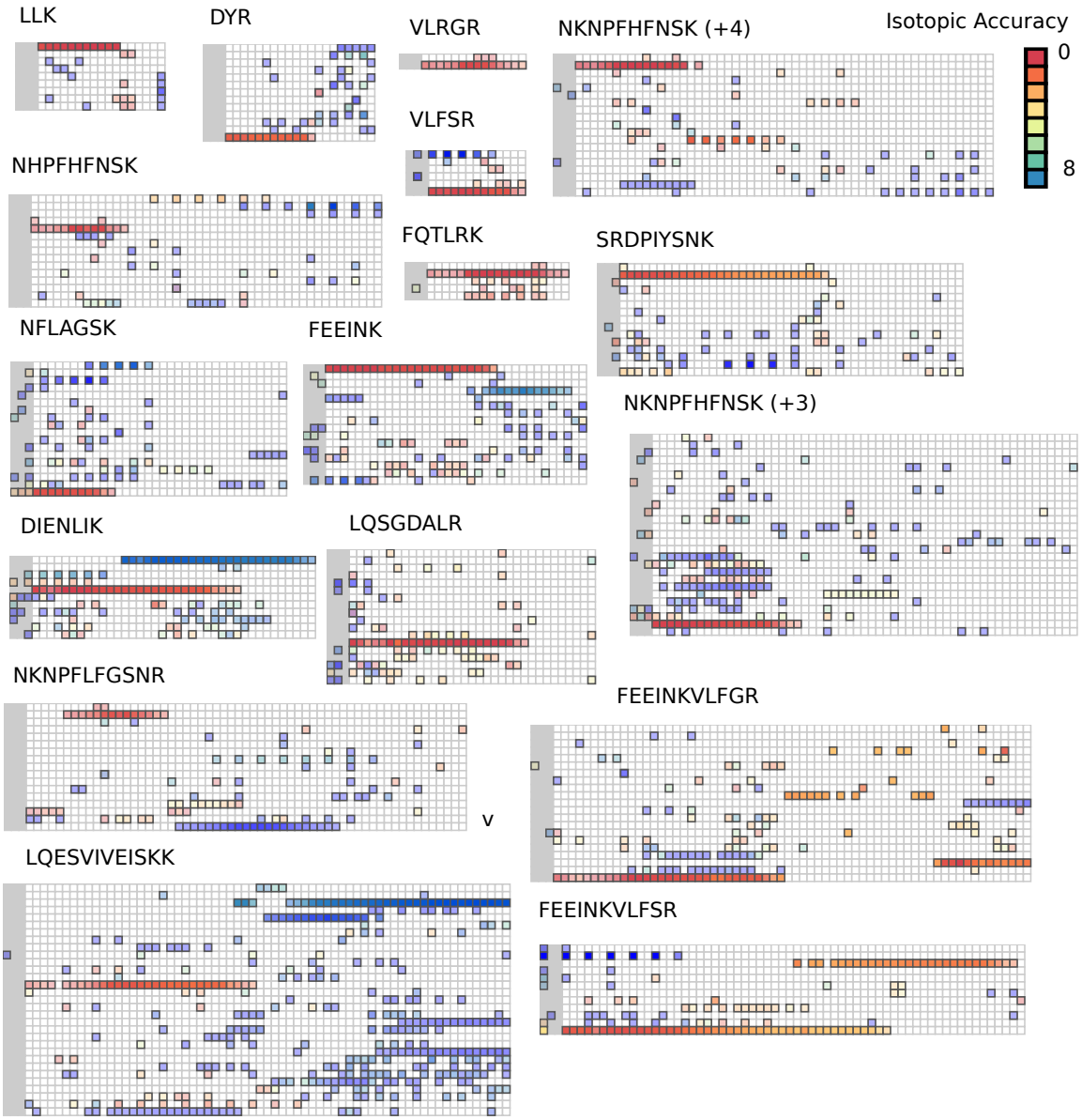


Figure C.2: Candidate Feature Sets: training data

Appendix D

Validating RIA LC Bias

In order to investigate whether there was an isotope bias within the LC domain, we extracted peptide envelopes for 600 peptides from an unlabeled *E. coli* sample. For each peptide, all isotopomers were extracted using the procedure described in Appendix B.1. However, instead of summing all ions for each isotopomer across all scans, isotopomer distributions were calculated for each scan event. Relative ion abundances, b, were then generated by dividing the raw ion count by the total number of ions extracted for that scan, such that

$$b_{i,j,k} = \Pr\{X_{i,k} = j\} = \frac{I_{i,j,k}}{\sum_j I_{i,j,k}} \tag{D.1}$$

For the j^{th} isotopomer, we are interested in how $b_{i,j,k}$ changes at each scan event in time. For clarity, we can arrange $b_{i,j,k}$ into a the column vector $b_{i,j}$, where the k^{th} element in the vector is the relative isotope abundance for the j^{th} isotopomer of peptide i at scan k. Additionally, we transform the scan index into retention times through the function $R(s_k)$, because scan events are not necessarily equally spaced in time.

A weighted linear regression was performed by attempting to find parameters $\alpha_{i,j}$

and $\beta_{i,j}$, such that

$$\begin{aligned} b_{i,j,1} &= \alpha_{i,j} R(s_1) + \beta_{i,j} \\ b_{i,j,2} &= \alpha_{i,j} R(s_2) + \beta_{i,j} \\ \vdots &= \vdots \\ b_{i,j,M} &= \alpha_{i,j} R(s_M) + \beta_{i,j} \end{aligned} \tag{D.2}$$

Or, in matrix notation:

$$b_{i,j} = A_{i,j} x_{i,j} \qquad A_{i,j} = \begin{pmatrix} R(s_1) & 1 \\ R(s_2) & 1 \\ \vdots & \vdots \\ R(s_M) & 1 \end{pmatrix} \quad x_{i,j} = \begin{pmatrix} \alpha_{i,j} \\ \beta_{i,j} \end{pmatrix} \tag{D.3}$$

Because each equation in (D.2) comes from from scans consisting of different sample sizes, $N_{i,k}$, we perform a weighted linear regression for solving (D.2) using the weighting matrix, W_i, such that

$$W_i = \begin{pmatrix} N_{i,s_1} & 0 & \dots & 0 \\ 0 & N_{i,s_2} & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & N_{i,s_M} \end{pmatrix} \tag{D.4}$$

The optimal solution for parameters stored in $x_{i,j}$ is thus

$$x^*_{i,j} = (A^T_{i,j} W_i A_{i,j})^{-1} A^T_{i,j} W_i b_{i,j} \tag{D.5}$$

Each optimal slope parameter, $\alpha^*_{i,j}$, was then compared to adjacent heavy and light isotopomers for peptide i. For example, if peptide i contained three isotopomers of

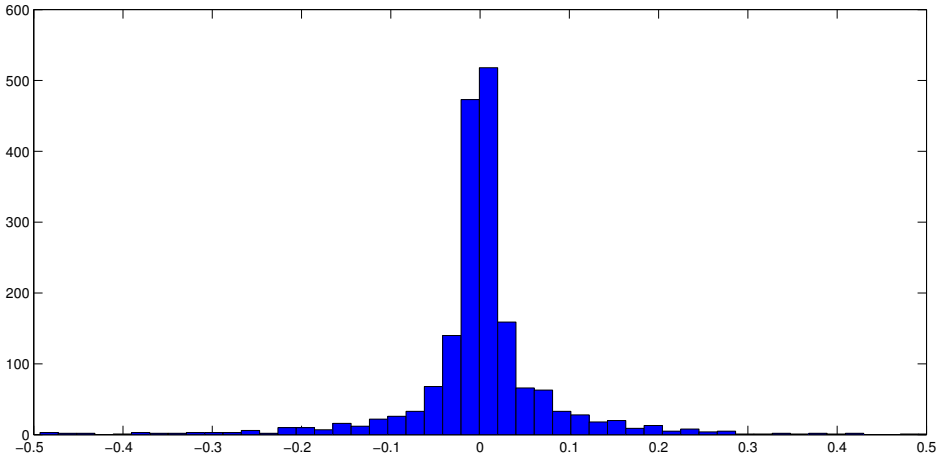


Figure D.1: Histogram of Weighted RIA Slope Differences:

mass $\{\mathrm{m}+0, \mathrm{m}+1, \mathrm{m}+2\}$, we would measured the difference in $\alpha^*_{i,j}$ between m+1 and m+0, as well as between m+2 and m+1. This was done because if there were an isotope bias, where heavy isotopes eluted off the column slower than lighter isotopes, we would observed distributions enriched with heavier isotopes in later scans. This would result in a paired difference in $\alpha^*_{i,j}$ that is greater than zero.

All paired differences in $\alpha^*_{i,j}$ are displayed in Figure D.1. A one-sample t-test was performed to demonstrate that the paired differences come from a normal distribution and have a mean of zero. The t-test resulted in a $p-$value of 0.53 with a mean and standard deviation of of -0.0016 ± 0.1115 (CI:$[-0.0067, 0.0034]$, DF: 1822), which provides no reason to infer that the data do not come from a normal distribution with a mean of zero. Thus, there is no observable isotope effect for RIA quantification.